ON CERTAIN INVARIANTS OF RANDOM DIGRAPHS AND UNIFORM HYPERGRAPHS

by

Kunal Dutta

THE INSTITUTE OF MATHEMATICAL SCIENCES, CHENNAI

A Thesis submitted to the Board of Studies in Mathematical Sciences

In partial fulfillment of requirements for

The degree of

DOCTOR OF PHILOSOPHY

of

HOMI BHABHA NATIONAL INSTITUTE

April 2014

Homi Bhabha National Institute

Recommendations of the Viva Voce Board

As members of the Viva Voce Board, we certify that we have read the dissertation prepared by **Kunal Dutta** entitled, "On Certain Invariants of Random Digraphs and Uniform Hypergraphs" and we recommend that it may be accepted as fulfilling the requirement for the Degree of Doctor of Philosophy.

_____ **Date :**

Chairman: V. Arvind

_____ **Date :**

Convener: C. R. Subramanian

_____ **Date :**

Member: Anish Sarkar

_____ **Date :**

Member: Sitabhra Sinha

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to HBNI.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it may be accepted as fulfilling the dissertation requirement.

_____ **Date :**

Guide: C. R. Subramanian

DECLARATION

I hereby declare that the investigation presented in this thesis has been carried out by me, under the guidance of Prof. C. R. Subramanian. The work is original and has not been submitted earlier as a whole or in part for a degree/diploma at this or any other Institution or University.

Kunal Dutta

Acknowledgements

I sincerely thank my thesis advisor *Prof.* C. R. Subramanian for his guidance and help at every stage of the thesis process, from evoking an interest in the topic of random graphs, showing its beauty, spending time with me whenever I walked in for innumerable discussions, to actually working out the calculations and correcting my considerable errors. I am especially grateful to him for being patient with me and hearing out my often silly or useless ideas and doubts. Words are not enough to thank his willingness to work through polishing this thesis, working practically round the clock for several days.

I am grateful to the academic faculty and administration of the Institute of Mathematical Sciences for providing excellent facilities and environment for research. Thanks *Profs.* Arvind, Venkatesh, Meena, Jam, Kamal, Saket and Vikram for all your encouragement and help. Thanks also to Mr. Vishnu Prasad, Mr. Ahmed, Mr. Shankar, Ms. Jayanthi and Ms. Latha, Mr. Gopi, Mr. Padmanabhan, Mr. Parthiban and all the office staff.

I'd like to thank my collaborators *Profs.* C.R. Subramanian, Amritanshu Prasad, Dhruv Mubayi and Wesley Calvert. I'd also like to thank Pooja Singla, C.P. Anilkumar, N.R. Aravind, and Tomász Łuczak for helpful discussions. Amri, Wesley thanks for the awesome lunch-time, coffee-time and other discussions!

I must mention the help received from my friend and former colleague Arindam Chakraborty, who informed me about theoretical computer science and IMSc, and encouraged me to apply. I am grateful to all my friends and colleagues at IMSc, especially in Theoretical Computer Science, for their unstinting help and encouragement whenever I needed. Thanks Anisa, Prajakta, Tanumoy, Neeraj, Arka, Narayanan, Partha, Rajarshi, Anil, Joydeep, Nutan, Pooja, Somnath, Bireswar, Sarbeswar, Ramchandra and Sreejith, Karteek, Yadu, Arijit, and Nitin!

I thank my parents *Drs.* Suvimal and Kaveri Dutta for being such wonderful human beings - for their love, affection and kindness, the guidance and values they gave me, the freedom they gave to pursue a field of my choice and for various things too numerous to describe in general. I thank my sister Anumita for her wonderful company and being there when I needed. I would also like to thank my uncles, aunts and cousins, Radha-masi for all their kind encouragement, ungrudging help and valuable advice. I thank my grandfathers *Prof.* S. C. Dutt and *Dr. Ing.* S. K. Nag - they were largely the sources of my inspiration - and also my grandmothers.

Lastly, I am thankful to the One, Who has guided, advised and been with me throughout my life - in, through, and beyond all the above.

Abstract

This thesis studies four problems on graphs using the Probabilistic Method. The first two are finding the maximum size of an induced acyclic tournament and acyclic subgraph respectively, in random directed graphs. The third one deals with finding the maximum size of an induced path, cycle or tree, in a random graph, while the last one is about an improved lower bound on the independence number of certain uniform hypergraphs.

Given a simple directed graph $D = (V, A)$, let the size of the largest induced acyclic tournament be denoted by $mat(D)$. Let $D \in \mathcal{D}(n, p)$ (with $p = p(n)$) be a *random* instance, obtained by choosing each of the $\binom{n}{2}$ possible undirected edges independently with probability $2p$ and then orienting each chosen edge in one of two possible directions with probability $1/2$. We show that for such a random instance, $mat(D)$ is asymptotically almost surely (a.a.s.) one of only 2 possible values, namely either b^* or $b^* + 1$, where $b^* = \lfloor 2(\log_r n) + 0.5 \rfloor$ and $r = p^{-1}$.

It is also shown that if, asymptotically, $2(\log_r n) + 1$ is not within a distance of $w(n)/(\ln n)$ (for any sufficiently slow $w(n) \to \infty$) from an integer, then $mat(D)$ is $\lfloor 2(\log_r n) + 1 \rfloor$ a.a.s. As a consequence, it is shown that $mat(D)$ is 1-point concentrated for all n belonging to a subset of positive integers of density 1 if p is independent of n. It is also shown that there are functions $p = p(n)$ for which $mat(D)$ is provably *not* concentrated in a single value. We also establish thresholds (on p) for the existence of induced acyclic tournaments of size i which are sharp for $i = i(n) \to \infty$.

We also analyze a polynomial time heuristic and show that it produces a solution whose size is at least $\log_r n + \Theta(\sqrt{\log_r n})$. Our results are valid as long as $p \geq 1/n$. All of these results also carry over (with some slight changes) to a related model which allows 2-cycles.

For the next problem, given a simple directed graph $D = (V, A)$, let the size of the largest induced acyclic subgraph ***(dag)*** of D be denoted by $mas(D)$. Let $D \in \mathcal{D}(n, p)$ be a *random* instance as in the previous chapter. We obtain improved bounds on the range of concentration, upper and lower bounds of $mas(D)$. Our main result is that

$$mas(D) \;\; \geq \;\; \lfloor 2 \log_q np - X \rfloor$$

where $q = (1-p)^{-1}$, $X = 1$ if $p \geq n^{-1/3+\epsilon}$ ($\epsilon > 0$ is any constant), $X = W/(\ln q)$ if $p \geq C/n$, and C, W are suitably large constants. This improves the previously known lower bounds given by Spencer and Subramanian [61, 63] where there is an $O(\ln \ln np / \ln q)$ term instead of X. We also obtain a slight improvement on the upper bound, using an upper bound on the number of acyclic orientations of an undirected graph. We also analyze a polynomial-time heuristic to find a large induced dag and show that it produces a solution whose size is at least $\log_q np + \Theta(\sqrt{\log_q np})$. Our results also carry over to the related model $\mathcal{D}_2(n, p)$.

The next problem deals with random *undirected* graphs. We study the concentration of the largest induced paths, trees and cycles (holes) in the random graph model $\mathcal{G}(n, p)$, and prove a 2-point concentration for the size of the largest induced path and hole, for all $p \geq n^{-1/2} \ln^2 n$. As a corollary, we obtain an improvement over a result of Erdős and Palka [29] concerning the size of the largest induced tree in a random graph.

In the last problem, we consider the independence number of K_r-free graphs and linear k-uniform hypergraphs in terms of the degree sequence, and obtain new lower bounds for them. This answers some old questions raised by Caro and Tuza [21]. Our proof technique is an extension of a method of Caro and Wei [20, 72], and we also give a new short proof of the main result of [21] using this approach. As byproducts, we also obtain some non-trivial identities involving binomial coefficients, which may be of independent interest.

Contents

List of Figures

List of Publications from this Thesis

Journals

- Kunal Dutta, D. Mubayi, C. R. Subramanian: New lower bounds for the independence number of sparse graphs and hypergraphs.
  SIAM Journal of Discrete Mathematics (2012) **26**(3) 1134-1147.

- Kunal Dutta, C. R. Subramanian: Induced acyclic tournaments in random digraphs: Sharp concentration, thresholds and algorithms.
  Accepted in *Discussiones Mathematicae Graph Theory*.

Conferences

- Kunal Dutta, C. R. Subramanian: Largest Induced Acyclic Tournament in Random Digraphs: A 2-point concentration.
  Proceedings of LATIN 2010 (9th Latin American Theoretical Informatics Symposium), Oaxaca, Mexico, April 2010.

- Kunal Dutta, C. R. Subramanian: Induced acyclic subgraphs in random digraphs : Improved bounds.
  Proceedings of AofA'10 (21st International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms), Vienna, Austria, June 2010.

- K. Dutta, C. R. Subramanian: On induced acyclic subgraphs in sparse random digraphs.
  Proceedings of EuroComb'11 (European Conference on Combinatorics, Graph Theory and Applications), Budapest, Hungary, August 2011.

Preprints (Submitted)

- K. Dutta, C. R. Subramanian: Improved bounds on induced acyclic subgraphs in random digraphs.

Chapter 1

Introduction

1.1 Graph Theory

Many problems occuring in diverse areas such as fields related to computer science, VLSI design, operations research, economics and even statistical physics, deal with the combinatorics of the interconnections between certain elements or nodes. A node could be a resistor, transistor, capacitor etc. in an electrical network, or a city in a network of intercity roadways or a web router in a computer network. For example, consider the following problems:

(i) A postman has to visit certain houses in an area to deliver mail, and return. In what order should he visit the houses to minimize his travel? What if some of the routes connecting these houses are much longer than the straight-line distance, while shortcuts exist between others?

(ii) Or, how should m jobs be distributed amongst n persons where each person is qualified for a certain subset of jobs, and each job requires a certain number of persons?

All these problems can be modelled by representing certain elements of the problem instance as nodes, and specifying interconnections between the nodes. The common feature in all these is that the internal structure of a node is not needed for modelling the problem. Only the interconnections (or lack of them) between nodes are important.

A *graph* is a very simple structure that has been found to be extremely useful in modelling these, and many other such problems occurring in various disciplines. A graph $G = (V, E)$ is a set V of elements called vertices, together with a collection E of pairs of vertices called *edges*. Graph theory is a branch of discrete mathematics that deals with the study of graphs as

abstract objects. Graph theory (along with Topology) originated with Leonhard Euler's 1736 solution of the famous *Königsberg Bridges* problem. Over the last two-and-a-half centuries, graph theory has been systematically built up from a collection of problem-solving "tricks" to an important branch of Discrete Mathematics. In the last sixty years or so, the emergence of Computer Science and the significant role of graph theory in modelling problems in computer science has led to a tremendous increase in interest in the area and acceleration in its growth. Many computational problems in Computer Science, Operations Research and several other areas can be modelled as graphs, and graph algorithms are used to solve these problems.

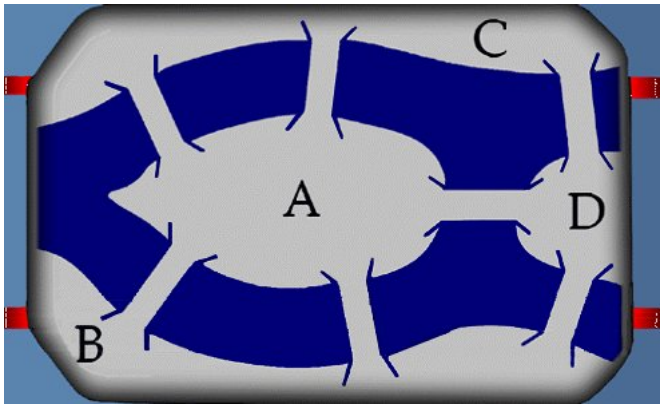

Figure 1.1: The Bridges of Königsberg

1.2 Random Graphs

Random Graphs is the probabilistic/statistical analysis of graphs. In this branch of Graph Theory, instead of looking at individual graphs, one looks at probability distributions over n-vertex graphs, and studies their properties and invariants from a probabilistic or statistical point of view. The theory of random graphs can be thought of as a confluence of Probability Theory and Graph Theory. The concept of a *random* graph was first used by Erdös in 1947 [25] in a famous result that proved a lower bound on the diagonal Ramsey numbers $R(k, k)$. Since then, it remained more or less an interesting "trick" until 1959, when Erdős and Rényi [26, 27] published a series of seminal papers that established the study of random graphs (random structures in general) as a separate field in itself.

Studying random graphs has applications in many areas of research, from graph theory, algorithms, operations research, computational complexity and statistical mechanics, to game theory and economics. These applications can be roughly classified into three categories:

- Showing the existence of one (or many) graphs having certain properties.
- Giving insights into the structure or behaviour of "most" (in some sense) graphs - or sometimes, all "large enough" graphs.
- Enabling a stochastic analysis of certain graph algorithms, network models etc.

The Probabilistic Method The rise of random graphs as an area of study has been hand-in-hand with the rise of a branch of combinatorics called "Probabilistic Combinatorics". Probabilistic Combinatorics uses ideas from probability theory to solve combinatorial problems. In fact, Erdős 's 1947 paper [25] is also cited as one of the first papers to use the probabilistic method of solving combinatorial problems. Initially regarded by the majority Combinatorics community as just "a fanciful way of counting", it was later shown to be more powerful than straightforward counting, as one could leverage the existing techniques in probability theory to solve problems that would have been very difficult to solve using direct counting techniques. Today, the Probabilistic Method has found applications that are both wide and deep, in areas ranging from number theory, additive combinatorics, the geometry of Banach spaces, and graph theory, to - perhaps most significantly - theoretical computer science. It has also fuelled much development in Discrete Probability, particularly in the discovery of new inequalities like the Lovász Local Lemma and concentration inequalities.

1.3 Thesis Outline

The outline of the rest of this thesis is as follows. Some preliminaries of Discrete Probability and graphs, directed graphs, hypergraphs and some random models are presented in Chapters 2 and 3 respectively. The work in Chapters 4-7 was done in collaboration with my advisor Prof. C. R. Subramanian. Chapter 8 was done in collaboration with Profs. C. R. Subramanian and Dhruv Mubayi.

The thesis is concerned with the study of a few graph invariants of a random digraph or a random graph. We present both theoretical and algorithmic results related to these invariants. For random digraphs, we study and determine the most likely values that the size of a largest induced acyclic tournament or the size of a largest induced acyclic subgraph can take. A tournament is a directed graph obtained by orienting the edges of a complete undirected graph. For random graphs, we study the most likely values that the sizes of a largest induced path, a largest induced cycle or a largest induced tree can take. For the case

of tournaments (in random digraphs) and induced paths and cycles (in random graphs), it is shown that the most likely values are from a set of two consecutive positive integers. In addition, we also obtain new lower bounds on the size of a largest independent set in an arbitrary (not a random instance) graph or a linear hypergraph in terms of its degree sequence. An outline of the presentation of these results is given below.

In Chapter 4 we present several theoretical and algorithmic results relating to largest induced acyclic tournaments in random directed graphs. These include a general two-point concentration result, special cases where one-point concentration can or *cannot* be obtained and threshold results, as well as algorithms to find large acyclic tournaments induced in random digraphs, alongwith their analyses.

In Chapters 5 and 6 we present results on a related topic: the size of a largest induced acyclic *subgraph* present in a random directed graph. These include concentration results as well as lower and upper bounds. An interesting case for the lower bound is when the variance of the random variable under question becomes much larger than the square of its expectation, and most second-moment based techniques seem to fail. A combination of Talagrand's inequality with the Paley-Zygmund inequality is used to handle this case. We focus on lower bounds in Chapter 5. In Chapter 6, we cover upper bounds as well as the algorithmic question of *finding* a large induced acyclic subgraph. As in Chapter 4, heuristics are suggested and analyzed.

Next, in Chapter 7 we study a group of problems in random *graphs* - that of determining the size of a largest induced path, cycle (hole) or tree. The concentration of the sizes of largest induced paths, trees and cycles are studied in the random graph model $\mathcal{G}(n,p)$. A 2-point concentration is proved for the size of the largest induced path and hole, for all $p \geq n^{-1/2}(\ln n)^2$. As a corollary, an improvement is obtained over a result of Erdős and Palka [29] concerning the size of the largest induced tree in a random graph.

Chapter 8 deals with a different problem: finding a degree-sequence based lower bound on the independence number (size of a largest independent set) of general and linear k-uniform hypergraphs. A new proof of a theorem given by Caro and Tuza [21] is given, and some old questions asked by them are answered. The key idea is an extension of the random-permutation technique used in the probabilistic proof of Turan's theorem given by Bopanna-Caro-Wei [9] to hypergraphs, particularly linear hypergraphs.

Finally, in Chapter 9, some open problems and potential future directions are presented.

Notation : Throughout, we use standard notation. $\Re^+$ denotes the positive real numbers, $\mathcal{N}$ denotes the set of natural numbers. Given a natural number $n \in \mathbb{Z}^+$, we indicate the set $\{1, \ldots, n\}$ by $[n]$. We also use standard notations like $O(.)$, $\Omega(.)$, $o(.)$ and $\omega(.)$ with the usual meanings. We ignore floors and ceilings wherever they are not crucial. For a sequence of events $\{E_n\}_{n\geq 1}$ defined over a corresponding sequence of probability spaces $\{\Omega_n\}_{n\geq 1}$, we use the phrase E_n *holds asymptotically almost surely (a.a.s)* to mean that the sequence of probabilities $\{P_n\}_{n\geq 1}$ associated with the events $\{E_n\}$ tends to 1 as n tends to infinity.

Chapter 2

Basics of Discrete Probability

2.1 Basic Definitions

We provide a quick overview of standard notions, facts and results from probability theory. We focus throughout only on discrete probability spaces. Throughout the thesis, a *sample space* Ω refers to a finite or a countably infinite set. A σ-field over Ω is a collection $\mathcal{F}$ of subsets of Ω, called *events*, such that (i) empty set $\emptyset$ belongs to $\mathcal{F}$, (ii) $\mathcal{F}$ is closed under complementations, and (iii) $\mathcal{F}$ is closed under the union of members of any countable subcollection of itself.

A *probability space* is a triple (Ω, Σ, P), where Ω is a sample space, Σ is a σ-field over Ω, and $P : \Sigma \to [0, 1]$ is a function such that (i) $P(\Omega) = 1$, (ii) P is countably additive, that is, for any countable collection $\{A_i\}_{i \in I}$ of mutually disjoint events from Σ, $P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$.

If $\mathcal{E}_2$ is an event with non-zero probability, then the *conditional probability* of $\mathcal{E}_1$ given $\mathcal{E}_2$ (denoted by $\mathbf{Pr}[\mathcal{E}_1|\mathcal{E}_2]$) is the probability of $\mathcal{E}_1$ in the probability space conditioned by $\mathcal{E}_2$ and is defined to be

$$\mathbf{Pr}[\mathcal{E}_1|\mathcal{E}_2] \;=\; \mathbf{Pr}[\mathcal{E}_1 \cap \mathcal{E}_2]/\mathbf{Pr}[\mathcal{E}_2]$$

A finite collection of events $\{\mathcal{E}_i | i \in I\}$ is *totally independent* if for all subsets $S \subseteq I$,

$$\mathbf{Pr}\left[\bigcap_{i \in S} \mathcal{E}_i\right] \;=\; \prod_{i \in S} \mathbf{Pr}[\mathcal{E}_i]$$

These events are *k-wise independent* if every subcollection of k events is totally independent. For $k = 2$, this notion is often referred to as *pairwise independence*.

A *real random variable* X (associated with a probability space $(\Omega, \mathcal{F}, P)$) is a real-valued function $X : \Omega \to \Re$, such that $\forall x \in \Re$, the set $\{\omega \in \Omega | X(\omega) \leq x\}$ is an event in $\mathcal{F}$. By our assumption on the probability spaces, each of the random variables we will consider is a *discrete random variable* , that is, it takes values only from a range which is either finite or a countable subset of $\Re$. An *indicator random variable* is a discrete random variable whose range is $\{0, 1\}$. Throughout, we will study the behavior of several random variables defined over a common probability space. Two random variables X and Y are independent if, for any two events $\mathcal{E}_X$ and $\mathcal{E}_Y$ associated only with X and Y respectively, $\mathcal{E}_X$ and $\mathcal{E}_Y$ are independent. This definition is naturally extendible to any finite collection of random variables.

The *expectation* or *first moment* (denoted by $E[X]$) of a random variable X is defined by $\mathbf{E}[X] = \sum_x x\mathbf{Pr}[X = x]$ where $\mathbf{Pr}[X = x] = \mathbf{Pr}[\{\omega | X(\omega) = x\}]$ and the summation is over the range of X.

In general, the *k-th moment* m_X^k and the *k-th central moment* μ_X^k of a random variable X are defined as follows:

$$\begin{aligned} m_X^k &= \mathbf{E}[X^k] \\ \mu_X^k &= \mathbf{E}[(X - \mathbf{E}[X])^k] \end{aligned}$$

For $k = 2$, the central moment is often referred to as the *variance*, and is denoted by either $Var(X)$ or $\sigma^2 = \sigma^2(X)$ and its positive square root is often called the *standard deviation*, denoted by $\sigma = \sigma(X)$.

The *Covariance* of two random variables X, Y is defined to be

$$COV(X, Y) = \mathcal{E}[(X - \mathcal{E}[X])(Y - \mathcal{E}[Y])] = \mathcal{E}[XY] - \mathcal{E}[X]\mathcal{E}[Y]$$

Clearly, when $X = Y$, the covariance reduces to the variance of X.

2.2 Basic Relations Between Probabilistic Operators and Inequalities

The following theorem is a simple but powerful observation and follows from the definition of the notion of expectation.

Theorem 2.2.1 *(Linearity of Expectation): Let $X_1, X_2, ..., X_k$ be arbitrary random variables, and $h(X_1, .., X_k)$ be a linear function. Then*

$$E[h(X_1, ..., X_k)] \quad = \quad h(E[X_1], E[X_2], ..., E[X_k])$$

Proposition 2.2.2 *(Boole-Bonferroni Inequalities): Let $\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_n$ be arbitrary events. Then for even k*

$$\mathbf{Pr}\left[\bigcup_{i=1}^{n} \mathcal{E}_i\right] \quad \geq \quad \sum_{j=1}^{k} (-1)^{j+1} \sum_{i_1 < i_2 < ... < i_j} \mathbf{Pr}\left[\bigcap_{r=1}^{j} \mathcal{E}_{i_r}\right]$$

and for odd k

$$\mathbf{Pr}\left[\bigcup_{i=1}^{n} \mathcal{E}_i\right] \quad \leq \quad \sum_{j=1}^{k} (-1)^{j+1} \sum_{i_1 < i_2 < ... < i_j} \mathbf{Pr}\left[\bigcap_{r=1}^{j} \mathcal{E}_{i_r}\right]$$

The case $k = 1$ is used very often, and is referred to as the union bound.

2.2.1 Variance of Sums of (Indicator) Random Variables

When the random variable in question is the sum of other random variables, it is often useful to express the second central moment in terms of such moments of the summands as follows: Let $X = \sum_i X_i$ be the sum of a finite number of other random variables X_i, $i \in I$. Then,

Proposition 2.2.3

$$\sigma^2(X) \quad = \quad \sum_i \sigma^2(X_i) + \sum_{i \neq j} COV(X_i, X_j)$$

where the second summation is over ordered pairs (i, j). The special case when the X_i's are indicator random variables is especially useful and will be used several times in the coming chapters. The following well-known upper bounds on tail probabilities will often be applied later in the thesis.

Theorem 2.2.4 *(Markov's Inequality): Let Y be a non-negative random variable. Then for all $t \in \Re^+$, we have $t\mathbf{Pr}[Y \geq t] \leq E[Y]$ and hence*

$$\mathbf{Pr}[Y \geq t] \quad \leq \quad E[Y]/t$$

Theorem 2.2.5 *(Chebyshev's Inequality): Let X be a random variable with expectation μ_X and standard deviation σ_X. Then for all $t \in \Re^+$,*

$$\mathbf{Pr}[|X - \mu_X| \geq t\sigma_X] \quad \leq \quad (1/t^2)$$

The proof follows simply by applying Markov's inequality to the random variable $(X - \mu_X)^2$.

The Paley-Zygmund Inequality is a slightly weaker version of the one-sided version of the previous theorem.

Theorem 2.2.6 *(Paley-Zygmund Inequality): Let X be a random variable which takes non-negative values. Then, for every $\theta : 0 < \theta < 1$:*

$$\mathbf{Pr}[X > \theta E[X]] \quad \geq \quad (1-\theta)^2 \frac{E[X]^2}{E[X^2]}$$

Taking the limit as $\theta \to 0$, the Paley-Zygmund Inequality gives the following:

$$\mathbf{Pr}[X > 0] \geq E[X]^2 / E[X^2].$$

Theorem 2.2.7 *(Chernoff-Hoeffding Bounds): Let $X_1, X_2, ...X_n$ be independent indicator random variables such that, $\mathbf{Pr}[X_i = 1] = p_i$ for each i. Then, for $X = \sum_{i=1}^{n} X_i$, $\mu = E[X] = \sum_{i=1}^{n} p_i$,*

$$\begin{aligned} \mathbf{Pr}[X \geq (1+\delta)\mu] \quad &\leq \quad \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu} \text{ for any } \delta \geq 0 \\ \mathbf{Pr}[X \leq (1-\delta)\mu] \quad &\leq \quad e^{-\mu\delta^2/2} \text{ for any } \delta, \ \ 0 \leq \delta \leq 1 \end{aligned}$$

The above two inequalities are often combined (for every $\epsilon \in [0,1]$) as:

$$\mathbf{Pr}[|X - \mu| \geq \epsilon\mu] \leq 2e^{-\mu\epsilon^2/3}$$

2.3 Advanced Inequalities

2.3.1 Lovász Local Lemma

For a finite collection of totally independent events, the probability of the intersection of all these events is exactly the product of the probabilities of the individual events. Often, the events under consideration are not independent and we need to estimate the probability of the intersection which may become complicated due to the dependencies between events. However, when each event is influenced by only a few other events, we can give a non-zero lower bound on the required probability. The following lemma, proved by Erdős and Lovász [28] in 1975 can be extremely useful in such situations.

Lemma 2.3.1 *Let $\mathcal{A} = \{E_1, E_2, ..., E_n\}$ be a collection of events such that for all $1 \leq i \leq n$, E_i is totally independent of all events but those in $D_i \subseteq \mathcal{A} \backslash \{E_i\}$.*
If there exists a real sequence $\{x_i\}_{i=1}^n$, $x_i \in [0,1)$, such that

$$\begin{aligned} \forall i \in [n],\ \mathbf{Pr}[E_i] &\leq x_i \prod_{j:E_j \in D_i} (1 - x_j), \text{ then} \\ \mathbf{Pr}\left[\bigcap_{i=1}^{n} \overline{E}_i\right] &\geq \prod_{i=1}^{n} (1 - x_i) > 0 \end{aligned}$$

2.3.2 Concentration Inequalities

For the case when the indicator variables X_i's are not independent, Chernoff's bounds cannot be applied. However, if the variables fulfil certain weaker conditions, some concentration results can still be obtained. Before presenting the first theorem, we mention the conditions which need to be fulfilled:

Definition A collection of random variables $\{X_1, X_2, ...\}$ is a *discrete-time martingale* if it satisfies the following conditions:

(i) $\forall n, E[|X_n|] < \infty$, and

(ii) $\forall n,\ E[X_{n+1}|X_1, X_2, ..., X_n] = X_n$.

Doob's Martingale Let $\Omega = \prod_{i=1}^n \Omega_i$ form probability space with a probability distribution P over Ω. Suppose $f : \Omega \to \Re$ is $f = f(X_1, \ldots, X_n)$, then the sequence $Y_i = \{\mathcal{E}[f|X_1, \ldots, X_i]\}_{i=0}^n$, where $Y_0 = \mathcal{E}[f]$ by definition, forms a martingale, known as a *Doob Martingale*.

Note that in the Doob Martingale defined above, $Y_0 = \mathcal{E}[f]; Y_n = f$.

Theorem 2.3.2 *(Azuma's Inequality): Suppose $\{X_k : k = 0, 1, 2, ...\}$ is a martingale, and $|X_k - X_{k-1}| \leq c_k$. Then for all $n > 0$, $t \in \Re^+$,*

(i) $\mathbf{Pr}(X_n - X_0 \geq t) \leq e^{-t^2/(2\sum_{k=1}^n c_k^2)}$.

(ii) $\mathbf{Pr}(X_n - X_0 \leq -t) \leq e^{-t^2/(2\sum_{k=1}^n c_k^2)}$.

Theorem 2.3.3 (see [9, 40]) *Suppose that $Z_1, \ldots, Z_N$ are independent random variables taking their values in some sets $\Gamma_1, \ldots, \Gamma_N$ respectively. Suppose further that $X = f(Z_1, \ldots, Z_N)$, where $f : \Gamma_1 \times \ldots \Gamma_N \to \mathcal{R}^+$ is a function such that for some function $\psi : \mathcal{N} \to \mathcal{N}$, the following two conditions hold:*

(i) If $z, z' \in \Gamma = \prod_{i=1}^N \Gamma_i$ differ only in one component, then $|f(z) - f(z')| \leq 1$.

(ii) If $z \in \Gamma$ and $r \in \mathcal{R}^+$ with $f(z) \geq r$, then there exists a set $J \subseteq \{1, \ldots, N\}$ with $|J| \leq \psi(r)$, such that we have $f(y) \geq r$ for every $y \in \Gamma$ with $y_i = z_i$ whenever $i \in J$. (f is said to be ψ-certifiable in such a case).

Then for every $r \in \mathcal{R}^+$ and $t \geq 0$,

$$\mathbf{Pr}[X \leq r - t]\mathbf{Pr}[X \geq r] \quad \leq \quad e^{-t^2/4\psi(r)}$$

In particular, if m is a median of X, then for every $t \geq 0$,

$$\mathbf{Pr}[X \leq m - t] \leq 2e^{-t^2/4\psi(m)}$$

and

$$\mathbf{Pr}[X \geq m + t] \leq 2e^{-t^2/4\psi(m+t)}.$$

A median of X is any value m such that $\mathbf{Pr}[X \leq m] \leq 0.5$ and $\mathbf{Pr}[X \geq m] \leq 0.5$. □

Chapter 3

Basics of (Random) Graph Theory

3.1 Basic Definitions

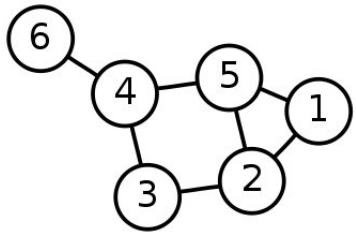

Figure 3.1: A simple undirected graph

We provide a quick review of some basic notions and facts about graphs, digraphs and hypergraphs. For a set V and a nonnegative integer k, we use the notation $\binom{V}{k}$ to denote the set of all k-sized subsets of V. We use 2^V to denote the collection of all subsets of V.

A (simple, undirected) *graph* $G = (V, E)$ is an ordered pair where V is a set of elements known as *vertices*, and $E \subseteq \binom{V}{2}$ is a collection of elements known as (undirected) *edges*.

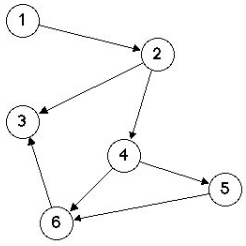

Figure 3.2: A directed graph

A *directed graph* or digraph $D = (V, A)$ is a graph whose edge set consists of ordered pairs $A \subset V \times V$ called *arcs* or directed edges. For a given unordered pair $\{u, v\} \in \binom{V}{2}$, we allow one or both of the arcs (u, v) and (v, u). We do not allow two copies of the same arc (u, v). If, for every $\{u, v\} \in \binom{V}{2}$, at most one of the two arcs $(u, v), (v, u)$ is allowed, then it is called a *simple* digraph or also as an *oriented graph*. An oriented graph D is also a digraph obtained by orienting the edges of an undirected graph G. In that case, we say that D has G as its underlying undirected graph. A *tournament* is a simple digraph $D = (V, A)$ whose underlying undirected graph is $G = (V, E)$ where $E = \binom{V}{2}$.

For a graph $G = (V, E)$ or a directed graph (simple or otherwise) $D = (V, A)$, a path P in G (or D) is a sequence $(u_0, e_1, u_1, \ldots, e_k, u_k)$ such that u_i's are distinct (except possibly that $u_0 = u_k$) and each e_i is an edge (or arc) in E (or in A) joining u_{i-1} and u_i. If $u_0 = u_k$, then P forms a cycle (or a directed cycle). For graphs or simple digraphs, every cycle should involve at least k vertices and k edges for some $k \geq 3$. But, for a general digraph, we can have cycles on just two vertices as, for example, the cycle formed by (u, v) and (u, v) assuming both of these arcs are present in A. The length of a path or a cycle (undirected or directed) is the number of vertices in it.

An graph G (or digraph D) is *acyclic* if G (or D) contains no cycle (or directed cycle). An acyclic graph on n vertices can have at most $n - 1$ edges in it. But, an acyclic digraph on n vertices can have as many as $\binom{n}{2}$ arcs.

An *independent set* in an undirected graph $G = (V, E)$ is a subset A of V such that $\binom{A}{2} \cap E = \emptyset$. The maximum size of an independent set in G is a well-studied numerical invaraint (a value being the same for isomorphic copies) of graphs and is known as the independence number of G and is denoted by $\alpha(G)$.

A *hypergraph* (alternatively, a *set system*) $H = (V, E)$ is a pair, consisting of a set V of vertices, and a collection $E \subset 2^V$ of subsets of V which are known as *edges* or hyperedges. A hypergraph (or set system) is said to be k-uniform (for some $k \geq 1$) if $E \subseteq \binom{V}{k}$. A graph is a 2-uniform hypergraph. An *independent set* in a k-uniform hypergraph $G = (V, E)$ is a subset A of V such that $\binom{A}{k} \cap E = \emptyset$. Independence number (denoted by $\alpha(G)$) is similarly defined by for hypergraphs.

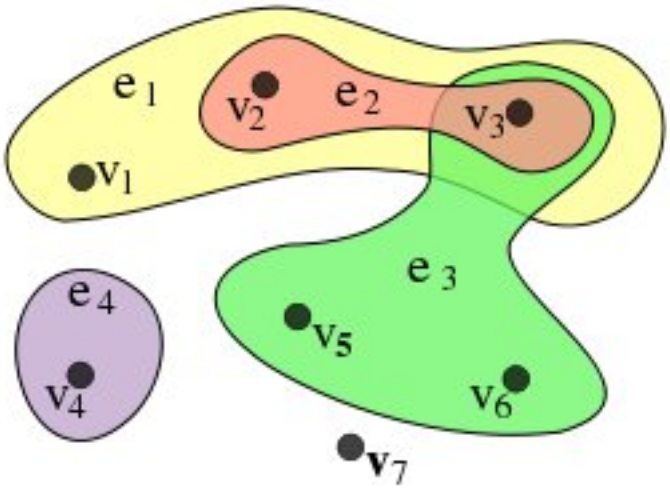

Figure 3.3: A hypergraph

3.2 Some Models of Random Graphs

In this section, we give a brief overview of various models for defining random graphs (or digraphs). Before talking about the behaviour or characteristics of random graphs, we need to define what we mean by "random", that is, we need to decide on a probability space from which such graphs are to be chosen.

Please note that unless stated otherwise, all graphs in this thesis are labelled graphs. For convenience, (without loss of generality) we take the vertex set to be $V = \{1, 2, ..., n\}$, and indicate the set of all n-vertex undirected graphs by $\mathcal{G}^n$. We use N to denote the quantity $\binom{n}{2}$.

1. $\mathcal{G}(n, M)$: In this model, the probability space is uniform distribution over all graphs having n vertices and M edges. Consider the set $\mathcal{S}$ of all graphs over V having M edges. The size of $\mathcal{S}$ is clearly $\binom{N}{M}$. Convert $\mathcal{S}$ into a probability space $\mathcal{G}(n, M)$ by making all graphs in it equiprobable. Each element of $\mathcal{S}$ is chosen with probability $\binom{N}{M}^{-1}$.

2. $\mathcal{G}(n, p)$: This is closely related to the $\mathcal{G}(n, M)$ model. Instead of fixing the *number* of edges, each potential edge in $\binom{V}{2}$ is included as an edge with the same *probability* p. The random choices are independent for different potential edges. Each graph over V and having m edges is chosen with probability $p^m(1-p)^{N-m}$. Thus, the number of edges is itself a random variable, but all graphs with a given number of edges are

equiprobable. It is easily seen that $\mathcal{G}(n, 1/2)$ is the uniform distribution over all n-vertex simple graphs. This model and the previous one have been studied in great detail with many interesting observations forming the theory of random graphs.

3. $\mathcal{D}(n, p)$: This is a model of simple directed graphs whose study forms a major part of this thesis. It was introduced by Subramanian in 2003 (see [63]). Let the vertex set be $V = \{1, 2, ..., n\}$. Choose each *undirected edge* joining distinct elements of V independently with probability $2p$. For each chosen $\{u, v\}$, independently orient it in one of the two directions $\{u \to v, v \to u\}$ in D with equal probability $= 1/2$. The resulting directed graph is an orientation of a simple graph, that is, there are no 2-cycles. Each simple digraph on V having m arcs is chosen with probability $p^m(1-2p)^{N-m}$.

There is also a closely related model, which has been well-studied in the literature (see e.g. [51]), which we denote by $\mathcal{D}_2(n, p)$ following the notation of [61]:

4. $\mathcal{D}_2(n, p)$: It is similar to the $\mathcal{D}(n, p)$ model, except that 2-cycles are allowed. The vertex set is again $V = [n]$, and each possible arc $(i, j) \in V \times V$, $(i \neq j)$ is chosen independently with probability $p = p(n)$. Each digraph on V having m arcs is chosen with probability $p^m(1-p)^{n(n-1)-m}$.

5. $\mathcal{G}(n, d)$: In this model, the graph is constrained to be d-regular. The uniform distribution over all n-vertex, d-regular graphs is chosen.

There are several other less common models. We mention briefly some of them:

6. $\mathcal{G}\{K(n, m); p\}$: Labelled bipartite graphs with vertex sets U and W, $|U| = n$, $|W| = m$, in which each $U - W$ edge is chosen independently with probability p.

7. $\vec{\mathcal{G}}_{k-out}$. Each vertex in V chooses k other vertices uniformly and independently from the $\binom{n-1}{k}$ choices, and directs arcs from itself towards these vertices.

8. "Small worlds"- the Albert-Barabási model: In order to model various large networks like the world-wide web, acquaintance networks, power-grid and telephone networks etc., Barabási and Albert [11] defined a "preferential attachment" model. In this model, vertices are added to the graph sequentially, and when a new vertex v is added, the probability of an edge between v and a pre-existing vertex u is proportional to the pre-existing degree of u. Thus, a vertex having more neighbours is likely to keep gaining neighbours at a faster rate than a vertex having fewer neighbours. Most of the

initial work was of an experimental/simulation-based nature. Bollobás and Riordan [18] were the first to define a mathematical model with the required properties and prove some theoretical results like precise bounds on the diameter.

This thesis will be concerned with only the three random models $\mathcal{G}(n,p)$ (for undirected graphs), $\mathcal{D}(n,p)$ (for simple digraphs) and $\mathcal{D}_2(n,p)$ (for digraphs). Also, we will be mostly concerned with the asymptotic behavior of the models as n becomes large. Precisely, we will consider a sequence $\{\mathcal{G}(n,p) : n \geq 1\}$ (similarly for $\mathcal{D}(n,p)$ and $\mathcal{D}_2(n,p)$) of random graphs where we allow $p = p(n)$ to depend on n. We will be interested in the "typical behavior" of $G \in \mathcal{G}(n,p)$ as $n \to \infty$. By "typical", we mean a statement which holds with probability $q(n)$ where $q(n) \to 1$. In that case, we say that a.a.s., the random graph G exhibits the behavior. These are described precisely later in the following chapters.

3.3 Some Random Graph Phenomena

3.3.1 Thresholds

One of the most intriguing discoveries made by Erdős and Rényi during their seminal work on random graphs (see [30, 31]) was that of *threshold* or *phase transition* phenomena. Basically, they discovered that for many graph properties, as the number of edges $M = M(n)$ is increased from zero to $\binom{n}{2}$, the random graph $G \in \mathcal{G}(n,M)$ goes from a.a.s. *not having* the property to a.a.s. having it (or vice versa), for a very small change in the number of edges. By the contiguity of $\mathcal{G}(n,p)$ and $\mathcal{G}(n,M)$, the same could be said for the $\mathcal{G}(n,p)$ model, in terms of the edge probability $p = p(n)$ increasing from zero to 1.

A property $\mathcal{P}$ is a collection of labelled graphs which is closed under isomorphism. The random graph $G \in \mathcal{G}(n,p)$ is said to have the property $\mathcal{P}$ if $G \in \mathcal{P}$ a.a.s. A graph property $\mathcal{P}$ is said to have a threshold $p_0 = p_0(n)$ if there exists some $q = q(n)$ such that (i) G doesn't have $\mathcal{P}$ a.a.s. if $p \leq p_0(n) - q(n)$ and (ii) G has $\mathcal{P}$ a.a.s. if $p \geq p_0(n) + q(n)$, or vice versa. The case when the random graph $\mathcal{G}(n,p)$ goes from a.a.s. not having $\mathcal{P}$ to a.a.s. having $\mathcal{P}$ is called a $0-1$ threshold, and the opposite case is known as a $1-0$ threshold.

3.3.2 Concentration of Measure

A closely related phenomenon, which will be the focus of much of this thesis, is that of *concentration of measure*. As the name suggests, it means the presence of almost all the probability mass of a random variable in a very small interval. Given a sequence of random

variables $X = X(n)$ over a sequence of probability spaces $\Omega = \Omega(n)$, we say that X is concentrated in the range $[a, b]$ ($a = a(n), b = b(n)$) if the probability that X lies outside the range $[a, b]$ goes to zero as n tends to infinity:

$$\lim_{n\to\infty} \mathbf{Pr}(X \notin [a, b]) = 0.$$

A *graph invariant* is a real-valued function $f(G)$ (G is a graph or a digraph) such that $f(G) = F(H)$ whenever G and H are isomorphic. Clique number $\omega(G)$ or the independence number $\alpha(G)$ are examples of graph invariants. Given a graph invariant $f = f(G)$, one gets a corresponding random variable, $X = f(G) : G \in \mathcal{G}(n, p)/\mathcal{D}(n, p)$. An important section of the theory of random graphs deals with the asymptotic concentrations of various graph invariants for different ranges of the edge probability $p = p(n)$.

Chapter 4

Largest Induced Acyclic Tournaments in Random Digraphs

4.1 Introduction

A *tournament* is a simple directed graph whose underlying undirected graph is a complete graph. A tournament is acyclic if and only if it is transitive, that is, the underlying edge relation is a linear order. Given a directed graph $D = (V, A)$, we want to find the maximum size of an induced acyclic tournament in D, denoted by $mat(D)$. This study is motivated by the following reasons: firstly, this forms a toy problem, so to say, for studying some of the basic techniques used to prove concentration of random variables. An understanding of these techniques and their strong and weak points, developed in this chapter, shall help us in tackling the harder problem that lies ahead - that of determining the concentration of $mas(D)$ - the size of the largest induced acyclic subgraph of a digraph D.

Secondly, the algorithmic version of this problem is important for several applications in theoretical computer science. Although known to be NP-hard in the worst case, and even up to any reasonable approximation, studying $mat(D)$ for random digraphs gives an understanding of the *average case* hardness of this problem. This is useful in many instances, for designing algorithms that are much more efficient on average-case inputs.

Both the above points are discussed in detail later. In this chapter, we study the problem of determining $mat(D)$ for random digraphs both from an analytical and an algorithmic point of view.

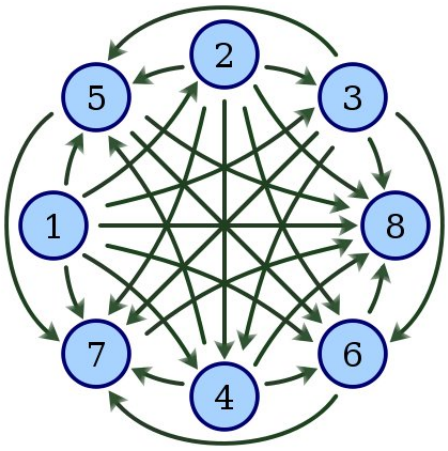

Figure 4.1: An acyclic tournament on 8 vertices

Note: Throughout this chapter, r denotes p^{-1}. $\Re^+$ denotes the positive real numbers, and standard "O" notation is used, including $O(f(n))$ and $\Omega(f(n))$ to denote the set of functions upper and lower bounded respectively, by a constant times $f(n)$, $o(f(n))$ to denote the set of functions growing asymptotically slower (with n) than $f(n)$, and $\omega(f(n))$ to denote the set of functions asymptotically faster than $f(n)$.

We study the $\mathcal{D}(n,p)$ model of a simple random digraph (see Chapter 3) introduced in [63]. In what follows, $p = p(n) \leq 0.5$ is a real number.

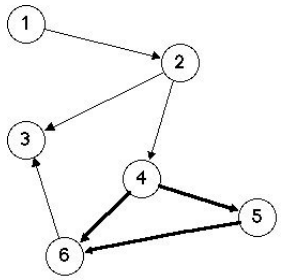

Figure 4.2: In Figure 3.2 the largest induced acyclic tournament is $\{4,5,6\}$

4.1.1 Analytical aspects

Subramanian [63] first studied the related problem of determining $mas(D)$, the size of a largest induced acyclic subgraph in a random digraph $D = (V, E)$, and later Spencer and Subramanian [61] obtained the following result.

Theorem 4.1.1 *[61] Let $D \in \mathcal{D}(n,p)$ and $w = np$. There is a sufficiently large constant W such that : If p satisfies $w \geq W$, then, a.a.s,*

$$mas(D) \quad \in \quad \left[\left(\frac{2}{\ln q}\right)(\ln w - \ln\ln w - O(1)), \left(\frac{2}{\ln q}\right)(\ln w + 3e)\right]$$

where $q = (1-p)^{-1}$.

Thus, with high probability $(1-o(1))$, $mas(D)$ lies in an integer band of width $O\left(\frac{\ln\ln w}{\ln q}\right)$. But this upper bound on width is asymptotically $\Theta(r \ln\ln w)$, and hence can become large for small values of p. However, if we focus on more restricted subgraphs, namely, induced acyclic tournaments, then the optimum size can be shown (see Theorem 4.1.2 below) to be one of two consecutive values a.a.s. In other words, we obtain a 2-point concentration for $mat(D)$. This is one of our main results in this chapter.

Theorem 4.1.2 *Let $\{\mathcal{D}(n,p) : p = p(n), n \geq 1\}$ be an infinite sequence of probability distributions. Let $w = w(n)$ be any sufficiently slowly growing function of n (say, any w with $w < \sqrt{n}$ always) such that $w \to \infty$ as $n \to \infty$. Let $D \in \mathcal{D}(n,p)$. Then, a.a.s., the following holds:*

(i) Suppose $p \geq 1/n$. Define

$$d = 2\log_r n + 1 = \frac{2(\ln n)}{\ln r} + 1; \quad b^* \; = \; \lfloor d - 1/2 \rfloor = \lfloor 2(\log_r n) + 0.5 \rfloor.$$

Then, $mat(D)$ is either b^* or $b^* + 1$.

(ii) $mat(D) \in \{2,3\}$ if $1/(wn) \leq p < 1/n$.

(iii) $mat(D) = 2$ if $wn^{-2} \leq p < 1/(wn)$.

(iv) $mat(D) \leq 2$ if $1/(wn^2) \leq p < wn^{-2}$.

(v) $mat(D) = 1$ if $p < (wn^2)^{-1}$.

Similar two-point concentration results are known for maximum clique size $\omega(G)$ of a random undirected graph $G \in \mathcal{G}(n,p)$ for $p \leq 0.5$ (see [17, 14, 40]). The chromatic number $\chi(G)$ is another parameter which has been shown to be 2-point concentrated for sparse random undirected graphs (see [47, 7, 1]). However, unlike the case of $mat(D)$, there is no explicit closed form expression for $\omega(G)$. With some assumptions about $p = p(n)$, one can also prove (proof presented in Section 3) a stronger one-point concentration (Theorem 4.1.3 below) on $mat(D)$ for all large values of n.

Theorem 4.1.3 *Let $\mathcal{D}(n,p)$, d be as defined in Theorem 4.1.2. Let $w = w(n)$ be any function so that as $n \to \infty$, $w(n) \leq 0.5(\ln n)$ and $w \to \infty$. If $p \geq 1/n$ is such that d satisfies $\frac{w}{\ln n} \leq \lceil d \rceil - d \leq 1 - \frac{w}{\ln n}$ for all large values of n, then a.a.s. $mat(D) = \lfloor d \rfloor$.*

As a consequence, we also obtain the following concentration result. For any choice of $p = p(n)$ and any given definition of $f(n) = 1 - o(1)$, let $N_{f,p}$ denote the set of natural numbers n such that $mat(D)$ takes a specific value with probability at least $f(n)$. Let us call $p = p(n)$ a *constant function* if, for some $a \in [0, 0.5]$, $p(n) = a$ for every n. Then,

Corollary 4.1.4 *For every constant function $p = p(n)$, there exists a function $f = f(n) = 1 - o(1)$ such that the set $N_{f,p}$ is a subset of natural numbers having density 1.*

Our proof (presented in Subsection 4.4.1) of the above corollary is direct and does not take recourse to the Borel-Cantelli Lemma which is applied in similar one-point concentration proofs. Perhaps, similar direct proofs are possible in other cases where the Borel-Cantelli lemma has been used, as for example, in proving a one-point concentration result for the clique number $\omega(G)$ of random undirected graphs.

It is interesting to note that the bounds on $\lceil d \rceil - d$ assumed in Theorem 4.1.3 are essentially tight. We give an example of a function $p = p(n)$ such that the assumptions in Theorem 4.1.3 do not hold, and prove that $mat(D)$ is **not** 1-point concentrated:

Theorem 4.1.5 *For any fixed $j \in \mathbb{Z}^+$ (with $j \geq 3$) and $c \in \mathcal{R}^+$, let $D \in \mathcal{D}(n,p)$, $p = n^{-2/(j-1+\frac{c}{\ln n})}$. Then, for every sufficiently large n, each of the two events (i) $mat(D) = j-1$ and (ii) $mat(D) = j$, occurs with probability lower bounded by a positive constant.*

The proof of this theorem is provided in Subsection 4.4.2 and is based on applying Lovász Local Lemma and Paley-Zygmund Inequality.

We also establish a threshold (on p) for the existence of induced acyclic tournaments of size i. For every fixed i, the threshold is coarse and is a sharp one if $i = i(n)$ varies with n and is any suitably growing function which goes to ∞ as $n \to \infty$. These are stated in the following theorem whose proof is presented in Subsection 4.4.3.

Theorem 4.1.6 *For every (positive) integer valued function $i = i(n)$ such that $i(n) \in \{1, \ldots, \lfloor 2\log_2 n \rfloor\}$ for every n, there exist functions $p_i = p_i(n) \in [0,1]$ and $q_i = q_i(n) \in [0,1]$ such that : If $D \in \mathcal{D}(n,p)$ with $1/n \leq p = p(n) \leq 0.5$, then a.a.s.*

(a) if $p \geq p_i + q_i$ then $mat(D) \geq i$, while

(b) if $p \leq p_i - q_i$ then $mat(D) < i$.

Also, if $i = i(n) \to \infty$ is a growing function of n, then the threshold $p_i(n)$ is a sharp threshold in the sense that $q_i(n) = o(p_i(n))$.

The proof of Theorem 4.1.7 (see Subsection 4.1.2) suggests a correspondence between cliques in arbitrary undirected graphs and acyclic tournaments in specific orientations of these graphs. A quantitative statement of this relationship can be obtained when random graphs are compared to random digraphs. See Lemma 4.2.1 of Subsection 4.2.1 for the statement and its proof.

Outline : The presentation of the results is organized as follows: Firstly, in Section 4.2, we discuss some connections between the acyclic tournament number $mat(D)$ and the clique number $\omega(G)$. In Section 4.3, we provide the proof of Theorem 4.1.2. In Section 4.4, Theorem 4.1.3, Corollary 4.1.4 and Theorem 4.1.3 are proved. The proofs of the Theorems 4.1.2 and 4.1.3 are based on the Second Moment Method. We also present the proof of Theorem 4.1.6 in Section 4.4. 4.1.2.

4.1.2 Algorithmic aspects

By MAT(D, k), we denote the following computational problem : Given a simple directed graph $D = (V, A)$ and k, determine if $mat(D) \geq k$. By MAT(D), we denote its optimization version. That is, given D, find an induced acyclic tournament of maximum size. MAT(D, k) is known to be NP-complete [36], even if D is restricted to be a tournament [59]. Also, MAT(D) is known to be hard to approximate ([49]) when the input is an arbitrary digraph: For some $\epsilon > 0$, a polynomial-time approximation algorithm with an approximation ratio of $O(n^{\epsilon})$ is not possible unless $P = NP$.

Below we strengthen both of these results as follows. We show that MAT(D) is hard and inapproximable even when D is restricted to be acyclic (a dag), as shown in Theorem 4.1.7.

Theorem 4.1.7 *MAT(D, k) is NP-complete when D is restricted to be acyclic. Also, for every $\epsilon \geq (\log n)^{-\gamma}$, for some constant $0 < \gamma < 1$, the optimization problem MAT(D) is not efficiently approximable with an approximation ratio of $n^{1-\epsilon}$ unless $NP \subseteq ZPTIME(2^{(\log n)^{O(1)}})$, even if D is restricted to be acyclic.*

Proof : We reduce the NP-complete Maximum Clique problem MC(G, k) to the MAT(D, k) problem as follows. Given an instance $(G = (V, E), k)$ of the first problem, compute an

instance $f(G) = (G' = (V, A), k)$ in polynomial time where

$$A = \{(u, v) : uv \in E, u < v\}.$$

Clearly, G' is a dag and it is easy to see that a set $V' \subseteq V$ induces a clique in G if and only if V' induces an acyclic tournament in G'. This establishes that MAT(D, k) is NP-hard even if D is restricted to be a dag.
The inapproximability of MAT(D) follows from the following observation. Note that the reduction $G \to f(G)$ is an L-reduction in the sense of [54], since $|f(G)| = |G|$ and $\omega(G) = mat(G')$. Hence, any inapproximability result on maximum clique in undirected graphs (for example [38, 42]), implies a similar inapproximability for the MAT(D) problem. ■

However, the *average* case version of the problem - finding $mat(D)$ for a random digraph D - offers some hope. In this version, we seek to design efficient algorithms for computing an optimal solution which succeed a.a.s. over a random digraph. We use the model $\mathcal{D}(n, p)$ defined before for studying random digraphs.

We show (see Theorem 4.5.1) that a.a.s. *every* maximal induced acyclic tournament is of size which is at least nearly half of the optimal size. Hence any greedy heuristic obtains a solution whose approximation ratio is a.a.s. 2. This is similar to the case of cliques in undirected random graphs (see e.g. [14]).

We also study another heuristic which combines greedy and brute-force approaches as follows. We first apply the greedy heuristic to get a partial solution whose size is nearly $\log_r n - c\sqrt{\log_r n}$ for some arbitrary constant c. Amongst the remaining vertices, let C be the set of vertices such that each vertex in C can be individually and "safely" added to the partial solution. Then, in the subgraph induced by C we find an optimal solution by brute-force and combine it with the partial solution. It is shown in Theorem 4.6.1 that this modified approach produces a solution whose size is at least $\log_r n + c\sqrt{\log_r n}$. This results in an additive improvement of $\Theta(\sqrt{\log_r n})$ over the simple greedy approach. The improvement is due to the fact we stop using greedy heuristic at a point where it is possible to apply brute-force efficiently. This approach is similar to (and was motivated by) the one used in [44] for finding large independent sets in $\mathcal{G}(n, 1/2)$.

As a consequence, we see that the problem of finding an optimal induced acyclic tournament can be approximated within a ratio of $2 - o(1)$ a.a.s. for random digraphs. This is in sharp contrast to the worst-case version where, by Theorem 4.1.7, it is very unlikely to be approximable even with a large multiplicative ratio.

Outline : The presentation of the algorithmic results is as follows. Theorem 4.5.1 is stated and proved in Section 4.5. Theorem 4.6.1 is stated and proved in Section 4.6.

4.1.3 Non-simple random digraphs

Each of the concentration and algorithmic results mentioned before also carry over (with some slight changes) to a related random model $\mathcal{D}_2(n,p)$ where we allow 2-cycles to be present and each of the potential arcs is chosen independently. These are presented in Section 6. In Section 7, we present some observations on the concentration of the maximum size of an induced tournament (not necessarily acyclic) for the two models of random directed graphs. Finally, in Section 8, we conclude with a summary and some open problems.

Notations : Throughout, we use standard notation. $\Re^+$ denotes the positive real numbers, $\mathcal{N}$ denotes the set of natural numbers. We often use the short notations $p = p(n)$, $w = w(n)$ to denote functions (real or integer valued) over $\mathcal{N}$. We also use standard notations like $O(.)$, $\Omega(.)$, $o(.)$ and $\omega(.)$ with usual meanings.

4.2 $mat(D)$ versus $\omega(G)$

In this section, we explore some connections between the maximum acyclic tournament number $mat(D)$ and the clique number $\omega(G)$. A few such connections, such as the reduction of $CLIQUE(G,k)$ to $MAT(D,k)$ have already been shown previously. The first additional result is the connection between the probability distribution of $mat(D), D \in \mathcal{D}(n,p)$ and $\omega(G), G \in \mathcal{G}(n,p)$, for some $0 < p = p(n) \leq 1/2$.

4.2.1 Comparision of Probability distributions

The following lemma relates the probabilities in the two models $\mathcal{D}(n,p)$ and $\mathcal{G}(n,p)$ for having, respectively, acyclic tournament number and clique number equal to a given value. Its proof is similar to the proof of an analogous relationship involving $mas(D)$ and $\alpha(G)$ (maximum size of an independent set in G) established in [63].

Lemma 4.2.1 *For any positive integer b, for a random digraph $D \in \mathcal{D}(n,p)$,*

$$Pr[mat(D) \geq b] \quad \geq \quad Pr[\omega(G) \geq b].$$

where $G \in \mathcal{G}(n,p)$.

Proof : Given a linear ordering σ of vertices of D and a subset A of size b, we say that $D[A]$ is consistent with σ if for every $\sigma_i, \sigma_j \in A$ with $i < j$, $D[A]$ has the arc (σ_i, σ_j).

Let τ denote an arbitrary but fixed ordering of V. Once we fix τ, the spanning subgraph of D formed by arcs of the form $(\tau(i), \tau(j))$ $(i < j)$ is having the same distribution as $\mathcal{G}(n,p)$.

Hence, for any A, the event of $D[A]$ being consistent with τ is equivalent to the event of A inducing a clique in $\mathcal{G}(n,p)$. Hence,

$$\begin{aligned}
\mathbf{Pr}(\ mat(D) \geq b\ ) &= \mathbf{Pr}(\ \exists A,\ |A| = b,\ D[A]\ \text{ is an acyclic tournament }\ ) \\
&= \mathbf{Pr}(\ \exists A,\ |A| = b,\ \exists \sigma,\ D[A]\ \text{ is consistent with }\ \sigma\ ) \\
&= \mathbf{Pr}(\ \exists \sigma,\ \exists A,\ |A| = b,\ D[A]\ \text{ is consistent with }\ \sigma\ ) \\
&\geq \mathbf{Pr}(\ \exists A,\ |A| = b,\ D[A]\ \text{ is consistent with }\ \tau\ ) \\
&= \mathbf{Pr}(\ \omega(G) \geq b\ ).
\end{aligned}$$

Hence it is natural that we have a bigger lower bound for $mat(D)$ than we have for $\omega(G)$. ■

Note : Recall that we first draw an undirected $G \in \mathcal{G}(n, 2p)$ and then choose uniformly randomly an orientation of $E(G)$. Hence, for any fixed $A \subseteq V$ of size b with $b = \omega(1)$,

$$\mathbf{Pr}(\ D[A] \text{ is an acyclic tournament } \mid G[A] \text{ induces a clique }\ ) \ = \ \frac{b!}{2^{\binom{b}{2}}} \ = \ o(1).$$

However, there are so many cliques of size b in G that one of them manages to induce an acyclic tournament.

4.2.2 Lower Bounds

Another intriguing connection between $mat(D)$ and $\omega(G)$ stems from the degree-sequence based lower bound expressions that can be obtained for them. In [9], Alon and Spencer gave a probabilistic proof of the following degree-sequence based lower bound on the independence number $\alpha(G)$ of an n-vertex graph G:

$$\alpha(G) \geq \sum_{v \in V} \frac{1}{d(v)+1}$$

where $d(v)$ is the degree of the vertex $v \in V[G]$. Applying this bound to the complement graph $\bar{G}$, one obtains a lower bound on the clique number of the graph G:

$$\omega(G) \geq \sum_{v \in V} \frac{1}{n - d(v)} \tag{4.1}$$

A similar expression can be obtained for acyclic tournaments in digraphs:

Theorem 4.2.2 *Given a simple digraph $D = (V, E)$ on n vertices, with $d^+(v), d^-(v)$ denoting the out-degree and in-degree, respectively, of the vertex v, the maximum acyclic tournament of D is of size*

$$mat(D) \geq \max\left\{\sum_{v\in V}\frac{1}{n-d^-(v)}, \sum_{v\in V}\frac{1}{n-d^+(v)}\right\} \tag{4.2}$$

Proof Choose uniformly at random a linear ordering '$<$' over V and define two sets I_i and I_o as follows :

(i) I_i is the set of those $v \in V$ satisfying : for every u such that $u < v$, u is an in-neighbor of v.

(ii) I_o is the set of those $v \in V$ satisfying : for every u such that $u < v$, u is an out-neighbor of v.

Writing each of $|I_i|$ and $|I_o|$ as a sum of indicator variables (one for each v) and applying Linearity of Expectation, it can be easily verified that $\mu_i = E[|I_i|] = \sum_{v\in V}\frac{1}{n-d^-(v)}$ and also that $\mu_o = E[|I_o|] = \sum_{v\in V}\frac{1}{n-d^+(v)}$. Also, $E[\max\{|I_i|, |I_o|\}] \geq \max\{\mu_i, \mu_o\}$. Hence there exists a linear ordering for which either I_i or I_o is of size at least $\max\{\mu_i, \mu_o\}$. In any case, each of I_i and I_o always induces an acyclic tournament. This proves the theorem. ∎

4.3 Analysis of $\mathcal{D}(n, p)$

Let U be any fixed subset of V of size b. The following two easy-to-verify claims play a role in the analysis.

Claim 4.3.1 *A directed acyclic graph $H = (U, A)$ has at most one (directed) hamilton path.*

Proof : Order the vertices of U along a hamilton path P (if any exists) of H. An arc $(u, v) \in A$ is a forward arc if u comes before v in P and is a backward arc otherwise. Since H is acyclic, any arc $(v, u) \in A$ must be a forward arc, since otherwise the segment of P from u to v along with (v, u) forms a cycle in H.

Now if there is another hamilton path Q in H, $Q \neq P$, then walking along P, consider the first vertex a where Q differs from P. Then in the path Q, a is visited after some vertex a' that comes after a in P. But this implies that (a', a) is a backward arc in H contradicting the observation earlier that H has no backward arc. ∎

Claim 4.3.2 $\mathbf{Pr}[D[U]$ *is an acyclic tournament* $] = b!\, p^{\binom{b}{2}}$

Proof : By Claim 4.3.1, any acyclic tournament on U has exactly one hamilton path P which is also a linear ordering of U. Also, there are exactly $b!$ choices for P. For every fixed linear ordering σ of U, let $\mathcal{E}(U,\sigma)$ denote the event that $D[U]$ is an acyclic tournament with the hamilton path σ. This event happens whenever each of the $\binom{b}{2}$ forward arcs with respect to σ are present in $D[U]$, which is a collection of $\binom{b}{2}$ identical and independent events. Hence, we have

$$\mathbf{Pr}(\mathcal{E}(U,\sigma)) \;=\; p^{\binom{b}{2}}.$$

Hence, $\mathbf{Pr}[D[U] \text{ is an acyclic tournament }] \;=\; \sum_{\sigma} \mathbf{Pr}(\mathcal{E}(u,\sigma)) \;=\; b!\; p^{\binom{b}{2}}$. ∎

Before we proceed further, we introduce some notations which play an important role in the analysis. Define $\delta = \lceil d \rceil - d$. Then, it follows that

$$b^* = \begin{cases} d - 2 + \delta & \text{if } \; \delta > 1/2; \\ d - 1 + \delta & \text{if } \; \delta \leq 1/2. \end{cases}$$

For a given b, let $m = \binom{n}{b}$ and let $\{A_1, \ldots, A_m\}$ denote the set of all b-sized subsets of V. For $i \in [m]$, let X_i denote the random variable that indicates whether $D[A_i]$ induces an acyclic tournament or not. Let $X(b) = X(n,b)$ denote the number of induced acyclic tournaments of size b in D. Since there are $\binom{n}{b}$ sets of size b, it follows by Linearity of Expectation that

$$E[X(n,b)] = \sum_i E[X_i] = \binom{n}{b}\; b!\; p^{\binom{b}{2}}.$$

We are only interested in the behavior of $E[X(n,b)]$ for $b \in [1, b^*+2]$. From the definition of b^*, it follows that $b^*+2 \leq \lceil d \rceil + 1 \leq \frac{2(\ln n)}{\ln r} + 3 \leq 3(\ln n)$ for sufficiently large n since $p \leq 1/2$. As a result, we have

$$[1 - o(1)] \cdot f(n,p,b)^b \;\leq\; E[X(n,b)] \;\leq\; f(n,p,b)^b \quad \ldots\ldots \quad (A)$$

$$\text{where} \quad f : (\Re^+)^3 \to \Re^+ \text{ such that } \quad f(n,p,b) \;=\; n\, p^{(b-1)/2}.$$

Setting $f(n,p,b) = 1$ and solving for b, we see that

$$f(n,p,b) > 1 \;\text{ if }\; b < d; \quad f(n,p,d) = 1; \quad f(n,p,b) < 1 \;\text{ if }\; b > d.$$

4.3.1 Proof of $mat(D) \leq b^* + 1$

First, we focus on proving the upper bound of Theorem 4.1.2. This is done by proving that

$$\mathbf{Pr}(X(b^*+2) > 0) \;\leq\; E[X(b^*+2)] \;=\; o(1).$$

Recall that b^* can be expressed in terms of d and δ in two different ways depending on the value of δ.

Case I: $\delta > 1/2$

$$\begin{aligned} E[X(b^*+2)] &= E[X(d+\delta)] \\ &\leq f(n,p,d+\delta)^{d+\delta} \\ &= \left(f(n,p,d)\cdot p^{\delta/2}\right)^{d+\delta} \\ &= p^{\delta(d+\delta)/2} \;=\; p^{\delta d/2}\cdot p^{\delta^2/2} \\ &= n^{-\delta}\cdot p^{\delta(1+\delta)/2} \;\leq\; n^{-\delta} \text{ since } p \leq 1 \text{ and } \delta \geq 0 \\ &\leq n^{-1/2} \;=\; o(1). \end{aligned}$$

Case II: $\delta \leq 1/2$

$$\begin{aligned} E[X(b^*+2)] &= E[X(d+1+\delta)] \\ &\leq f(n,p,d+1+\delta)^{d+1+\delta} \\ &= \left(f(n,p,d)\cdot p^{(1+\delta)/2}\right)^{d+1+\delta} \\ &= p^{(1+\delta)(d+1+\delta)/2} \;=\; p^{(1+\delta)d/2}\cdot p^{(1+\delta)^2/2} \\ &= n^{-(1+\delta)}\cdot p^{(1+\delta)(2+\delta)/2} \;\leq\; n^{-(1+\delta)} \text{ since } p \leq 1 \text{ and } \delta \geq 0 \\ &\leq n^{-1} \;=\; o(1). \end{aligned}$$

This establishes the upper bound.

4.3.2 Proof of $mat(D) \geq b^*$

Next, we focus on proving the lower bound of Theorem 4.1.2. For this, we first show that $E[X(b^*)] \to \infty$ as $n \to \infty$.

Case I: $\delta > 1/2$

$$\begin{aligned} E[X(b^*)] &= E[X(d-2+\delta)] \\ &\geq [1-o(1)]\cdot f(n,p,d-2+\delta)^{d-2+\delta} \\ &= [1-o(1)]\cdot \left(f(n,p,d)\cdot p^{(-2+\delta)/2}\right)^{d-2+\delta} \\ &= [1-o(1)]\cdot p^{(-2+\delta)(d-2+\delta)/2} \;=\; p^{(-2+\delta)d/2}\cdot p^{(-2+\delta)^2/2} \\ &= [1-o(1)]\cdot n^{2-\delta}\cdot p^{(2-\delta)(1-\delta)/2} \\ &\geq [1-o(1)]\cdot n^{2-\delta}\cdot p^{3/8} \quad \text{since } p \leq 1 \text{ and } \delta > 1/2 \\ &\geq n^{1/2} \;\to\; \infty \quad \text{as} \quad n \;\to\; \infty. \end{aligned}$$

Case II: $\delta \leq 1/2$

$$\begin{aligned}
E[X(b^*)] &= E[X(d-1+\delta)] \\
&\geq [1-o(1)] \cdot f(n,p,d-1+\delta)^{d-1+\delta} \\
&= [1-o(1)] \cdot \left(f(n,p,d) \cdot p^{(-1+\delta)/2}\right)^{d-1+\delta} \\
&= [1-o(1)] \cdot p^{(-1+\delta)(d-1+\delta)/2} \;=\; p^{(-1+\delta)d/2} \cdot p^{(-1+\delta)^2/2} \\
&= [1-o(1)] \cdot n^{1-\delta} \cdot p^{(-1+\delta)\delta/2} \\
&\geq [1-o(1)] \cdot n^{1-\delta} \text{ since } p \leq 1 \text{ and } \delta \leq 1/2 \\
&\geq n^{1/2} \;\to\; \infty \quad \text{as} \quad n \;\to\; \infty.
\end{aligned}$$

For the sake of notational simplicity, we use X to denote $X(b^*)$ and use b to denote b^* for the rest of this section. Now, we need to show that $X > 0$ with high probability. We use the well-known Second Moment Method to establish this. Let $Var(X)$ denote the variance of X.

Recall that X_i denotes the indicator random variable for the i-th b-size subset of V. Using standard arguments (see [9]), it can be seen that

$$Var(X) \;\leq\; E[X] + \sum_{i \neq j} COV(X_i, X_j) \tag{4.3}$$

where the second sum is over ordered pairs and $COV(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j)$ is the covariance between X_i and X_j. Note that X_i and X_j are independent whenever $|A_i \cap A_j| \leq 1$ and in that case $COV(X_i, X_j) = 0$. Otherwise, with $|A_i \cap A_j| = l$, we have

$$\begin{aligned}
COV(X_i, X_j) &\leq E(X_i X_j) \;=\; E(X_i)E(X_j | X_i = 1) \\
&= b! p^{\binom{b}{2}} \cdot (b!/l!) \cdot p^{\binom{b}{2} - \binom{l}{2}}
\end{aligned} \tag{4.4}$$

where the last equality follows from Claim 4.3.2. Also, for any fixed i, the number of b-sized subsets A_j such that $|A_i \cap A_j| = l$ is exactly $\binom{b}{l}\binom{n-b}{b-l}$. As a result,

$$\begin{aligned}
\sum_{i \neq j} COV(X_i, X_j) &= \sum_i \sum_{j \,:\, 2 \leq |A_i \cap A_j| \leq b-1} COV(X_i, X_j) \\
&\leq \sum_i b! p^{\binom{b}{2}} \cdot \left(\sum_{2 \leq l \leq b-1} \binom{b}{l} \binom{n-b}{b-l} \left(\frac{b!}{l!}\right) \cdot p^{\binom{b}{2} - \binom{l}{2}} \right) \\
&= E[X] \cdot \left(\sum_{2 \leq l \leq b-1} \binom{b}{l} \binom{n-b}{b-l} \left(\frac{b!}{l!}\right) \cdot p^{\binom{b}{2} - \binom{l}{2}} \right)
\end{aligned}$$

$$
\begin{aligned}
&= E[X]^2 \cdot \left(\sum_{2 \le l \le b-1} \frac{(b)_l}{(l!)^2} \binom{n-b}{b-l} \binom{n}{b}^{-1} \cdot p^{-\binom{l}{2}} \right) \\
&= E[X]^2 \cdot M \qquad (4.5)
\end{aligned}
$$

where $M = M(n,p,b)$ is as defined above. Applying Chebyshev's Inequality and (4.3), it follows that

$$
Pr[X=0] \;\le\; Var(X)(E[X])^{-2} \;\le\; \left(E[X] + \sum_{i \neq j} COV(X_i, X_j) \right) (E[X])^{-2} \qquad (4.6)
$$

Combining (4.6) and (4.5), we notice that

$$
\mathbf{Pr}(X=0) \;\le\; (E[X])^{-1} + M \;=\; o(1) \qquad (4.7)
$$

provided $M = M(n,b) = o(1)$ since it has already been shown that $E[X] \to \infty$. Thus, we only need to show that $M = o(1)$ to complete the arguments.

Now, we focus on showing that $M = o(1)$. Notice that

$$
\begin{aligned}
M &= \sum_{2 \le l \le b-1} \frac{(b)_l}{(l!)^2} \cdot \binom{n-b}{b-l} \cdot \binom{n}{b}^{-1} \cdot p^{-\binom{l}{2}} \\
&\le \sum_{2 \le l \le b-1} \binom{b}{l}^2 \cdot \frac{1}{(n)_l} \cdot p^{-\binom{l}{2}} \;=\; \sum_{2 \le l \le b-1} \binom{b}{l}^2 \cdot \frac{1+o(1)}{n^l} \cdot p^{-\binom{l}{2}} \\
&= (1+o(1)) \sum_{2 \le l \le b-1} \binom{b}{l}^2 p^{l(d-l)/2} \;=\; (1+o(1)) \sum_{2 \le l \le b-1} F_l
\end{aligned}
$$

where the last-but-one equality follows using $f(n,p,d) = 1$.
Let t_l be the ratio between successive terms: $t_l = F_{l+1}/F_l$. Now take the ratio of ratios: $s_l = t_{l+1}/t_l$.

$$
\begin{aligned}
t_l &= F_{l+1}/F_l = \frac{\binom{b}{l+1}^2 p^{(l+1)(d-l-1)/2}}{\binom{b}{l}^2 p^{l(d-l)/2}} \;=\; \left(\frac{b-l}{l+1} \right)^2 p^{-l+(d-1)/2} \\
s_l &= \left(\left(\frac{b-l-1}{b-l} \right) \left(\frac{l+1}{l+2} \right) \right)^2 p^{-1}
\end{aligned}
$$

First we state the following easy-to-prove fact regarding any sequence of positive real numbers.

Fact 4.3.3 *For a sequence of positive real numbers $a_1, \ldots, a_n$, if $s_i = a_{i+2}a_i/a_{i+1}^2 \ge 1$ for all $1 \le i \le n-2$, then for all $i \in [n]$, we have $a_i \le \max\{a_1, a_n\}$.*

Proof Consider the ratio $b_i := a_{i+1}/a_i$. Since $s_i = b_{i+1}/b_i \geq 1$ we get that for all $i \in [n-1]$, $b_{i+1} \geq b_i$. There are only two possibilities for b_1:

- $b_1 \geq 1$. Then $b_i \geq 1$ for all $i \in [n-]$, implying that $a_i \leq a_n$ for all $i \in [n]$. Hence we are done.
- $b_1 < 1$. In this case there is a unique $k : 1 < k \leq n$ such that $b_k \geq 1$, but for all $j < k$, $b_j < 1$. Then for all $j < k$, we have $a_j < a_1$, whereas for all $j \geq k$ the situation reduces to the first case, and again we get $a_j \leq a_n$.

Thus in all the above cases, either $a_j \leq a_1$ or $a_j \leq a_n$. Therefore for all $i \in [n]$, $a_i \leq (a_1 + a_n)$.

Claim 4.3.4 *(i) If $p \leq 1/4$, then $s_l \geq 1$ for every b with $2 \leq l \leq b-3$. (ii) If $p > 1/4$, then $s_l \geq 1$ for every b with $2 \leq l \leq b-4$ and also $t_{b-2} > 1$.*

From the above (Fact 4.3.3 and Claim 4.3.4), the proof of Theorem 4.1.2 follows easily, as we get that for all $l : 2 \leq l \leq b-1$, for all $p \geq 1/n$, $F_l \leq \max\{F_2, F_{b-1}\}$. Now $F_2 = \binom{2}{2}^2 p^{2.(d-2)/2} = p^{2\left(\frac{d-1}{2} - \frac{1}{2}\right)} = \frac{p^{-1}}{n^2} = O(1/n)$ and $F_{b-1} = \binom{b}{b-1}^2 p^{(b-1)(d-b+1)/2} \leq b^2 \left(p^{(b-1)/2}\right) = b^2 \left(\frac{p^{(b-d)/2}}{n}\right) \leq b^2 \left(\frac{r^{(d-b)/2}}{n}\right) \leq b^2 \left(\frac{r^{3/4}}{n}\right) = O\left(\frac{(\ln n)^2}{n^{1/4}}\right)$. Therefore, $M = (1 + o(1)) \cdot \sum_{l=2}^{b-1} F_l = O\left(\frac{(\ln n)^3}{n^{1/4}}\right) = o(1)$.

Proof of Claim 4.3.4 Case (i) : Assume that $p \leq 1/4$ and $2 \leq l \leq b-3$. Then, we have $(b-l-1)/(b-l) \geq 2/3$ and $(l+1)/(l+2) \geq 3/4$. This implies that $s_l \geq p^{-1}/4 \geq 1$.
Case (ii) : Assume that $p > 1/4$ and $l \leq b-4$. It can be verified that the square term in s_l is at least $1/2$ and $p^{-1} \geq 2$, so $s_l \geq 1$. Now $t_{b-2} = (2/(b-1))^2 p^{-(b-2)+(d-1)/2} \geq (4r^b p^2)/(nb^2) \geq \frac{4np^{2.5}}{b^2} \to \infty$, using our assumption that $p > 1/4$. ∎

We have thus completely established that $M = o(1)$ for all $p \geq 1/n$, thereby establishing that $\mathbf{Pr}(X = 0) = o(1)$. Hence, a.a.s., $mat(D) \in \{b^*, b^* + 1\}$ for the stated range of p. The remaining parts of Theorem 4.1.2 are straightforward to derive and are given below for the sake of completeness.

For $1/wn \leq p < 1/n$,

$$E[X(n,4)] \;=\; \binom{n}{4} \cdot 4! \cdot p^{\binom{4}{2}} \;\leq\; n^4 p^6 \;\leq\; (1/n^2) \;=\; o(1).$$

Now, an acyclic tournament of size 2 is simply an edge which a.a.s. exists since :

$$Pr[mat(D) < 2] \;=\; Pr[D \text{ is the empty graph}] \;=\; (1-2p)^{\binom{n}{2}} \;\leq\; e^{-n(n-1)p} \;=\; o(1),$$

since $p \geq 1/wn \geq w/n^2$. Hence, when $1/wn \leq p \leq 1/n$, $mat(D) \in \{2, 3\}$, a.a.s.

For $wn^{-2} \leq p < 1/wn$,

$$E[X(n,3)] \;=\; \binom{n}{3} \cdot 3! \cdot p^{\binom{3}{2}} \;\leq\; n^3 p^3 \;=\; o(1) \text{ since } np = o(1).$$

The proof for $mat(D) \geq 2$ is the same as in the previous case, since $n^2 p = \omega(1)$, and hence, at least one arc will exist, a.a.s. So when $w/n^2 \leq p \leq 1/wn$, $mat(D) = 2$, a.a.s.

For $(wn^2)^{-1} \leq p \leq w/n^2$, $E[X(n,3)] = o(1)$, as in the previous case, and so $mat(D) = 1$ or 2, a.a.s.. When $p < (wn^2)^{-1}$, $mat(D) = 1$ since D a.a.s. has no directed edge.

This completes the proof of Theorem 4.1.2. ∎

4.4 One-point Concentration and threshold results

Recall the definition of d, δ from the proof of Theorem 4.1.2. The proof of Theorem 4.1.3 proceeds by considering the following 2 cases:

Case $0 < w/\ln n \leq \delta \leq 1/2$ **:** In this case, $\lfloor d \rfloor = b^*$. From theorem 4.1.2, it only remains to show that $\mathbf{Pr}[mat(D) \geq b^* + 1] \to 0$ as $n \to \infty$. We again use the first moment method to show that $E[X(b^* + 1)] = o(1)$.

By our assumption about p, $\delta \leq 1/2$. Hence, by definition, $b^* + 1 = d + \delta$. Thus,

$$\begin{aligned} E[X(b^*+1)] &\leq f(n,p,d+\delta)^{d+\delta} \\ &= (p^{\delta/2})^{d+\delta} \;=\; p^{\delta(d+\delta)/2} \\ &= n^{-\delta} \cdot p^{\delta(1+\delta)/2} \\ &\leq n^{-\delta} \;\leq\; n^{-w/\ln n} \\ &= e^{-w} \;\to\; 0 \quad \text{as } n \to \infty \end{aligned}$$

Case $1/2 < \delta \leq 1 - w/\ln n < 1$ **:** Here, $\lfloor d \rfloor = b^* + 1$. The proof proceeds by verifying that $Pr[X(b^*+1) = 0] = o(1)$, and hence, $mat(D) \geq b^* + 1$ a.a.s. Together with the upper bound on $mat(D)$ when $p \geq 1/n$ in Theorem 4.1.2, this gives the desired result. Briefly, this can be seen as follows:
From (4.7), it suffices to show that

(i) $E[X(n, b^* + 1)] \to \infty$ as $n \to \infty$, and

(ii) $M = M(n, p, b^* + 1) = o(1)$.

To prove (i), we notice that

$$\begin{aligned} E[X(b^*+1)] &\geq [1-o(1)].f(n,p,d+\delta-1)^{d+\delta-1} \\ &= [1-o(1)].(p^{(\delta-1)/2})^{d+\delta-1} \;=\; [1-o(1)].p^{(\delta-1)(d+\delta-1)/2} \\ &= [1-o(1)].n^{1-\delta}\cdot p^{-\delta(1-\delta)/2} \\ &\geq [1-o(1)].n^{w/\ln n} \\ &= [1-o(1)].e^{w} \;\to\; \infty \quad \text{as } n\to\infty \end{aligned}$$

To prove (ii), we need to go along the proof of Theorem 4.1.2, and evaluate $M(n,p,b^*+1)$. An easy check reveals that $M(n,p,b^*+1) = M(n,p,b^*)\cdot O((\ln n)^3) + T_{b^*}$,
where

$$T_{b^*} \;=\; \binom{b^*+1}{b^*}^2 \left((n)_{b^*}p^{\binom{b^*}{2}}\right)^{-1} \;\leq\; 2\binom{b^*+1}{b^*}^2 \left(np^{(b^*-1)/2}\right)^{-b^*}$$

Now, from the proof of Theorem 4.1.2, we have that $M(n,p,b^*) = O\left(\frac{(\ln n)^3}{n^{1/4}}\right)$. Therefore $M(n,p,b^*+1) = O\left(\frac{(\ln n)^6}{n^{1/4}}\right) + T_{b^*}$.
Next, using the definition of b^* when $\delta > 1/2$, we have

$$\begin{aligned} T_{b^*} &\leq 2(b^*+1)^2(np^{(b^*-1)/2})^{-b^*} \\ &= 2(b^*+1)^2(np^{(d-3+\delta)/2})^{-b^*} \\ &= 2(b^*+1)^2(p^{-1+\delta/2})^{-b^*} \\ &= 2(b^*+1)^2 p^{b^*(1-\delta/2)} \\ &= o(1) \end{aligned}$$

Thus it is verified that both $M.O(\ln n)^3$ and T_{b^*}, and hence their sum, are $o(1)$. ∎

4.4.1 Proof of Corollary 4.1.4

Let p be fixed but arbitrary. It follows from the definitions of d and δ, that for every n, we have $n = r^{\frac{k-(1+\delta)}{2}}$ for some nonnegative integer k. Also, it follows from Theorem 4.1.3 that for every sufficiently large n, $mat(D)$ is concentrated on one value if $\frac{w}{\ln n} \leq \delta \leq 1 - \frac{w}{\ln n}$. Hence, for every such n, we must have

$$r^{\frac{k-2}{2}+\frac{w}{2\ln n}} \leq n \leq r^{\frac{k-1}{2}-\frac{w}{2\ln n}}.$$

For every $k \geq 2$, we define two values as follows.

$$m_{k,l} = \min\{n : n \geq r^{\frac{k-2}{2}+\frac{w}{2\ln n}}\};\;\; m_{k,h} = \max\{n : n \leq r^{\frac{k-1}{2}-\frac{w}{2\ln n}}\}.$$

It follows that $mat(D)$ is just one value for every *sufficiently large* $n \in R$ where $R = \bigcup_{k\geq 2} R_k$ and $R_k = \{n : m_{k,l} \leq n \leq m_{k,h}\}$. Hence it suffices to show that R is a subset of density 1 of the set $\mathcal{N}$ of positive integers. Now, $\mathcal{N} - R = \bigcup_{k\geq 3} S_k$ where $S_k = \{n \in \mathcal{N} : m_{k-1,h} < n < m_{k,l}\}$.

For every $k \geq 3$,

$$|R_k| \approx r^{\frac{k-2}{2}} \left(r^{\frac{1}{2} - \frac{w}{2\ln m_{k,h}}} - r^{\frac{w}{2\ln m_{k,l}}} \right) \text{ and } |S_k| \approx r^{\frac{k-2}{2}} \left(r^{\frac{w}{2\ln m_{k,l}}} - r^{-\frac{w}{2\ln m_{k-1,h}}} \right).$$

By choosing w suitably, we can ensure that $w/\ln n \to 0$ as $n \to \infty$. Also, $m_{k,h}$ and $m_{k,l}$ grow exponentially in k. Hence, for every sufficiently large k,

$$|R_k| \approx r^{\frac{k-2}{2}} \left(r^{\frac{1}{2}} - 1 \right) \text{ and } |S_k| = O\left(r^{\frac{k-2}{2}} \left(\frac{w(r^{k/2})}{k \ln t} \right) \right) = o\left(r^{\frac{k-2}{2}} \right).$$

Thus, for all sufficiently large k, we have $\sum_{j\leq k+1} |S_j| = O(|S_{k+1}|) = o(|R_k|)$. As a result, we have

$$\frac{|R \cap [n]|}{n} \to 1 \text{ as } n \to \infty.$$

This shows that R has density 1 as a subset of $\mathcal{N}$. This completes the proof of Corollary 4.1.4. ∎

4.4.2 Proof of Theorem 4.1.5

The proof is based on an application of Lovász Local Lemma stated below.

Lemma 4.4.1 *Let $\mathcal{A} = \{E_1, E_2, ..., E_m\}$ be a collection of events over a probability space such that each E_i is totally independent of all but the events in $\mathcal{D}_i \subseteq \mathcal{A}\backslash\{E_i\}$.*
If there exists a real sequence $\{x_i\}_{i=1}^m$, $x_i \in [0, 1)$, such that

$$\begin{aligned} \forall i \in [m],\ Pr[E_i] &\leq x_i \prod_{j:E_j\in\mathcal{D}_i} (1 - x_j), \text{ then} \\ Pr\left[\bigcap_{i=1}^m \overline{E}_i\right] &\geq \prod_{i=1}^m (1 - x_i) > 0. \end{aligned}$$

Hence, with positive probability, none of the events occur. ∎

First, notice that for the given value of p, $d = (j + c/(\ln n))$, $b := b^* = j - 1$ and $\delta = 1 - c/\ln n > 1/2$, for sufficiently large n. By Theorem 4.1.2, we know that for the given probability $p = p(n)$, $mat(D) \in \{b, b+1\}$ a.a.s. Therefore to prove Theorem 4.1.5, it suffices

to show that there exist constants $0 < c_1 \leq c_2 < 1$ such that $c_1 \leq \mathbf{Pr}(mat(D) = b+1) \leq c_2$ for all sufficiently large n. This is proved below. For various symbols like, d, δ and b^*, we use the same meanings used in the proof of Theorem 4.1.2.

Consider the expected number of acyclic tournaments of size $b+1$:

$$\begin{aligned} E[X_{b+1}] &\approx (np^{b/2})^{b+1} = \left(np^{(d+\delta-2)/2}\right)^{b+1} = (p^{(-1+\delta)/2})^{b+1} \\ &= (p^{-c/2(\ln n)})^{d-1+\delta} = (r^{c/2(\ln n)})^{d-1+\delta} = (e^{c/2(\log_r n)})^{d-1+\delta} \\ &= (e^c)^{1+(\delta/2(\log_r n))} \approx e^{c'} \end{aligned}$$

for some constant $c' > 0$. If the expectation had been a constant less than 1, a simple application of Markov's inequality would have established the upper bound on the probability.

(i) **Proof of $\mathbf{Pr}(mat(D) = b+1) \leq c_2$:**

We apply Lovász Local Lemma 4.4.1 to prove this claim. For every i, $1 \leq i \leq N := \binom{n}{b+1}$, define E_i to be the event that A_i induces an acyclic tournament, where A_i is the i-th $(b+1)$-set in some fixed ordering of all $(b+1)$-subsets of V. For every i, $\mathbf{Pr}(E_i) = q := (b+1)!p^{\binom{b+1}{2}} = o(1)$. Choose $x_i = x = 25q$ for each i. Construct the dependency graph on N events by joining E_i and E_j if $|A_i \cap A_j| \geq 2$. It can be seen that each E_i is totally independent of all other E_j's which are not adjacent to E_i. Note that the dependency graph is regular with the uniform degree of any E_i being given by $deg(E_i) = \sum_{2 \leq k \leq b} \binom{b+1}{k}\binom{n-b-1}{b+1-k}$. It is easy to see that $deg(E_i) \leq \binom{b+1}{2}\binom{n-2}{b-1} \approx Y$ where $Y := N(b^4/2n^2)$. Using $q = o(1)$ and $Nq \approx e^{c'}$, it follows that $Yq = Nqb^4/(2n^2) = o(1)$ and also that $\ln(1-25q) = -25q[1+o(1)]$. To apply Local Lemma, it suffices to prove that

$$q \leq 25q(1-25q)^Y$$

Equivalently, it suffices to prove that

$$1 \leq 25e^{Y(\ln(1-25q))} = 25e^{-25Yq[1+o(1)]} = 25[1-o(1)]$$

The above clearly holds true. Now applying 4.4.1, one gets that

$$\mathbf{Pr}[\cap_i \overline{E_i}] \geq \prod_{i=1}^{N}(1-x) = (1-25q)^N \approx e^{-25Nq} \geq e^{-25e^{c'}}.$$

Therefore, $\mathbf{Pr}(mat(D) = b+1) \leq \mathbf{Pr}[mat(D) \geq b+1] = \mathbf{Pr}[\cup_i E_i] \leq c_2$ where $c_2 := 1 - e^{-25e^{c'}}$.

(ii) **Proof of $\mathbf{Pr}(mat(D) = b+1) \geq c_1$:**

To prove this, we use the following version of Paley-Zygmund Inequality (see [33]).

$$\mathbf{Pr}[X_{b+1} > 0] \geq E[X_{b+1}]^2 / E[X_{b+1}^2] \tag{4.8}$$

Notice that the RHS of the previous inequality 4.8 is exactly $1/(1+z)$, where $z = Var(X_{b+1})/E[X_{b+1}]^2$. As in the proof of Theorem 4.1.3, $z \leq E[X_{b+1}]^{-1} + M(n,p,b+1)$, and $M(n,p,b+1) \leq M(n,p,b).(\ln n)^3 + T_b$. Now, $M(n,p,b) = O\left(\frac{(\ln n)^6}{n^{1/4}}\right) = o(1)$, and it was shown that $T_b = o(1)$. Therefore, we get that $z \leq e^{-c'} + o(1) \approx e^{-c'}$, and therefore $1/(1+z)$ in 4.8 is at least c_1' where c_1 is the constant defined by $c_1 = 1/(1+e^{-c'})$. This proves that $\mathbf{Pr}(mat(D) \geq b+1) \geq c_1$. As a result, $\mathbf{Pr}(mat(D) = b+1) = \mathbf{Pr}(mat(D) \geq b+1) - \mathbf{Pr}(mat(D) \geq b+2) \geq c_1 - o(1) \approx c_1$.

Hence there exist constants $c_1, c_2 \in (0,1)$ such that,

$$\begin{aligned} c_1 \leq \quad & \mathbf{Pr}[mat(D) = b+1] \quad \leq c_2 \quad \text{and hence} \\ 1 - o(1) - c_2 \leq \quad & \mathbf{Pr}[mat(D) = b] \quad \leq 1 - o(1) - c_1. \end{aligned}$$

Thus, $mat(D)$ is **not** concentrated at any *single* point. ∎

4.4.3 Proof of Theorem 4.1.6

First, we prove the following lemma, from which the theorem follows as an easy consequence:

Lemma 4.4.2 *Let $i = i(n) \in \{1, \ldots \lfloor 2\log_2 n \rfloor\}$ be any fixed function of n. Let $D \in \mathcal{D}(n,p)$ and let $w = w(n)$ be any function of n so that as $n \to \infty$, $w \to \infty$ and $w \leq (0.5 \ln n)$. Then, asymptotically almost surely, the following are true :*

(i) If $p \geq n^{-2/(i-1+\frac{w}{\ln n})}$, then $mat(D) \geq i$.

(ii) If $p \leq n^{-2/(i-1-\frac{w}{\ln n})}$, then $mat(D) < i$.

Proof The probability of having an induced acyclic tournament of size i only increases with increasing p. From the one-point concentration result of Theorem 4.1.3, it follows that if p

is such that d (defined before) satisfies $d \geq i + \frac{w}{\ln n}$, then a.a.s. $mat(D) \geq i$. Similarly, if p is such that d satisfies $d \leq i - \frac{w}{\ln n}$, then a.a.s. $mat(D) < i$. However,

$$\begin{aligned} d \geq i + \frac{w}{\ln n} &\Leftrightarrow \log_r n \geq \frac{i-1}{2} + \frac{w}{2\ln n} \\ &\Leftrightarrow n \geq p^{-\left(\frac{i-1}{2} + \frac{w}{2\ln n}\right)} \\ &\Leftrightarrow p \geq n^{-2/\left(i-1+\frac{w}{\ln n}\right)} \end{aligned}$$

Similarly, we have

$$d \leq i - \frac{w}{\ln n} \Leftrightarrow p \leq n^{-2/\left(i-1-\frac{w}{\ln n}\right)}$$

This completes the proof of the lemma. ∎

From the above lemma, Theorem 4.1.6 can be proved as follows. We choose $w(n) = \sqrt{\ln n}$ and it satisfies the conditions of the lemma above. We set $lb_i(n) = n^{-2/(i-1-\frac{w}{\ln n})}$ and $ub_i(n) = n^{-2/(i-1+\frac{w}{\ln n})}$, and define $p_i(n) = (ub_i(n) + lb_i(n))/2$, and $q_i(n) = (ub_i(n) - lb_i(n))/2$.

If $i(n) \to \infty$, we choose $w(n) = i(n)/4$ so that $w(n) \to \infty$. Also, it can be verified that $lb_i(n) = ub_i(n)[1 - o(1)]$ and hence $q_i(n) = o(p_i(n))$, so we have a sharp threshold for such $i = i(n)$. ∎

Remark: In the above proof, notice that the ratio $\frac{q_i}{p_i} \leq \frac{ub_i - lb_i}{ub_i}$ which is

$$1 - e^{-\frac{w}{(i-1)^2 - (w/\ln n)^2}} = 1 - \left(1 - O\left(\frac{w}{(i-1)^2 - (w/\ln n)^2}\right)\right) = O(1/i)$$

for $w = i/4$.

4.5 Finding an induced acyclic tournament

In this section, we obtain a lower bound on the size of any maximal induced acyclic tournament. As a consequence, it follows that the following simple greedy heuristic GRDMAT(D) (described below) for finding a large induced acyclic tournament inside a random digraph, a.a.s., produces an acyclic tournament of size within a constant factor ($\geq 1/2$) of the optimal. It is easy to inductively verify that GRDMAT(D) always outputs a maximal acyclic tournament.

GRDMAT($D = (V, E)$)

1. $A := \emptyset$.
2. **while** $\exists\, u \in V \setminus A$ such that $D[A \cup \{u\}]$ induces an acyclic tournament **do**
3. Add u to A. (* ties in the choice of u are broken arbitrarily *)
4. **end**
5. Return $D[A]$ and halt.

The following theorem proves a lower bound on the size of any maximal acyclic tournament. In what follows, we assume that $p \geq n^{-1/4}$ mainly to focus on the interesting range of p. If p is smaller, then $mat(D) \leq 9$ a.a.s. and one can find provably optimal solutions in polynomial time.

Theorem 4.5.1 *Given $D \in \mathcal{D}(n,p)$ with $p \geq n^{-1/4}$ and any $w = w(n)$ such that $w(n) \to \infty$ as $n \to \infty$, with probability $1 - o(1)$, every maximal induced acyclic tournament is of size at least $\lceil \delta \log_r n \rceil$, where $\delta = 1 - \frac{\ln(\ln np + w)}{\ln n}$.*

Proof : Without loss of generality, we assume that $p \geq n^{-1/4}$ so that $\log_r n \geq 4$. Hence $d = \delta(\log_r n) > 3$ and $\lfloor d \rfloor \geq 3$.
For any induced acyclic tournament $D[A]$ of size $|A| = b$, $b \leq d = \delta(\log_r n)$, and any vertex $u \in V \setminus A$, the probability that u can be added to A is given by:

$$\mathbf{Pr}[D[A \cup \{u\}] \text{ is an acyclic tournament}] \;=\; (b+1)p^b$$

The above equality is true since $D[A \cup \{u\}]$ induces an acyclic tournament if and only if u can be added to any of the $b+1$ positions in the unique hamilton path of $D[A]$ in such a way that each of the edges joining u with vertices in A is present and is oriented in the proper direction. Also, this probability decreases with increasing b.

This event depends only on the edges joining u with the vertices in A, and hence, is independent of events corresponding to other vertices in $V \setminus A$. Therefore, the probability that $D[A]$ is a maximal acyclic tournament is given by

$$\begin{aligned} \mathbf{Pr}(D[A] \text{ is maximal }) &= \mathbf{Pr}[\forall u \in V \setminus A,\, u \text{ can't be added to } A] \\ &= \left(1 - (b+1)p^b\right)^{n-b} \end{aligned}$$

As b increases, this probability increases and hence achieves its maximum (for $b \leq d$) at $b = \lfloor d \rfloor$. Hence, for an induced acyclic tournament $D[A]$ of size $\lfloor d \rfloor$, we have (using $(d+1)(n-d) \geq nd$) :

$$\begin{aligned}\mathbf{Pr}(D[A] \text{ is maximal }) &\leq \left(1 - (\lfloor d \rfloor + 1)p^{\frac{\delta(\ln n)}{\ln r}}\right)^{n-\lfloor d \rfloor} \leq \left(1 - \frac{d+1}{n^\delta}\right)^{n-d} \\ &\leq e^{-\frac{(d+1)(n-d)}{n^\delta}} \leq e^{-dn^{1-\delta}}.\end{aligned}$$

For any *fixed* set A of size $b \leq d$, let $\mathcal{E}(A)$ denote the event that $D[A]$ is a maximal induced acyclic tournament.

$$\mathbf{Pr}(\mathcal{E}(A)) \leq b! p^{\binom{b}{2}} e^{-dn^{1-\delta}}$$

Thus,

$$\begin{aligned}\mathbf{Pr}(\exists A, |A| = b : \mathcal{E}(A)\ ) &\leq \binom{n}{b} b! p^{\binom{b}{2}} e^{-dn^{1-\delta}} \\ &\leq \left(np^{(b-1)/2}\right)^b e^{-dn^{1-\delta}} = (f(n,p,b))^b e^{-dn^{1-\delta}} \qquad (4.9)\end{aligned}$$

where we recall that $f(n,p,b) = np^{(b-1)/2}$.
Note that for each $b \leq \lfloor d \rfloor$,

$$\frac{f(n,p,b+1)^{b+1}}{(f(n,p,b))^b} = f(n,p,b+1)p^{b/2} = np^b \geq np^d = n^{1-\delta} = (\ln np) + w.$$

Hence

$$\sum_{b \leq d} (f(n,p,b))^b \leq (f(n,p,\lfloor d \rfloor))^{\lfloor d \rfloor} \sum_{b \leq \lfloor d \rfloor} (\ln np + w)^{-(\lfloor d \rfloor - b)} = 2(f(n,p,\lfloor d \rfloor))^{\lfloor d \rfloor}.$$

As a result, taking the union bound over all choices of A, we see that (using $\lfloor d \rfloor \geq 3$)

$$\mathbf{Pr}(\exists A,\ |A| \leq \lfloor d \rfloor : \mathcal{E}(A)) \leq 2\left(f(n,p,\lfloor d \rfloor)\right)^{\lfloor d \rfloor} e^{-dn^{1-\delta}} \leq 2\left(np\right)^{\lfloor d \rfloor} e^{-dn^{1-\delta}}$$

For $\delta = 1 - \frac{\ln(\ln np + w)}{\ln n}$, this probability is :

$$\mathbf{Pr}[\exists A,\ |A| \leq d : \mathcal{E}(A)] \leq 2 \cdot e^{d(\ln np - (\ln np + w))} = 2 \cdot e^{-dw} \leq 2e^{-w} \to 0 \text{ as } n \to \infty.$$

Hence, every maximal induced acyclic tournament is of size at least $\lceil d \rceil$. □

4.6 Another efficient heuristic with improved guarantee

We present below another efficient heuristic which will be analyzed and be shown to have an additive improvement of $\Theta\left(\sqrt{\log_r n}\right)$ over the guarantee given (in Section 4) on the size of any maximal solution. It is similar to a heuristic presented in [44] for finding large independent sets in $G \in \mathcal{G}(n, 1/2)$. We show that, for every large constant $c > 0$, one can find in polynomial time an acyclic tournament of size at least $\lfloor \log_r n + c\sqrt{\log_r n} \rfloor$.

The idea is to construct greedily a solution A of size $g(n, p, c) = \lceil \log_r n - c\sqrt{\log_r n} \rceil$ and then add an optimal solution (found by an exhaustive search) in the subgraph induced by those vertices each of which can be safely and individually added to A to get a bigger solution. We will show that exhaustive search can be done in polynomial time and yields (a.a.s.) a solution of size $2c\sqrt{\log_r n}$. As a result, we finally get a solution of the stated size. The algorithm is described below.

ACYTOUR$(D = (V, E), p, c)$

1. Choose and fix a linear ordering σ of V.
2. $c' = 1.2c$; $A = \emptyset$; $B = V$.
3. **while** $B \neq \emptyset$ and $|A| < g(n/2, p, c')$ **do**
4. Let u be the σ-smallest vertex in B.
5. **If** $D[A \cup \{u\}]$ induces an acyclic tournament **then** add u to A.
6. remove u from B. **endwhile**
7. **if** $|A| < g(n/2, p, c')$ **or** $|B| < n/2$, **then** Return FAIL and halt.
8. $C = \{u \in B : \forall v \in A,\ v \to u \in E\}$; $r = p^{-1}$; $\mu = |B|p^{|A|}$.
9. **if** $|C| \notin [(0.9)\mu, (1.1)\mu]$ **then** Return FAIL.
10. **for each** $X \subset C$: $|X| = \lfloor 2c'\sqrt{\log_r n/2} \rfloor - 1$ **do**
11. **if** $D[X]$ is an acylic tournament **then** Return $D[A \cup X]$ and halt. **endfor**
12. Return FAIL.

We analyze the above algorithm and obtain the following result.

Theorem 4.6.1 *Let $D \in \mathcal{D}(n,p)$. For every sufficiently large constant $c \geq 1$: if p is such that $n^{-1/c^2} \leq p \leq 0.5$, then, with probability $1-o(1)$, ACYTOUR(D) will output an induced acyclic tournament of size at least $b' = \lfloor (1+\epsilon')\log_r n \rfloor$, where $\epsilon' = c/\sqrt{\log_r n}$.*

Proof : Recall our assumption that c is sufficiently large.

Correctness : First, we prove the correctness. Note that $D[A]$ is always an induced acyclic tournament. Also, each $u \in C$ is such that $D[A \cup \{u\}]$ is an acyclic tournament with u as the unique sink vertex (having zero out-degree). Hence, any acyclic tournament $D[X]$ present as a subgraph in $D[C]$ can be safely added to A so that $D[A \cup X]$ also induces an acyclic tournament.

Analysis : Consider the following events defined as

$$\begin{aligned}
\text{Failure at step } 7 : \mathcal{E}_1 &:= |A| < g(n/2,p,c') \textbf{ or } |B| < n/2 \ ; \\
\text{Failure at step } 9 : \mathcal{E}_2 &:= \overline{\mathcal{E}_1} \cap \mathcal{E}_2' \text{ where } \mathcal{E}_2' := |C| \notin [(0.9)\mu, (1.1)\mu] \\
\text{Failure at step } 12 : \mathcal{E}_3 &:= \overline{\mathcal{E}_1} \cap \overline{\mathcal{E}_2} \cap \mathcal{E}_3' \text{ where } \mathcal{E}_3' := mat(D[C]) < \lfloor 2c'\sqrt{\log_r n/2} \rfloor
\end{aligned} \tag{4.10}$$

If none of the events $\{\mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3\}$ holds, then the algorithm will succeed and output a solution whose size is

$$\begin{aligned}
|A \cup X| &\geq \log_r(n/2) - c'\sqrt{\log_r(n/2)} + 2c'\sqrt{\log_r n/2} - 2 \\
&\geq (1+\epsilon')(\log_r n) + (c'-c)\sqrt{\log_r n/2} - 2.5 - \log_r 2 \\
&\geq (1+\epsilon')(\log_r n) + (0.2c)\sqrt{\log_r n/2} - 3.5 \\
&\geq (1+\epsilon')(\log_r n).
\end{aligned}$$

The probability of this happening is

$$\mathbf{Pr}(\overline{\mathcal{E}_1} \cap \overline{\mathcal{E}_2} \cap \overline{\mathcal{E}_3}) = 1 - \mathbf{Pr}(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3)$$

The events $\mathcal{E}_1, \mathcal{E}_2'$ and $\mathcal{E}_3'$ are totally independent since they are determined by pairwise disjoint sets of potential edges. Also, the events $\mathcal{E}_1, \mathcal{E}_2$ and $\mathcal{E}_3$ are mutually exclusive and hence

$$\begin{aligned}
\mathbf{Pr}(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3) &= \mathbf{Pr}(\mathcal{E}_1) + \mathbf{Pr}(\mathcal{E}_2) + \mathbf{Pr}(\mathcal{E}_3) \\
&\leq \mathbf{Pr}(\mathcal{E}_1) + \mathbf{Pr}(\mathcal{E}_2' \mid \overline{\mathcal{E}_1}) + \mathbf{Pr}(\mathcal{E}_3' \mid \overline{\mathcal{E}_1} \cap \overline{\mathcal{E}_2})
\end{aligned} \tag{4.11}$$

Let V_1 denote the set of first $n/2$ vertices of σ. Then, by Theorem 4.5.1, any maximal tournament in $D[V_1]$) is of size at least $\log_r(n/2) - \log_r(\ln(n/2) + \ln\ln(n/2)) \geq g(n/2,p,c') = \lceil \log_r(n/2) - c'\sqrt{\log_r(n/2)} \rceil$, with probability $1-o(1)$. Hence, $\mathbf{Pr}(\mathcal{E}_1) = o(1)$.

For any fixed vertex $u \in B$,

$$\mathbf{Pr}(u \in C) \;=\; \mathbf{Pr}(\forall\, v \in A,\; (v,u) \in E) \;=\; p^{|A|}.$$

Hence

$$\mu \;=\; E[|C|] \;=\; |B| \cdot p^{|A|}.$$

Since $|C|$ is the sum of $|B|$ identical and independent indicator random variables, by applying Chernoff-Hoeffding bounds (see [53, 9]), we get that

$$\mathbf{Pr}(|C| \notin [(0.9)\mu, (1.1)\mu] \;\leq\; 2e^{-\mu/300}.$$

Since $|A| = g(n/2, p, c')$, we deduce that

$$\mu \;\approx\; |B| \cdot 2r^{c'\sqrt{\log_r n/2}}/n,$$

after justifiably ignoring the effect of the ceiling function used in the definition of $g(n/2, p, c)$. More precisely, since we know that $\mathcal{E}_1$ has not occured, $|B| \geq n/2$ and hence

$$r^{c'\sqrt{\log_r n/2}} \;\leq\; \mu \;\leq\; 2r^{c'\sqrt{\log_r n/2}} \tag{4.12}$$

It is easy to verify that $\mu \to \infty$ as $n \to \infty$. Hence $\mathbf{Pr}(\mathcal{E}_2' \mid \overline{\mathcal{E}_1}) \;=\; o(1)$.

Given that neither of $\mathcal{E}_1$ and $\mathcal{E}_2$ holds, it follows that $|C| \;\geq\; (0.9)\mu \;\approx (0.9) \cdot r^{c'\sqrt{\log_r n/2}}$. Hence, using $r \geq 2$ and applying Theorem 4.1.2,

$$mat(D[C]) \;\geq\; \lfloor 2c'\sqrt{\log_r n/2} + 0.5 + 2\log_r 0.9 \rfloor \;\geq\; \lfloor 2c'\sqrt{\log_r n/2} \rfloor - 1$$

with probability $1 - o(1)$. This establishes that $\mathbf{Pr}(\mathcal{E}_3' \mid \overline{\mathcal{E}_1} \cap \overline{\mathcal{E}_2}) = o(1)$. It then follows from (6.2) that ACYTOUR(D) outputs a solution of required size with probability $1 - o(1)$.

Time Complexity : It is easy to see that the running time is polynomial except for the *for* loop of lines 10 and 11. The maximum number of iterations of the **for** loop is at most

$$\binom{|C|}{|X|} \;\leq\; \binom{(1.1)\mu}{\lfloor 2c'\sqrt{\log_r(n/2)} \rfloor} \;\leq\; \binom{(1.1) \cdot 2r^{c'\sqrt{\log_r(n/2)}}}{\lfloor 2c'\sqrt{\log_r(n/2)} \rfloor} \;=\; O\left(r^{4c'^2(\log_r n)}\right) \;=\; O\left(n^{O(1)}\right).$$

where the upper bound on μ is the one obtained in 4.12. Since each iteration takes polynomial time, the algorithm finishes in polynomial time always. ∎

Remark In Theorem 4.6.1, we assume that $p \geq n^{-1/c^2}$. This is because if $p \leq n^{-1/c^2}$, then $mat(D) \leq \lceil 2c^2 + 1 \rceil$ a.a.s. and hence even a provably optimal solution can be found in polynomial time a.a.s..

4.7 $mat(D)$ for non-simple random digraphs

We also consider another model introduced in [61] which does not force the random digraph to be simple and allows cycles of length 2.

Model $D \in \mathcal{D}_2(n,p)$ **:** Choose each *directed* edge $u \to v$ joining distinct elements of V independently with probability p.

Note that if $D \in \mathcal{D}_2(n,p)$ and $D' \in \mathcal{D}_2(n,1-p)$, then for every b, we have

$$\mathbf{Pr}(mat(D) = b) \;=\; \mathbf{Pr}(mat(D') = b).$$

Hence, for the rest of this section, without loss of generality, we assume that $p \leq 0.5$ and use q to denote $1-p$.

The maximum size of any induced acyclic tournament is determined by those unordered pairs $\{u,v\}$ such that exactly one arc between u and v is present. Hence, if $D \in \mathcal{D}_2(n,p)$ and $D' \in \mathcal{D}(n,pq)$, then for every b, we have

$$\mathbf{Pr}(mat(D) = b) \;=\; \mathbf{Pr}(mat(D') = b).$$

Hence, we can obtain the following analogues of Lemma 4.2.1, Theorems 4.1.2, 4.1.3, 4.1.6, 4.5.1, 4.6.1 and Corollary 4.1.4.

Lemma 4.7.1 *For any positive integer b, for a random digraph $D \in \mathcal{D}_2(n,p)$,*

$$Pr[mat(D) \geq b] \;\;\geq\;\; Pr[\omega(G) \geq b].$$

where $G \in \mathcal{G}(n,pq)$.

Theorem 4.7.2 *Let $D \in \mathcal{D}_2(n,p)$ with $p \geq 1/n$. Define*

$$d = 2\log_{(pq)^{-1}} n + 1 = \frac{2(\ln n)}{\ln(pq)^{-1}} + 1; \quad b^* \;=\; \lfloor d - 1/2 \rfloor.$$

Then, a.a.s. as $n \to \infty$, $mat(D)$ is either b^* or b^*+1.

Theorem 4.7.3 *Let $D \in \mathcal{D}_2(n,p)$. Let $w = w(n)$ be any function so that as $n \to \infty$, $w \leq 0.5(\ln n)$ and $w \to \infty$. If $p = p(n)$, $p \geq 1/n$, is such that d (defined in Theorem 4.7.2) satisfies $\frac{w}{\ln n} \leq \lceil d \rceil - d \leq 1 - \frac{w}{\ln n}$ for all large values of n, then $mat(D)$ is a.a.s equal to $\lfloor d \rfloor$.*

Corollary 4.7.4 *Let $D \in \mathcal{D}_2(n,p)$. For every constant function $p = p(n)$, there exists a function $f = f(n) = 1 - o(1)$ such that the set $N_{f,p}$ is a subset of natural numbers having density 1.*

Our goal is to obtain a threshold statement in terms of $p = p(n)$. Firstly, observe that Theorem 4.1.6 can be applied straightaway to get a threshold statement (for $\mathcal{D}_2(n,p)$ model) in terms of the parameter pq. However, to get a threshold in terms of p more work needs to be done. Before stating the analogue of the threshold theorem, we need some definitions:

Let $w = w(n)$ be a sufficiently slow-growing function of n, such that $w = \omega(1)$ and $w = o(\ln n)$. Let $i = i(n)$ be a suitably growing function which goes to ∞ as $n \to \infty$. Define $a = n^{-2/(i-1+\frac{w}{\ln n})}$, and $b = n^{-2/(i-1-\frac{w}{\ln n})}$. Let $f(x,y)$ denote the function $x^2 - x + y$.

Theorem 4.7.5 *Let $i = i(n) \in \{1, \ldots \lfloor \log_4 n \rfloor\}$ (for every n) be any fixed function of n. Then, there exist functions $c = c_i(n) \in [0,1]$ and $d = d_i(n) \in [0,1]$ such that : If $D \in \mathcal{D}_2(n,p)$ with $p \geq 1/n$, then, asymptotically almost surely, the following are true :*

(i) If $p \geq c$, then $pq \geq a$ and hence $mat(D) \geq i$.

(ii) If $p \leq d$, then $pq \leq b$ and hence $mat(D) < i$,

where c, d are the real positive roots in the range $[0, 1/2]$ of the quadratic equations $f(x,a) = 0$ and $f(y,b) = 0$ respectively. Also, if $i = i(n)$ is a growing function, then $c - d = o(c)$.

Hence it follows that we obtain thresholds (sharp if $i = i(n)$ increases) for the existence of induced acyclic tournaments of size i.

Proof : Notice that $pq = p(1-p) = p - p^2$ and hence if $pq = y$, $y \in \Re^+$, then $p^2 - p + y = 0$, i.e. $f(p,y) = 0$. Now, taking y to be a and b respectively, we get that if $p = c$, then $pq = a$; if $p = d$, then $pq = b$. Also, since pq is increasing when $x \in [0, 1/2]$, $p \geq c$ implies $pq \geq a$, and $p \leq d$ implies $pq \leq b$. The Claims (i) and (ii) now follow by applying Lemma 4.4.2. It is easy to check that for each $y = a, b$, $f(x,y) = 0$ has 2 positive real roots only one of which lies in the range $[0, 1/2]$.

Now, for a sharp threshold we need to show that $(c - d) = o(c)$, i.e. $1 - d/c = o(1)$. This is proved as follows: If $d/c \geq (1 - 1/\sqrt{i})$, then we are done, since $1 - d/c \leq 1/\sqrt{i} = o(1)$. Therefore, assume that $d/c \leq (1 - 1/\sqrt{i})$. Now, $c \in [0, 1/2]$ and hence $c \leq 1/2$. By our assumption, $d \leq (1 - 1/\sqrt{i})c \leq \frac{1}{2}(1 - 1/\sqrt{i})$. Hence, $c + d \leq 1 - \frac{1}{2\sqrt{i}}$. Since c and d satisfy $f(c,a) = 0$ and $f(d,b) = 0$, after subtracting, we get $f(c,a) - f(d,b) = (c-d)(c+d-1) + a - b = 0$. Therefore, $c - d = \frac{a-b}{1-c-d}$. Now using the upper bound on $c + d$, we get $c - d \leq \frac{a-b}{1/(2\sqrt{i})} = 2(a-b)\sqrt{i}$. Observe that $a \leq c \leq 2a$, since $a = c - c^2$ and $c \in [0, 1/2]$. Therefore, $(c-d)/c \leq (c-d)/a \leq 2(a-b)\sqrt{i}/a$. But from the remark following the proof of Theorem 4.1.6, we have that $(a-b)/a = O(1/i)$. Therefore $(c-d)/c = O(\sqrt{i}/i) = O(1/\sqrt{i}) = o(1)$. Thus in this case too, the threshold is seen to be sharp. ∎

Theorem 4.7.6 *Given $D \in \mathcal{D}_2(n,p)$ with $pq \geq n^{-1/4}$ and any $w = w(n)$ such that $w(n) \to \infty$ as $n \to \infty$, with probability $1 - o(1)$, every maximal induced acyclic tournament is of size at least $d = \lfloor \delta \log_{(pq)^{-1}} n \rfloor$, where $\delta = 1 - \frac{\ln(\ln(npq)+w)}{\ln n}$.*

Theorem 4.7.7 *Let $D \in \mathcal{D}_2(n,p)$. For every sufficiently large constant $c \geq 1$: if $p \leq 0.5$ is such that $n^{-1/c^2} \leq pq \leq 0.25$, then, with probability $1 - o(1)$, ACYTOUR(D) will output an induced acyclic tournament of size at least $b' = \lfloor (1+\epsilon') \log_{(pq)^{-1}} n \rfloor$, where $\epsilon' = c/\sqrt{\log_{(pq)^{-1}} n}$.*

Remark However, in the case of $\mathcal{D}_2(n,p)$ model, we need to slightly modify the description of ACYTOUR(D) as follows : In the definition of C (Line 8), we also need to require that $(u,v) \notin E$ for each $v \in A$.

4.8 On the maximum size of induced tournaments

Suppose we drop the requirement of acyclicity of the induced tournament. It then reduces to the clique problem as follows. Let us first recall some basic facts about the distributions of $\omega(G)$ and $\alpha(G)$ for $G \in \mathcal{G}(n,p)$. $\omega(G)$ ($\alpha(G)$) denotes the maximum size of a clique (an independent set) in G. It is easy to verify that $\omega(G)$ for $G \in \mathcal{G}(n,p)$ and $\alpha(G)$ for $G \in \mathcal{G}(n,1-p)$ are identically distributed. Also, by the classical results of Bollobás and Erdős [17], and Grimmett and McDiarmid (see e.g. [14], Chapter 11), $\omega(G)$ is a.a.s. concentrated in just two values for every $p = p(n) \leq 1 - n^{-\epsilon}$ for some suitably small constant $\epsilon > 0$. But it does not seem to exhibit such sharp concentration behavior for larger values of p. In particular, if p is such that $p = 1 - n^{-2/3}$, $\omega(G)$ is only known (see [33]) to be concentrated in a band of $\Theta(n^{2/3})$.

This has implications to the concentration of the maximum size of an induced (need not be acyclic) tournament in a random digraph. We use $\omega(D)$ to denote the maximum size of an induced tournament in D. It is clear that $\omega(D)$ for $D \in \mathcal{D}(n,p)$ and $\omega(G)$ for $G \in \mathcal{G}(n,2p)$ are identically distributed for every $p = p(n) \leq 0.5$. Similarly, $\omega(D)$ for $D \in \mathcal{D}_2(n,p)$ and $\omega(G)$ for $G \in \mathcal{G}(n,2p(1-p))$ are identically distributed for every $p = p(n) \leq 1$.

But, unlike the case of $mat(D)$, the concentration of $\omega(D)$ is quite different between the two models $\mathcal{D}(n,p)$ and $\mathcal{D}_2(n,p)$. First, we focus on the model $\mathcal{D}_2(n,p)$. Since $2p(1-p) \leq 0.5$ for any $0 \leq p \leq 1$, $\omega(G)$ is 2-point concentrated for $G \in \mathcal{G}(n,2p(1-p))$, and hence we notice that $\omega(D)$ is always concentrated in just two values for any p.

If $D \in \mathcal{D}(n,p)$, then $\omega(D)$ is concentrated in just two values a.a.s. for any $p = p(n) \leq 0.25$. However, for $0.25 < p \leq 0.5$, $\omega(D)$ is not tightly concentrated and has the same distribution and concentration behavior as $\omega(G)$ for certain ranges of $p \geq 0.5$ (see the discussion before).

4.9 Summary

The problem of determining the size of the largest induced acyclic tournament $mat(D)$ in a random digraph was studied. We showed that a.a.s. $mat(D)$ takes one of only two possible values. The result is valid for all ranges of the arc probability p. The value of $mat(D)$ also has an explicit closed form expression (for all ranges of p) which does not seem to exist for clique number $\omega(G)$ of a random graph.

The results of this chapter and those of [63], [61] and [24] show that $mat(D)$ of a random digraph behaves like the clique number $\omega(G)$ of a random graph and maximum induced acyclic subgraph size $mas(D)$ behaves like the independence number $\alpha(G)$ of a random graph (see also the discussion above in Section 7).

We then showed that a.a.s. every maximal acyclic tournament is of a size which is at least nearly half of the optimal size. As a result, one immediately gets an efficient approximation algorithm whose approximation ratio is bounded by $2 + O((\ln \ln n)/(\ln n))$. We also considered and analyzed another efficient heuristic whose approximation ratio was shown to be $2 - O(1/\sqrt{\log_r n})$.

An interesting and natural open problem that comes to mind is the following.

Open Problem : Let p be a constant such that $0 < p \leq 0.5$. Design a polynomial time algorithm which, given $D \in \mathcal{D}(n,p)$, a.a.s. finds an induced acyclic tournament of size at least $(1+\epsilon)\log_r n$ for some positive constant ϵ.

Solving this problem could turn out to be as hard as designing an efficient algorithm which finds, given $G \in \mathcal{G}(n,1/2)$, a clique of size $(1+\epsilon)\log_2 n$ and the latter problem has remained open for more than three decades.

Chapter 5

Largest Induced Acyclic Subgraphs in Random Digraphs: Concentration and Lower Bounds

5.1 Introduction

A *directed acyclic graph* (**dag**) is a digraph without any directed cycles. Given a directed graph $D = (V, A)$, we want to find the maximum size (i.e. number of vertices) of an induced dag in D, denoted by $mas(D)$. In this chapter, we shall study this parameter theoretically for random digraphs. We shall first obtain some improved bounds on the concentration of the distribution of the digraph invariant $mas(D)$. Next, we shall look at the existing lower bounds of the parameter $mas(D)$ using the $\mathcal{D}(n, p)$ and $\mathcal{D}_2(n, p)$ models of random digraphs, and obtain improved lower bounds. The proofs of our results are presented in Section 5.4.

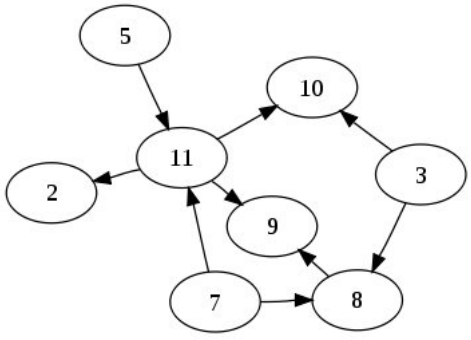

Figure 5.1: A directed acyclic graph

Notation : In what follows, $p \leq 0.5$ is a real number. Throughout the chapter, we use q to denote $(1-p)^{-1}$ and w to denote np.

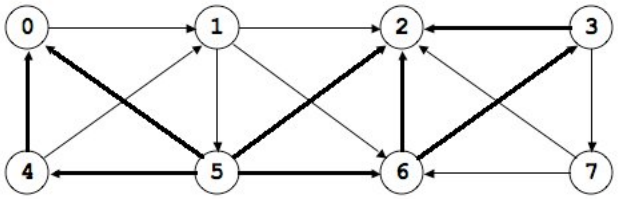

Figure 5.2: A maximum induced acyclic subgraph: {*0,2,3,4,5,6*}

5.1.1 Improved concentration results

The theoretical aspects of this problem were studied initially by Subramanian [63] and later Spencer and Subramanian [61] obtained the following result:

Theorem 5.1.1 *[61] Let $D \in \mathcal{D}(n,p)$ and $w = np$. There is a sufficiently large constant C such that : If p satisfies $w \geq C$, then a.a.s,*

$$mas(D) \quad \in \quad \left[\left(\frac{2}{\ln q}\right)(\ln w - \ln\ln w - O(1)), \left(\frac{2}{\ln q}\right)(\ln w + 3e)\right]$$

where $q = (1-p)^{-1}$.

The above theorem implies that $mas(D)$ is concentrated in an integer band of width $O\left(\frac{\ln\ln w}{\ln q}\right)$. For small p, this width can become quite large, for example, for $p = C/n$, it is $\Theta(n)$. For the concentration of the parameter $mas(D)$, we initially establish an "essentially" $\sqrt{\log_q w}$ $(= O(\sqrt{p^{-1}\ln w}))$ width for all ranges of p by using (see Chapter 2) Talagrand's Inequality. For a random variable X, define its median to be any value m such that $\mathbf{Pr}(X \leq m) \geq 0.5$ and $\mathbf{Pr}(X \geq m) \geq 0.5$.

Theorem 5.1.2 *Let $D \in \mathcal{D}(n,p)$ $(p = o(1))$ with $w = np$ and m being a median value of $mas(D)$. Then, for any $\beta = \beta(n)$ such that $\lim_{n\to\infty}\beta = \infty$, w.h.p.*

$$|mas(D) - m| \quad \leq \quad \beta\sqrt{\log_q w}$$

The proof of this theorem follows from a standard application of Talagrand's inequality and is presented in Section 5.3. Theorem 5.1.2 provides a sharper concentration than Theorem 5.1.1 for all p such that $p = o((\ln\ln n)^2/(\ln n))$. However, it does not give the location

of the concentration, i.e. the value of the median m. For small values of p, it will be better if we can obtain and prove explicit lower and upper bounds with a difference which is asymptotically at most the range of concentration proved by Theorem 5.1.2. However, the difference of the currently known bounds is much larger. This indicates that these bounds are not tight and it should be possible to find sharper upper and/or lower bounds. Conversely, if the precise location of the median of the distribution $mas(D)$ could be ascertained, a straightforward application of Talagrand's inequality would lead to sharp and explicit upper and lower bounds.

5.1.2 Improved explicit lower bounds

We also obtain improvements over known lower and upper bounds. The following theorem improves the lower bound of Theorem 5.1.1 in the range $p = \omega(n^{-1/2} \log n)$:

Theorem 5.1.3 *Let $D \in \mathcal{D}(n,p)$. There is a large positive constant W such that : For any $\beta = \beta(n) \to \infty$ and $p = p(n)$ with $n^{-1/2}(\ln n)\beta(n) \leq p \leq 1/2$, a.a.s.,*

$$mas(D) \quad \geq \quad \left(\frac{1}{\ln q}\right)(2 \ln w - W)$$

The gap between $2 \log_q np$ and the lower bound in Theorem 5.1.3 is $O(1/\ln q) = O(p^{-1})$, which can become large when p is small. This gap can be reduced to an absolute constant when p is larger:

Theorem 5.1.4 *For every small constant $\epsilon > 0$, the following is true : Given $D \in \mathcal{D}(n,p)$, if p is such that $n^{-1/3+\epsilon} \leq p \leq 1/2$, then a.a.s*

$$mas(D) \quad \geq \quad \frac{2 \ln w}{\ln q} - 1$$

The proofs of Theorems 5.1.3 and 5.1.4 are presented in Section 5.4. The theorems are established by a common proof (both being based on the Second Moment Method) with occasionally separate arguments to take care of the different assumptions and conclusions of the two theorems.

For $p \leq n^{-1/2} \log n$, the last two theorems do not apply. The basic problem is that the variance of the number of topologically ordered sets shoots up when p is in this range. However, as Theorem 5.1.2 shows, the concentration of $mas(D)$ is quite sharp around a median. To utilize this fact and overcome the problem of large variance, we combine Talagrand's inequality with a second moment-based inequality to get an approximate estimate of the

location of a median of $mas(D)$. As a result, we obtain the following theorem which is proved in Section 5.5.

Theorem 5.1.5 *There exist suitable positive constants C and W such that given p, with $C/n \leq p \leq 1/2$ and $D \in \mathcal{D}(n,p)$, a.a.s.*

$$mas(D) \geq \frac{2}{\ln q}(\ln w - W)$$

A similar technique was used by Alan Frieze for independent sets in random graphs (though he used Azuma's inequality, see [33]. Later Janson, Łuczak, and Rucínski obtained a slightly easier proof of Frieze's result by replacing Azuma's with Talagrand's inequality in the proof [40]).

Thus, this theorem extends the result of Theorem 5.1.3 to a much bigger range of p. As a consequence, the difference between the known upper and lower bounds is reduced to $O(1/\ln q)$ (which is $O(1/p)$, thereby improving the gap guaranteed by Theorem 5.1.1) for all $p \geq C/n$ (where C is some suitable constant). [1] Even though Theorem 5.1.3 appears to be subsumed by Theorem 5.1.5, we state it separately because the lower bound on p, namely $p = \omega(n^{-1/2}(\ln n))$, is the point upto which we can show that the variance of the random variable considered is asymptotically smaller than the square of its mean. Also, the estimated constant W for this range is smaller than the corresponding one for Theorem 5.1.5.

5.2 $mas(D)$ versus $\alpha(G)$

In this section, we explore some connections between the maximum acyclic subgraph number $mas(D)$ and the independence number $\alpha(G)$. As in the previous chapter, the results compare (a) the probability distributions of $\alpha(G), mas(D)$, where $G \in \mathcal{G}(n,p)$ and $D \in \mathcal{D}(n,p)$ respectively, and (b) lower bounds of $\alpha(G)$, $mas(D)$ in arbitrary graphs (and corresponding digraphs, in some sense - to be made precise later).

[1] This can be seen, for example, by comparing (i) the gap of $O\left(\frac{\ln\ln w}{\ln q}\right)$ between the upper and lower bounds for $mas(D)$ guaranteed by Theorem 5.1.1 alone, and (ii) the gap of $O\left(\frac{1}{\ln q}\right)$ between the upper bound of Theorems 5.1.1 together with the lower bound of Theorem 5.1.5. When $C \geq e^{e^W}$, where W is the constant of Theorem 5.1.5, the gap is strictly reduced.

5.2.1 Comparison of Probability distributions

The following lemma proved in [63] relates the probabilities in the models $\mathcal{D}(n,p)$ and $\mathcal{G}(n,p)$ for having, respectively, maximum acyclic subgraph number and independence number equal to a given value. We present this result with its proof for the benefit of the reader.

Lemma 5.2.1 *For any positive integer b, for a random digraph $D \in \mathcal{D}(n,p)$, $0 \leq p \leq 1/2$,*

$$Pr[mas(D) \geq b] \quad \geq \quad Pr[\alpha(G) \geq b].$$

where $G \in \mathcal{G}(n,p)$.

Proof : Given a linear ordering σ of vertices of D and a subset A of size b, we say that $D[A]$ is consistent with σ if there is no pair of vertices $\sigma_i, \sigma_j \in A$ with $i < j$, such that $D[A]$ has the arc (σ_j, σ_i).

Let τ denote an arbitrary but fixed ordering of V. Once we fix τ, the spanning subgraph of D formed by arcs of the form $(\tau(i), \tau(j))$ $(i < j)$ is having the same distribution as $\mathcal{G}(n,p)$. Hence, for any A, the event of $D[A]$ being consistent with τ is equivalent to the event of A inducing an independent set in $\mathcal{G}(n,p)$. Hence,

$$\begin{aligned}
\mathbf{Pr}(\ mas(D) \geq b\ ) &= \mathbf{Pr}(\ \exists A,\ |A| = b,\ D[A]\ \text{ is an induced acyclic subgraph }) \\
&= \mathbf{Pr}(\ \exists A,\ |A| = b,\ \exists \sigma,\ D[A]\ \text{ is consistent with }\ \sigma\ ) \\
&= \mathbf{Pr}(\ \exists \sigma,\ \exists A,\ |A| = b,\ D[A]\ \text{ is consistent with }\ \sigma\ ) \\
&\geq \mathbf{Pr}(\ \exists A,\ |A| = b,\ D[A]\ \text{ is consistent with }\ \tau\ ) \\
&= \mathbf{Pr}(\ \alpha(G) \geq b\ ).
\end{aligned}$$

Hence it is natural that we have a bigger lower bound for $mas(D)$ than we have for $\alpha(G)$. ■

5.2.2 Lower Bounds

Another intriguing connection between $mas(D)$ and $\alpha(G)$ stems from the degree-sequence based lower bound expressions that can be obtained for them. In [9], Alon and Spencer gave a probabilistic proof of the following degree-sequence based lower bound on the independence number $\alpha(G)$ of an n-vertex graph G:

$$\alpha(G) \geq \sum_{v \in V} \frac{1}{d(v)+1}$$

where $d(v)$ is the degree of the vertex $v \in V[G]$. A similar expression can be obtained for induced acyclic subgraphs in digraphs:

Theorem 5.2.2 *Given a simple digraph $D = (V, E)$ on n vertices, with $d^+(v), d^-(v)$ denoting the out-degree and in-degree, respectively, of the vertex v, the maximum induced acyclic subgraph of D is of size*

$$mas(D) \geq \max\left\{\sum_{v\in V} \frac{1}{d^-(v)+1}, \sum_{v\in V} \frac{1}{d^+(v)+1}\right\} \tag{5.1}$$

Proof Choose uniformly at random a linear ordering '$<$' over V and define two sets I_i and I_o as follows :

(i) I_i is the set of those $v \in V$ satisfying : for every in-neighbor u of v, we have $u < v$.

(ii) I_o is the set of those $v \in V$ satisfying : for every out-neighbor u of v, we have $u < v$.

Writing each of $|I_i|$ and $|I_o|$ as a sum of indicator variables (one for each v) and applying Linearity of Expectation, it can be easily verified that $\mu_i = E[|I_i|] = \sum_{v\in V} \frac{1}{d^-(v)+1}$ and also that $\mu_o = E[|I_o|] = \sum_{v\in V} \frac{1}{d^+(v)+1}$. Also, $E[\max\{|I_i|, |I_o|\}] \geq \max\{\mu_i, \mu_o\}$. Hence there exists a linear ordering for which either I_i or I_o is of size at least $\max\{\mu_i, \mu_o\}$. In any case, each of I_i and I_o always induces an acyclic subgraph of D. This proves the theorem. ∎

5.3 Proof of Theorem 5.1.2

A statement of Talagrand's inequality can be found in Chapter 2. To use Talagrand's however, we first need a lemma about the Lipschitz bound of the random variable $mas(D)$.

Lemma 5.3.1 *For any two digraphs D, D' which differ only in edges incident at a single vertex,*

$$|mas(D) - mas(D')| \quad \leq \quad 1$$

Proof : Let v be the vertex at whose incident edges D and D' differ. Remove the vertex v from D and D'. The resulting digraphs are now identical, and so $mas(D\backslash v) = mas(D'\backslash v)$. Now, restoring the vertex v to D and D' can only retain or increase the size of the optimal DAGs in D and D'. So either both $mas(D)$ and $mas(D')$ remain the same or both rise by 1, or only one of them rises. In all cases, we have $|mas(D) - mas(D')| \leq 1$. □

Proof (of Theorem 5.1.2): The proof follows from Theorem 2.3.3 and Lemma 5.3.1. We view D as a n-tuple $(E_1, \ldots, E_n)$ where E_i is the set of edges joining vertex i with some $j \leq i-1$. Let $mas(D)$ be the random variable X of Theorem 2.3.3. By Lemma 5.3.1, it follows that if two digraphs on V differ only in E_i, then their $mas(D)$ values differ by at most 1. Also, by definition X is $mas(D)$-certifiable (the E_i's associated with the vertices of any r-sized induced dag certify that $mas(D) \geq r$). Then X satisfies the requirements of Theorem 2.3.3 and hence applying the theorem, we have

$$Pr[|X - m| \geq t] \quad \leq \quad 2e^{-t^2/20 \log_q w}. \tag{5.2}$$

where m is a median of X. The denominator in the RHS exponent follows by observing that $\psi(m) = m$, and $m \leq 2 \log_q w + O(p^{-1}) \leq 2.5 \log_q w$. Taking $t = \beta\sqrt{\log_q w}$, where β is any asymptotically increasing function of n, the result follows. □

5.4 Proofs of Theorems 5.1.3 and 5.1.4:

We now present the proofs of Theorems 5.1.3 and 5.1.4. Before that, we introduce some facts, notations and definitions: **(F1)** For p, $0 \leq p \leq 0.5$, it is easy to verify that $p \leq \ln q$ and $\ln q \leq (1.5)p$. **(N1)** We use the standard notation $(n)_b$ to denote the expression $n(n-1)\ldots(n-b+1)$ defined for all positive integers n and b. **(D1)** Given a directed graph $D = (V, E)$, a *topological ordering* of a set $A \subseteq V$ is a permutation $\sigma : [|A|] \to A$ such that every arc in $D[A]$ is of the form $\sigma(i) \to \sigma(j)$, where $i < j$. **(D2)** A pair of vertices $\sigma(i), \sigma(j) \in A$ is said to be *consistent* with an ordering σ if they do not induce a backward arc in $D[A]$, i.e. an arc of the form $\sigma(i) \to \sigma(j)$, where $i > j$. **(D3)** Let S_n denote the set of all permutations of $[n]$. Given a permutation $\sigma \in S_n$, an *inversion* is defined to an (unordered) pair of elements $i, j \in [n]$, such that $i < j$ but $\sigma(i) > \sigma(j)$.

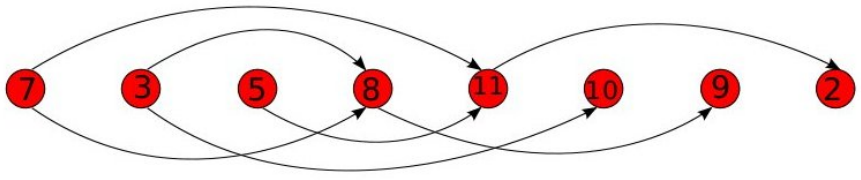

Figure 5.3: A topological ordering of the directed acyclic graph in Figure 5.1

We use the following (easily provable) identity from [62].

Lemma 5.4.1 *[62] For $\sigma \in S_n$, let $i(\sigma)$ denote the number of inversions in σ. Then*

$$\sum_{\sigma \in S_n} q^{i(\sigma)} \;=\; (1+q)(1+q+q^2)...(1+q+q^2+...+q^{n-1})$$

Proofs (of Theorems 5.1.3 and 5.1.4) The proofs of both theorems are essentially the same and we give a common proof highlighting at appropriate places where they differ. Given $D \in \mathcal{D}(n,p)$, consider the random variable

$$Y = Y(b) = |\{(A,\sigma) : A \subseteq V, |A| = b, \sigma : [b] \to A\}|.$$

where σ is a linear ordering of the vertices of A. Let $T_i = (A_i, \sigma_i), A_i \subseteq V, |A_i| = b$ be the i-th ordered b-set. Define an indicator random variable Y_i which is set to 1 if σ_i is a topological ordering for $D[A_i]$, and zero otherwise. Then,

$$Y = Y(b) = \sum_{i=1}^{(n)_b} Y_i \;;\; \text{Also, for each } i,\;\; E[Y_i] = Pr[Y_i = 1] = (1-p)^{\binom{b}{2}}.$$

Hence, by linearity of expectation, $E[Y] = \sum_{i=1}^{(n)_b} E[Y_i] = (n)_b(1-p)^{\binom{b}{2}}$. Define

$$b^* \;=\; \left\lfloor \frac{2\ln np}{\ln q} - X \right\rfloor \;=\; \frac{2\ln np}{\ln q} - X - \delta$$

where (i) $X = 1$ if $p \geq n^{-1/3+\epsilon}$ and $X = W/(\ln q)$ if $p \geq n^{-1/2}(\ln n)^2$ and (ii) δ, $0 \leq \delta < 1$, is defined to be the fractional part of the expression $2\log_q np - X$. We first prove that the first moment at b^* goes to infinity as $n \to \infty$.

Lemma 5.4.2 *At $b = b^*$, $E[Y] \to \infty$ as $n \to \infty$.*

Proof : The proof is by substituting the value of b^* in $E[Y(b^*)]$:

$$E[Y] \;=\; (n)_b(1-p)^{\binom{b}{2}} \;\geq\; (n-b)^b(1-p)^{\binom{b}{2}} \;=\; n^b(1-b/n)^b(1-p)^{\binom{b}{2}}$$

But for $p = \omega(n^{-1/2}(\ln n))$, $b/n \leq \frac{2(\ln np)}{n(\ln q)} \leq \frac{2(\ln np)}{np}$ and hence $b^2/n \leq \frac{4(\ln np)^2}{np^2} = o(1)$. Hence, $b^2/n \to 0$ as $n \to \infty$, so that $(1-b/n)^b \to 1$ as $n \to \infty$. Hence, as $n \to \infty$,

$$E[Y] \;\approx\; n^b(1-p)^{\binom{b}{2}} \;= (n(1-p)^{(b-1)/2})^b \;\geq (n(1-p)^{\ln np/\ln q})^b \;= (n/np)^b \;\geq p^{-\ln np/\ln q}$$

Since $p = \omega(n^{-1/2}(\ln n))$, $np = \omega(1)$, whereas $\ln q \leq \ln 2$. Therefore, $\ln np/\ln q = \omega(1)$ and so, $E[Y] \to \infty$ as $n \to \infty$. $\square$

Now we continue with the proof of the main theorems. It is based on the Second Moment method. By applying Chebyshev's Inequality to the random variable Y, we have

$$Pr[Y = 0] \;\leq\; Var(Y)/E[Y]^2 \;=\; (E[Y^2] - E[Y]^2)/E[Y]^2 \tag{5.3}$$

But $Y = \sum_i Y_i$ is the sum of indicator variables and hence from standard arguments, e.g. Proposition 2.2.3,

$$Var(Y) \;=\; \sum_i Var(Y_i) + \sum_{i \neq j} Cov(Y_i, Y_j) \;\leq\; E[Y] + \sum_{i \neq j} Cov(Y_i, Y_j) \tag{5.4}$$

where the sum is over ordered pairs (i, j) and $Cov(Y_i, Y_j) = E[Y_i Y_j] - E[Y_i]E[Y_j]$ denotes the covariance of the random variables Y_i and Y_j. Clearly, if $|A_i \cap A_j| < 2$, then Y_i and Y_j are independent and hence $Cov(Y_i, Y_j) = 0$. On the other hand, even if $|A_i \cap A_j| = b$, Y_i and Y_j could still be different random variables having non-zero covariance, since the permutations σ_i and σ_j could differ. Hence, only the pairs (i, j) for which $2 \leq |A_i \cap A_j| \leq b$ are of interest. Now,

$$\begin{aligned}
\sum_{i \neq j} Cov(Y_i, Y_j) \;&\leq\; \sum_{i \neq j} E[Y_i Y_j] \;=\; \sum_{i \neq j} E[Y_i] \cdot E[Y_j | Y_i = 1] \\
&=\; \sum_i E[Y_i] \left(\sum_{j : 2 \leq |A_i \cap A_j| \leq b} E[Y_j | Y_i = 1] \right) \\
&\leq\; \sum_i E[Y_i] E[Y].M \;=\; E[Y]^2.M
\end{aligned} \tag{5.5}$$

where M denotes $max_i \sum_{j : 2 \leq |A_i \cap A_j| \leq b} E[Y_j | Y_i = 1]/E[Y]$.

If it can be shown that $M = o(1)$, this implies that $Var(Y) \leq E[Y] + o\left(E[Y]^2\right)$ and hence that $\mathbf{Pr}(Y = 0) \leq (E[Y])^{-1} + o(1) = o(1)$ since $E[Y] \to \infty$. This establishes that with high probability, $Y = Y(b^*) > 0$ and hence $mas(D) \geq b^*$. Hence, it suffices to only show that $M = o(1)$.

We first find a combinatorial expression for M.

Lemma 5.4.3

$$M \;=\; \sum_{l=2}^{b} \frac{\binom{b}{l}\binom{n-b}{b-l}}{\binom{n}{b} l!} (1-p)^{-\binom{l}{2}} \prod_{i=1}^{l} \left(\frac{1 - r^i}{1 - r} \right)$$

where $r = (1 - 2p)/(1 - p)$.

Proof For T_i, T_j, $1 \leq i, j \leq (n)_b$, define $A_{i,j} := A_i \cap A_j$. For $u, v \in A_j$, if $\sigma_i^{-1}(u) < \sigma_i^{-1}(v)$, we say $u <_i v$. Similarly if $\sigma_j^{-1}(u) < \sigma_j^{-1}(v)$, say $u <_j v$. Since we are given $Y_i = 1$, we can treat σ_i (restricted to the set $A_{i,j}$) as the natural ordering induced by the vertex labelling, i.e. $u <_i v$ if and only if $u < v$. Let σ denote the restriction of σ_j to $A_{i,j}$. Given $u, v \in A_j$, suppose $u <_j v$. Then there are only 3 possibilities:

Case (i) At least one of $u, v \notin A_{i,j}$. In this case, the probability of not having the arc (v, u) is independent of the fact that $Y_i = 1$. Hence

$$\mathbf{Pr}[(v,u) \notin D[A_j] | Y_i = 1] = \mathbf{Pr}[(v,u) \notin D[A_j]] = 1 - p$$

The number of such pairs is clearly $\binom{b}{2} - \binom{l}{2}$.

Case (ii) Both $u, v \in A_{i,j}$ and $v <_i u$. Here, $Y_i = 1$ implies that $(u, v) \notin D[A_j]$. For $Y_j = 1$, we need $(v, u) \notin D[A_j]$. Hence the conditional probability becomes:

$$\mathbf{Pr}[(v,u) \notin D[A_j] | (u,v) \notin D[A_j]] = \frac{1-2p}{1-p}$$

The number of such pairs is clearly $i(\sigma)$ ($i(\sigma)$ defined as in Lemma 5.4.1) i.e. the number of inversions of σ_j restricted to $A_{i,j}$.

Case (iii) Both $u, v \in A_{i,j}$, and $u <_i v$. In this case, $Y_i = 1$ implies that $(v, u) \notin D[A_j]$. Hence,

$$\mathbf{Pr}[(v,u) \notin D[A_j] | Y_i = 1] = \mathbf{Pr}[(v,u) \notin D[A_j] | (v,u) \notin D[A_j]] = 1$$

For a fixed σ_i and σ_j, the number of such pairs is all the remaining pairs (from cases (i) and (ii)) i.e. $\binom{l}{2} - i(\sigma)$.

Since the event $Y_j = 1$ depends only on the presence or absence of the arcs (v, u), $u <_j v \in A_j$ and each arc is chosen independently of all the other arcs, the net probability expression obtained is

$$\begin{aligned}
E[Y_j | Y_i = 1] &= \mathbf{Pr}[Y_j = 1 | Y_i = 1] \\
&= \prod_{u <_j v \,;\, u,v \in A_j} \mathbf{Pr}[\, (v,u) \notin D[A_j] \mid Y_i = 1 \,] \\
&= \left(\prod_{Case(i)} (1-p) \right) \cdot \left(\prod_{Case(ii)} \frac{1-2p}{1-p} \right) \cdot \left(\prod_{Case(iii)} 1 \right) \\
&= (1-p)^{\binom{b}{2} - \binom{l}{2}} \cdot \left(\frac{1-2p}{1-p} \right)^{i(\sigma)} \cdot 1^{\binom{l}{2} - i(\sigma)}
\end{aligned}$$

As a result, we have

$$E[Y_j|Y_i=1] \;=\; Pr[Y_j=1|Y_i=1] \;=\; (1-p)^{\binom{b}{2}-\binom{l}{2}}\left(\frac{1-2p}{1-p}\right)^{i(\sigma)}$$

M is therefore:

$$M \;=\; max_i \sum_{2\le l\le b}\sum_{j:|A_i\cap A_j|=l} \mathbf{Pr}[Y_j=1|Y_i=1]/E[Y]$$

Once A_i is fixed, A_j such that $|A_i \cap A_j| = l$, can be chosen in $\binom{n-b}{b-l}\binom{b}{l}$ ways. Let S_k be the group of all permutations of a k-element set. Given A_i, A_j and a permutation σ over $A_{i,j}$, a permutation π over A_j (whose restriction to $A_i \cap A_j$ is σ) can be chosen in $b!/l!$ ways. Thus,

$$\begin{aligned} M &= \sum_{l=2}^{b}(n)_b^{-1}\binom{b}{l}\binom{n-b}{b-l}(b!/l!)(1-p)^{-\binom{l}{2}}\sum_{\sigma\in S_l}\left(\frac{1-2p}{1-p}\right)^{i(\sigma)} \\ &= \sum_{l=2}^{b}\binom{n}{b}^{-1}\binom{b}{l}\binom{n-b}{b-l}((1-p)^{-\binom{l}{2}}/l!)\sum_{\sigma\in S_l}\left(\frac{1-2p}{1-p}\right)^{i(\sigma)} \end{aligned}$$

The inner sum is $\sum_{\sigma\in S_l} r^{i(\sigma)}$ (r as defined in the statement of this lemma). Therefore applying Lemma 5.4.1 to the above expression, we get the required expression for M. □

Thus by the above lemma, we get

$$\begin{aligned} M &= \sum_{l=2}^{b}\binom{b}{l}^2\frac{(n-b)_{b-l}}{(n)_b}(1-p)^{-\binom{l}{2}}\prod_{i=1}^{l}\left(\frac{1-r^i}{1-r}\right) \\ &\le \sum_{l=2}^{b}\binom{b}{l}^2\frac{(1-l/n)^{-l}}{n^l}(1-p)^{-\binom{l}{2}}\prod_{i=1}^{l}\left(\frac{1-r^i}{1-r}\right) \end{aligned}$$

Since $l \le b$, we have $(1-l/n)^{-l} \le (1-b/n)^{-b} \to 1$ as $n \to \infty$.

$$\begin{aligned} M &\le \sum_{l=2}^{b}\binom{b}{l}^2 n^{-l}(1-p)^{-\binom{l}{2}}\prod_{i=1}^{l}\left(\frac{1-r^i}{1-r}\right) \\ &= \sum_{l=2}^{b}(1-r)^{-l}\binom{b}{l}^2 n^{-l}(1-p)^{-\binom{l}{2}}\prod_{i=1}^{l}(1-r^i) \;=\; \sum_{l=2}^{b}A_l \end{aligned}$$

We bound the value of A_l in the following two ways:

$$A_l \;\le\; \binom{b}{l}^2\left(\frac{1}{np(1-p)^{l/2}}\right)^l,$$

by substituting the value of r in terms of p and simplifying, and,

$$A_l \;\leq\; \binom{b}{l}^2 l! \left(\frac{1}{n(1-p)^{l/2}}\right)^l, \text{ since } r < 1.$$

Case I : $2 \leq l \leq t$, where $t = 2/\ln q$:

$$A_l \;\leq\; \binom{b}{l}^2 l! \left(\frac{1}{n(1-p)^{l/2}}\right)^l \;\leq\; \left(\frac{b^2}{n(1-p)^{l/2}}\right)^l \;\leq\; \left(\frac{b^2}{n(1-p)^{t/2}}\right)^l \;\leq\; \left(\frac{b^2 e}{n}\right)^l \;=\; o(1)$$

since $b^2 e/n \leq 1/(\log n)^2 = o(1)$.

Therefore, $\sum_{l=2}^{t} A_l \leq \sum_{l=2}^{\infty} (b^2 e/n)^l = s^2/(1-s) \leq s^2(1+2s) = o(1)$ where $s = b^2 e/n = o(1)$.

Case II: $t < l \leq b/2$:

$$\begin{aligned} A_l \;&\leq\; \binom{b}{l}^2 l! \left(\frac{1}{n(1-p)^{l/2}}\right)^l \;\leq\; \left(\frac{eb^2}{nl(1-p)^{l/2}}\right)^l \;\leq\; \left(\frac{eb^2}{nt(1-p)^{l/2}}\right)^l \\ &\leq\; \left(\frac{eb(\ln np)}{n(1-p)^{b/4}}\right)^l \;\leq\; \left(\frac{eb(\ln np)\sqrt{np}}{n}\right)^l \;\leq\; \left(\frac{eb\sqrt{p}(\ln np)}{\sqrt{n}}\right)^l. \end{aligned}$$

Now $b\sqrt{p}(\ln np) = O(p^{-1/2}(\ln n)^2) = O\left(n^{1/4}(\ln n)\right)$ because of our assumption about p. Hence, the summation can be upper-bounded by $(b/2)\left(\frac{eb\sqrt{p}(\ln np)}{\sqrt{n}}\right)^t$, which is less than $(b/2)(n^{-1/8})^t$, and is clearly $o(1)$.

Until now the proof arguments for both Theorems 5.1.3 and 5.1.4 are the same. The proofs vary for the remaining cases. First, we complete the

Proof of Theorem 5.1.3 : Our assumption is $p = \omega(n^{-1/2}(\ln n))$.
Case III : $b/2 < l \leq b$:

$$\begin{aligned} A_l \;&\leq\; \binom{b}{l}^2 \left(\frac{1}{np(1-p)^{l/2}}\right)^l \;\leq\; \left(\frac{e^2 b^2}{npl^2(1-p)^{l/2}}\right)^l \;\leq\; \left(\frac{4e^2}{np(1-p)^{b/2}}\right)^l \\ &\leq\; \left(\frac{4e^2(1-p)^{\delta/2}}{np(1/np)(1-p)^{-X/2}}\right)^l \;\leq\; \left(\frac{4e^2}{e^{W/2}}\right)^l \end{aligned}$$

for $W \geq 8$. The summation can therefore be upper-bounded by $(b/2)\left(4/e^2\right)^{-b/2}$, which is again $o(1)$, as $\lim_{n\to\infty} b = \infty$. This establishes that $M = o(1)$. So, we conclude that the probability that there is no topologically ordered set of size $b = \left\lfloor 2\log_q np - W/(\ln q) \right\rfloor$, goes to zero for a suitably chosen constant W (in fact $W = 8$ suffices). This proves the lower bound and hence completes the proof of Theorem 5.1.3. □

Proof of Theorem 5.1.4 : The proof is along the same lines and the only difference occurs when $l > b/2$. Case III (in the previous proof) gets split into two subcases :

Case IIIa : $b/2 < l \leq b - c'$, where $c' = 8/\ln q$:

$$\begin{aligned} A_l &\leq \binom{b}{l}^2 \left(\frac{1}{np(1-p)^{l/2}}\right)^l \leq \left(\frac{e^2 b^2}{npl^2(1-p)^{l/2}}\right)^l \leq \left(\frac{4e^2}{np(1-p)^{(b-c')/2}}\right)^l \\ &\leq \left(\frac{4e^2(1-p)^{(X+\delta)/2}}{np(1/np)(1-p)^{-c'/2}}\right)^l \leq \left(\frac{4e^2}{e^4}\right)^l \end{aligned}$$

Again the summation is upper bounded by $(b/2)\left(4/e^2\right)^{-b/2}$, which is $o(1)$.

Case IIIb: $b - c' < l \leq b$: Here we look at the ratio of successive terms:

$$\begin{aligned} A_{l+1}/A_l &= \left(\frac{b-l}{l+1}\right)^2 \left(\frac{1}{n(1-p)^l}\right)\left(\frac{1-r^{l+1}}{1-r}\right) \\ &\geq \frac{1}{2b^2}\left(\frac{(1-p)^{1+X+\delta}}{np(1/np)^2(1-p)^{-c'}}\right) \geq \left(\frac{np(1-p)^{c'+1+X+\delta}}{2b^2}\right) = \Omega\left(np^3(\ln np)^{-2}\right) \end{aligned}$$

For $p \geq n^{-1/3+\epsilon}$, $np^3(\ln np)^{-2} = \omega(1)$. Hence, the ratio $A_{l+1}^{l+1}/A_l \geq 1$ in the stated range of l. So the function A_l is increasing in the range $b - c' \leq l \leq b$, and the maximum value is therefore attained at $l = b$. For this value of l (using the assumption $X = 1$) :

$$\begin{aligned} A_b &\leq \binom{b}{b}^2 \left(\frac{1}{np(1-p)^{b/2}}\right)^b = \left((1-p)^{(1+\delta)/2}\right)^b = O\left((1-p)^{(\ln np)/(\ln q)}\right) \\ &= O\left(e^{-(\ln np)}\right) = O\left((np)^{-1}\right) = o(n^{-2/3}) \end{aligned}$$

Therefore, the summation is upper-bounded as follows:

$$\sum_{b-c'}^{b} A_l \leq c' A_b = o\left(p^{-1}n^{-2/3}\right) = o(1).$$

This establishes that $M = o(1)$ and we conclude that the probability that there is no topologically ordered set of size $b = \lfloor 2\log_q np - 1 \rfloor$, goes to zero. This proves the lower bound and hence completes the proof of Theorem 5.1.4. □

5.5 Proof of Theorem 5.1.5

When $p = o(n^{-1/2}\ln n)$, the variance of the number of topologically ordered sets shoots up when p is in this range, and so the second moment method fails. However, as Theorem 5.1.2

shows, the concentration of $mas(D)$ is quite sharp. To utilize this fact and overcome the problem of large variance, we use a technique which Alan Frieze first used for independent sets in random graphs (see [33], [40]). The technique involves combining a strong concentration inequality (such as Azuma's, Talagrand's etc.) which gives sharp concentration around an unknown value with a weak lower bound on the random variable in question - (here, $mas(D)$) - around a known value, to get past the "large variance" barrier. For the sake of simplicity, we sometimes treat some real valued expressions as sizes of induced dags without worrying about their possibly non-integer values. The proof arguments can be easily seen to carry over when we approximate these values by appropriate nearest integers.
We use the well-known Paley-Zygmund inequality to get a lower bound on $\mathbf{Pr}[Y(b) > 0]$:

$$\mathbf{Pr}[Y(b) > 0] \geq E[Y(b)]^2/E[Y(b)^2]$$

Specifically, we show

Lemma 5.5.1 *There exist positive constants C, W and c, such that if $p \geq C/n$ and $b \approx \frac{2}{\ln q}(\ln w - W)$, then*

$$\mathbf{Pr}[Y(b) > 0] \geq exp\left(-\frac{c(\ln w)^2 n}{(w)^{3/2}}\right)$$

Given Lemma 5.5.1, Theorem 5.1.5 follows easily, as shown below: Let $b^* = \frac{2}{\ln q}(\ln w - W')$, where $W' = 2W$, b and W as in Lemma 5.5.1. Now, by Theorem 2.3.3,

$$\mathbf{Pr}[mas(D) < b^*]\mathbf{Pr}[mas(D) \geq b] \;\leq\; exp(-(b-b^*)^2/4b) \;=\; exp(-\Omega\left(p^{-1}/\ln w\right))$$

Therefore, applying Lemma 5.5.1,

$$\begin{aligned}\mathbf{Pr}[mas(D) < b^*] \;&\leq\; \frac{exp(-\Omega(p^{-1}/\ln w))}{\mathbf{Pr}[mas(D) \geq b]} \;\leq\; exp(-\Omega(p^{-1}/\ln w) + cn(\ln^2 w)/(w)^{3/2}) \\ &=\; exp\left(-p^{-1}\Omega\left(\frac{1}{\ln w} - \frac{c(\ln^2 w)}{\sqrt{w}}\right)\right) \;=\; o(1)\end{aligned}$$

since $p \geq C/n$. □

The proof of Lemma 5.5.1 follows from standard calculations and arguments as briefly shown below.

Proof of Lemma 5.5.1 To get a lower bound on the RHS, i.e. $E[Y(b)]^2/E[Y(b)^2]$, it suffices to get an upper bound on the reciprocal, i.e. $E[Y(b)^2]/E[Y(b)]^2$. Recall from Section

5.4 the meanings of Y and $Y_i, 1 \leq i \leq m = (n)_b$. We have (as derived for $Var(Y)$ in the previous section) that

$$E[Y^2] \leq E[Y] + \sum_i E[Y_i]E[Y].M = E[Y] + E[Y]^2.M \quad (5.6)$$

where M denotes $max_i \sum_{j:0\leq|A_i\cap A_j|\leq b} E[Y_j|Y_i = 1]/E[Y]$ (here the maximization can be removed since the maximized expression is actually the same for all i as the random digraph model $\mathcal{D}(n,p)$ is homogenous). Note that M is defined essentially in the same way as in previous section except that $j \in \{0,1,\ldots,b\}$.[2]
Therefore,

$$E[Y^2]/E[Y]^2 = E[Y]^{-1} + M$$

Hence, the lower bound can be obtained by showing that $max\{E[Y]^{-1}, M\}$ is at most $exp(O(n\ln^2 w/(w)^{3/2}))$. Now, at $b = \frac{2}{\ln q}(\ln w - W)$, it is easy to check (by choosing W sufficiently large) that $E[Y] = \omega(1)$, and hence $E[Y]^{-1} = o(1)$. Therefore, the asymptotics of $E[Y^2]/E[Y]^2$ is mainly governed by the quantity M. We use the expression for M proved in Section 5.4.

Lemma 5.5.2 *(Section 5.4)*

$$M = \sum_{l=0}^{b} \frac{\binom{b}{l}\binom{n-b}{b-l}}{\binom{n}{b}l!}(1-p)^{-\binom{l}{2}} \prod_{i=1}^{l}\left(\frac{1-r^i}{1-r}\right)$$

where $r = (1-2p)/(1-p)$.

Write $M \leq \sum_{l=0}^{b} u_l.T_l$, where $u_l = \binom{b}{l}\binom{n-b}{b-l}(1-p)^{-\binom{l}{2}}/\binom{n}{b}$ and $T_l = \prod_{i=1}^{l}\left(\frac{1-r^i}{1-r}\right)/l!$
Notice that $T_l = \prod_{i=1}^{l}(1 + r + \ldots + r^{i-1})/i \leq 1$ for every l. Also,

$$u_l = \left(\frac{\binom{b}{l}\binom{n-b}{b-l}}{\binom{n}{b}}\right) e^{\binom{l}{2}\ln q} \leq \left(\frac{b^2 e}{nl} e^{(l-1)\ln q/2}\right)^l$$

We split the summation into three parts:

Case I : $0 \leq l \leq b/2$:

$$e^{(l-1)\ln q/2} \leq e^{b\ln q/4} \leq e^{\ln w/2}$$
$$\text{Hence, } u_l \leq \left(\frac{b^2 e}{nl}\sqrt{w}\right)^l \leq e^{\frac{b^2\sqrt{w}}{n}} \leq e^{c(\ln^2 w)n/w^{3/2}}$$

Here, we used the standard fact (which was also used in [33]) that $(A/x)^x$ is maximized when $x = A/e$ and hence $(A/x)^x \leq exp(A/e)$ (easily derivable using elementary calculus).

[2]There is a slight difference: the expression in Section 5.4 concerns the variance, so the summation is from $l = 2$ to b, as pairs of b-sets with intersection sizes zero and 1 do not contribute to the variance. Here, however, the sum is from $l = 0$ to b.

Case II: $b/2 < l \leq \frac{2}{\ln q}(\ln w - \ln\ln w - 3) + 1$:

$$u_l \;\leq\; \left(2ebe^{(l-1)\ln q/2}/n\right)^l \;\leq\; \left(\frac{4e\ln w}{w}e^{(l-1)\ln q/2}\right)^l \;\leq\; \left(\frac{4e\ln w}{w}\frac{w}{e^3\ln w}\right)^l \;<\; 1$$

Case III: $\frac{2}{\ln q}(\ln w - \ln\ln w - 3) + 1 < l \leq b$: For $l < b$,

$$\begin{aligned}\frac{u_l}{u_{l+1}} &= \frac{\binom{b}{l}\binom{n-b}{b-l}e^{\binom{l}{2}\ln q}}{\binom{b}{l+1}\binom{n-b}{b-l-1}e^{\binom{l+1}{2}\ln q}} = \frac{(l+1)(n-2b+l+1)}{(b-l)^2}e^{-l\ln q} \leq \frac{bn}{(b-l)^2}e^{-l\ln q} \\ &\leq \frac{bn}{(b-l)^2}e^6\ln^2 w/w^2\end{aligned}$$

Therefore,

$$\begin{aligned}u_l/u_b &\leq \left(\frac{1}{(b-l)!}\right)^2\left(bne^6\ln^2 w/w^2\right)^{b-l} \leq \left(\frac{bne^8\ln^2 w}{(b-l)^2w^2}\right)^{b-l} \leq e^{2\left(\frac{\sqrt{bn}e^3\ln w}{w}\right)} \\ &\leq e^{2\sqrt{2}e^3n((\ln w)/w)^{3/2}}\end{aligned}$$

For the third inequality, another well-known fact $(A/x^2)^x \leq exp(2\sqrt{A}/e)$ was used. Now,

$$\begin{aligned}u_b &= (1-p)^{-\binom{b}{2}}/\binom{n}{b} \leq e^{\ln q\binom{b}{2}}/\binom{n}{b} \leq \left(be^{(b-1)\ln q/2}/n\right)^b \\ &\leq \left(bpe^{-W}\right)^b = e^{b(\ln bp - W)}\end{aligned}$$

Also, for l in this range,

$$T_l \;=\; \frac{1}{l!}\prod_{j=1}^{l}\frac{(1-r^j)}{1-r} \;=\; \left((1-r)^{-l}/l!\right)\prod_{j=1}^{l}(1-r^j) \;\leq\; \left(\frac{e(1-p))}{lp}\right)^l \;=\; e^{l(1-\ln q-\ln(lp))}$$

Therefore

$$\begin{aligned}u_lT_l &\leq exp\left(\frac{2\sqrt{2}e^3\log^{3/2}w}{w^{3/2}}n + l(1-\ln q-\ln lp) + b(\ln bp - W)\right) \\ &= exp\left(\frac{2\sqrt{2}e^3\log^{3/2}w}{w^{3/2}}n + l(1-\ln q) - Wb + l\ln(bp/lp) + (b-l)\ln(bp)\right) \\ &< exp\left(\frac{2\sqrt{2}e^3\log^{3/2}w}{w^{3/2}}n + l(1+\ln 2-\ln q) - Wb + \frac{2\ln\ln w}{\ln q}\ln(bp)\right) \\ &< exp\left(\frac{2\sqrt{2}e^3\log^{3/2}w}{w^{3/2}}n\right)\end{aligned}$$

by adjusting the constant W so that $Wb > l(1 + \ln 2 - \ln q) + \frac{2(\ln\ln w)\ln(bp)}{\ln q}$. (In fact $W = 32$ suffices, though we have not made any attempt to obtain the best possible constant, for the sake of ease of presentation).
Thus, from the above three cases, we have that $M = exp(O(\frac{p^{-1}\ln^2 w}{\sqrt{w}}))$. Therefore,

$$\mathbf{Pr}[Y > 0] \;\geq\; E[Y]^2/E[Y^2] \;=\; 1/(M + o(1)) \;\geq\; exp(-\frac{cp^{-1}\ln^2 w}{\sqrt{w}})$$

where c is a fixed constant. This proves the Lemma 5.5.1 and completes the proof of Theorem 5.1.5. □

Conclusions

The problem of determining $mas(D)$, the size of the largest induced acyclic subgraph in a random directed graph $D = (V, E)$, was studied. The range of the concentration of $mas(D)$ was reduced from the previously known $O(p^{-1}\ln\ln np)$ to $O(\sqrt{\beta(n)\log_q np})$ for all $p = p(n) = \Omega(n^{-1})$, by applying Talagrand's inequality. Using the Second Moment method, the known lower bound was improved from $\frac{2}{\ln q}(\ln np - \ln\ln np - O(1))$ to $\frac{2}{\ln q}(\ln np - O(1))$ for all $p \geq C/n$, and in particular, to $\frac{2\ln np}{\ln q} - 1$, for $p \geq n^{-1/3+\epsilon}$ for any constant ϵ. Thus, for "nearly every" $p = p(n)$, the concentration band was improved to $O(\min\{p^{-1}, \sqrt{\beta(n)\log_q np}\})$.

Chapter 6

Largest Induced Acyclic Subgraphs in Random Digraphs: Upper Bounds and Algorithms

6.1 Introduction

In the previous chapter on largest induced acyclic subgraphs in random digraphs, we defined the digraph invariant $mas(D)$ and obtained some concentration results and lower bounds on it. In the current chapter, we shall initially focus on obtaining improved upper bounds. Next, we shall also consider the algorithmic question of *finding* efficiently a maximum-sized induced acyclic subgraph in a given n-vertex digraph $D = (V, E)$, drawn according to the distribution $\mathcal{D}(n, p)$. Finally, we shall briefly state the analogues of the theorems derived in the previous chapter and this chapter, for the case when the random digraph D is drawn using the model $\mathcal{D}_2(n, p)$.

In Section 6.2, using known upper bounds on the number of acyclic orientations of an undirected graph, we obtain the following slight improvement (not asymptotic) on the upper bound of Theorem 5.1.1.

Theorem 6.1.1 *Let $D \in \mathcal{D}(n, p)$. If p satisfies $n^{-1/2+\epsilon} \leq p \leq 0.5$ where $\epsilon > 0$ is any constant, then a.a.s (with $q = (1-p)^{-1}$)*

$$mas(D) \quad \leq \quad \left\lceil \frac{2}{\ln q}(\ln w + \ln(7e)) \right\rceil$$

To compare the bound in Theorem 6.1.1 with that in 5.1.1, notice that $2\ln(7e) \approx 5.9$, whereas $6e \approx 16.31$. Hence Theorem 6.1.1 gives an additive improvement of around $10.4/(\ln q)$ in

the upper bound. This method also indicates why we are unlikely to get any asymptotic improvement in the upper bound, by using the first moment method. Using a very different idea, we shall see this intuition rigorously proved later.

Theorems 5.1.2, 5.1.3, 5.1.4, 5.1.5 and 6.1.1 also carry over to a related model $\mathcal{D}_2(n,p)$ in which each possible directed arc exists independently with probability p. The theorems and a sketch of the corresponding proofs are given in Section 6.5.

6.1.1 The algorithmic aspects

By MAS(D,k), we denote the following computational problem : Given a simple directed graph $D=(V,A)$ and k, determine if $mas(D) \geq k$. The problem, as stated above, is a "decision version", having a Yes/No answer. The optimization version, MAS(D), requires us to find a maximum-sized induced acyclic digraph (size being the number of vertices).

However, MAS(D,k) is known to be NP-complete [36]. In fact, even finding an approximate solution to the optimization version MAS(D) is known to be hard [49] when the input is an arbitrary digraph: for some $\epsilon > 0$, a polynomial-time approximation algorithm with an approximation ratio of $O(n^{\epsilon})$ is not possible unless $P = NP$.

If we focus on random digraphs drawn from $\mathcal{D}(n,p)$, it was shown in [63] that a greedily built solution is of size at least $\epsilon(\log_q np)$ (for $p = \Omega(n^{-1})$), for every fixed $\epsilon < 1$. It was also predicted in [63] that ϵ can be made to approach 1 asymptotically. In Section 6.4, we improve this algorithmic result further by studying a heuristic which combines greedy and brute-force approaches as follows. We first apply the greedy heuristic to get a partial solution whose size is nearly $\log_q np - c\sqrt{\log_q np}$ for some arbitrary constant c. Then, in the subgraph induced by those vertices each of which can be safely added to the partial solution, we find an optimal solution by brute-force and combine it with the partial solution. It is shown that (for every fixed p) this modified approach produces a solution whose size is at least $\log_q np + c\sqrt{\log_q np}$. This results in an additive improvement of $\Theta(\sqrt{\log_q np})$ over the simple greedy approach. The improvement is mainly due to the fact we stop using the greedy heuristic at a point where it is possible to apply brute-force efficiently. This approach is similar to (and was motivated by) the "expose-and-merge" approach used in [52, 44] for finding large independent sets in $\mathcal{G}(n,1/2)$.

6.2 Upper Bound: Acyclic Orientations

In this section, we prove Theorem 6.1.1. Throughout this section, we define b to be $b = \lceil 2(\ln w + \ln(7e))/(\ln q) + 1 \rceil$. Consider an undirected graph $G = (V, E)$. Let $ao(G)$ denote the number of *acyclic orientations* of G. An orientation O of G is obtained by directing each edge $\{i, j\} \in E(G)$ from i to j (indicated by (i, j)) or vice-versa to get the digraph $D = (V, A)$, $A = O(G)$. An orientation O is acyclic if the resulting digraph D has no directed cycles.

Let $a(m)$ be the maximum number of acyclic orientations of any simple undirected graph Y, where Y has b vertices and m edges. Then the probability that the induced digraph on a *fixed* b-set R of $V[D]$, $D \in \mathcal{D}(n, p)$ is a dag is upper-bounded by:

$$\begin{aligned} \mathbf{Pr}[D[R] \text{ is acyclic}] &= \mathbf{E}_{Y \in \mathcal{G}(b,2p)} \frac{ao(Y)}{2^{|E(Y)|}} \\ &\leq \sum_{m=0}^{\binom{b}{2}} \frac{a(m)}{2^m} \binom{\binom{b}{2}}{m} (2p)^m (1-2p)^{\binom{b}{2}-m} \end{aligned} \tag{6.1}$$

We use the simple upper bound $a(m) \leq (d+1)^b$, where $d = 2m/b$ is the average degree of any such graph, obtained in [50]. Let Z denote the number of edges in the random subgraph of $\mathcal{G}(n, 2p)$ induced by R. Since $G[R]$ is drawn according to the distribution $\mathcal{G}(b, 2p)$, Z has expectation given by $\mathbf{E}[Z] = \binom{b}{2} 2p$.

The variable Z is in fact binomial and using Chernoff-Hoeffding large-deviation bounds ([53], [9]), the probability of its being much larger than the expected value can be tightly bounded:

$$\begin{aligned} \mathbf{Pr}[Z > 3\mathbf{E}[Z]] &\leq (e^2 3^{-3})^{\mathbf{E}[Z]} < e^{-2b(p/\ln q)\ln np} < (np)^{-2(p/\ln q)b} \leq n^{-2b(p/\ln q)(1/2+\epsilon)} \\ &= n^{-b(p/\ln q)(1+2\epsilon)} \end{aligned}$$

For $0 \leq p \leq 1/2$, $0.7 \leq p/\ln q \leq 1$ [63]. Now, the sum in (8) can be broken into two parts, as $m \leq 3\mathbf{E}[Z]$ and $m > 3\mathbf{E}[Z]$:

$$\begin{aligned} \mathbf{Pr}[D[R] \text{ is acyclic }] \leq & \sum_{m \leq \lfloor 3\mathbf{E}[Z] \rfloor} a(m) \binom{\binom{b}{2}}{m} (p)^m (1-2p)^{\binom{b}{2}-m} \\ & + \sum_{m > \lfloor 3\mathbf{E}[Z] \rfloor + 1} \frac{a(m)}{2^m} \binom{\binom{b}{2}}{m} (2p)^m (1-2p)^{\binom{b}{2}-m} \end{aligned}$$

$$
\begin{aligned}
&\leq a(\lfloor 3\mathbf{E}[Z]\rfloor)(1-p)^{\binom{b}{2}} + \left(\frac{b!}{2^{3\mathbf{E}[Z]}}\right)\mathbf{Pr}[Z > 3\mathbf{E}[Z]] \\
&\leq (6bp+1)^b(1-p)^{\binom{b}{2}} + \frac{b!}{2^{3\mathbf{E}[Z]}}n^{-b(0.7)(1+2\epsilon)} \\
&\leq (6bp+1)^b(1-p)^{\binom{b}{2}} + (b!)n^{-b[(0.7)(1+2\epsilon)+(6\ln 2)(\ln np/\ln n)(p/\ln q)]}
\end{aligned}
$$

Since $(\ln np/\ln n) \geq (1/2+\epsilon)$ and $(p/\ln q) \geq 0.7$, on applying the union bound over all b-sized subsets of V, we see that the probability that there exists an acyclic b-set $A \subset V$ is at most

$$
\binom{n}{b}[(6bp+1)^b(1-p)^{\binom{b}{2}} + (b!)n^{-b((0.7)(1+2\epsilon)+0.7)}]
$$

Let P_b denote $\mathbf{Pr}[\exists R,\ |R| = b,\ D[R] \text{ is acyclic}]$.

$$
\begin{aligned}
P_b &\leq \binom{n}{b}[(6bp+1)^b(1-p)^{\binom{b}{2}} + (b!)n^{-b(1+\epsilon)}] \leq \binom{n}{b}(6bp+1)^b(1-p)^{\binom{b}{2}} + n^{-b\epsilon} \\
&= \binom{n}{b}(6bp+1)^b(1-p)^{\binom{b}{2}} + o(1) \leq \left((6enp + en/b)(1-p)^{(b-1)/2}\right)^b + o(1) \\
&\leq \left((6enp + en(\ln q)/(2\ln w))(1-p)^{(b-1)/2}\right)^b + o(1) \\
&\leq \left(((6+o(1))enp)(1-p)^{(b-1)/2}\right)^b + o(1)
\end{aligned}
$$

The first expression in the previous inequality goes to zero since the base (which is raised to its b-th power) is bounded by a constant less than 1 for our definition of b. As a result, we get that $mas(D) \leq 2\lceil \log_q w + \log_q(7e)\rceil$. We observe that the additive second order term of this expression is marginally better than the corresponding term in the expression obtained in [61], although the asymptotics of both terms are the same, that is, $mas(D) = \frac{2(\ln np)}{\ln q} + O(p^{-1})$.

6.3 Upper Bound: Layered Construction

The central difficulty in obtaining an accurate estimate of the upper bound on $mas(D)$ is that of obtaining an exact closed-form expression for the probability that a given set of b vertices induces an acyclic subgraph. In this section, we shall see a new approach to this basic problem: the *layered construction* , which will yield us an expression for the exact probability, whose asymptotics can be easily estimated.

The Layered Construction The central idea is very simple: given a directed acyclic graph **dag** $A = (V, E)$ on n labelled vertices, obtain a partition $\mathcal{P} = \mathcal{P}(A)$ of the vertex set $V(A)$, (called the *layers* of A), as follows: The vertices of indegree zero form the first part

$P_1 = P_1(A)$ (called layer 1), remove P_1 from $V(A)$, and recurse on the remaining dag $A \setminus P_1$ to obtain $P_2, P_3, \ldots, P_k$, (called respectively layer 2, layer 3 etc.) where $k \leq n$.

Claim 6.3.1 *Given a dag A, the partition $\mathcal{P}(A)$ is uniquely determined.*

Proof The proof is obvious since at each step of the partitioning process, all, and only those, vertices which have in-degree zero, go into the part being assigned. Hence given the dag A, the part $P_1(A)$ is uniquely determined. Recursively, the parts $P_2(A), P_3(A), \ldots, P_k(A)$ are also uniquely determined. □

Claim 6.3.2 *Given a dag A and the layers of A, i.e. $\mathcal{P}(A)$, there do not exist $i, j \in [k]$, such that $j > i$ and there is an arc from some vertex in layer j to some vertex in layer i.*

Proof Suppose there exist some $i, j \in [k]$ such that $i < j$ and layer i has a vertex u which has an incoming arc from layer j, then during the construction of layer i, vertex u would not have in-degree zero, since layer j was yet to be constructed. Hence by construction, u could not be in layer i. Thus we get a contradiction. □

Claim 6.3.3 *Given a dag A and $\mathcal{P}(A)$, the subgraph induced by vertices in any given layer is an independent set.*

Proof If there exists a layer i which does not induce an independent set, then some vertex in layer i had non-zero indegree in the subgraph induced by layers $i, i+1, \ldots, k$, which is a contradiction. □

Claim 6.3.4 *Given a dag A and the layers of A, for each $1 < i \leq k$, each vertex in layer i has at least one in-coming arc from layer $i-1$.*

Proof Suppose not, then we have i such that the layer i has a vertex v which has zero in-degree from layer $i-1$. Then, even during the construction of layer $i-1$ and by Claim 6.3.2, v would have zero in-degree. By construction, this implies $v \in P_{i-1}$, which is a contradiction, since $v \in P_i$. □

Thus, from Claims 6.3.1-6.3.4, we can get an exact expression for the probability P_{dag} that the subgraph induced by some b vertices forms a dag. Before proceeding further, we need a definition:

Definition A *k-composition* of a positive integer r is a solution, in positive integers, of the equation

$$x_1 + x_2 + \ldots + x_k = r$$

where $1 \leq k \leq r$.

Applying the notion of k-compositions, we derive an exact expression for the probability of an arbitrary b-set inducing an acyclic subgraph. It is based on the following procedure for forcing a b-set to induce a dag :

Choose some $k \in [b]$, choose a k-composition of b. Create layers of sizes according to the chosen composition, choose vertices to lie in each of the layers, force each layer to induce an independent set, ensure arcs are only from lower-indexed layers to higher-indexed layers, and each vertex in layer i has at least one in-coming arc from some vertex in layer $i-1$, where $i = 2, \ldots, k$.

Let the set of k-compositions of b be denoted by $C_k(b)$. From the above method, we get an exact expression for the probability P_{dag}:

$$\begin{aligned} P_{dag} &= \sum_{k=1}^{b} \sum_{(b_i) \in C_k(b)} \binom{b}{b_1, b_2, \ldots, b_k} (1-2p)^{\sum_{i=1}^{k} \binom{b_i}{2}} (1-p)^{\binom{b}{2} - \sum_{i=1}^{k} \binom{b_i}{2}} \cdot \\ &\quad \prod_{i=2}^{k} \left(1 - \left(\frac{1-2p}{1-p}\right)^{b_{i-1}}\right)^{b_i} \\ &= (1-p)^{\binom{b}{2}} \sum_{k=1}^{b} \sum_{C_k(b)} \binom{b}{b_1, b_2, \ldots, b_k} \left(\frac{1-2p}{1-p}\right)^{\sum_{i=1}^{k} \binom{b_i}{2}} \cdot \\ &\quad \prod_{i=2}^{k} \left(1 - \left(\frac{1-2p}{1-p}\right)^{b_{i-1}}\right)^{b_i} \end{aligned}$$

Now the first moment method can be applied to find an upper bound on $mas(D)$: Let $X = X(n,p,b)$ be the number of induced dags of size b in the random digraph $D \in \mathcal{D}(n,p)$.

$$\begin{aligned} \mathbf{Pr}[D \text{ has a dag of size } b] &= \mathbf{Pr}[X > 0] \leq \mathcal{E}[X] = \binom{n}{b} \cdot P_{dag} \\ &= \binom{n}{b} (1-p)^{\binom{b}{2}} \sum_{k=1}^{b} \sum_{C_k(b)} \binom{b}{b_1, b_2, \ldots, b_k} \cdot \\ &\quad \left(\frac{1-2p}{1-p}\right)^{\sum_{i=1}^{k} \binom{b_i}{2}} \prod_{i=2}^{k} \left(1 - \left(\frac{1-2p}{1-p}\right)^{b_{i-1}}\right)^{b_i} \end{aligned}$$

Let $k = \lfloor 2(\ln w + \ln a) \rfloor$ and $b = \lfloor k/(\ln q) \rfloor + 1$, where the range of $a \in \Re^+$ is be specified below.

Lemma 6.3.5 *For $p = p(n)$ such that $n^{-0.5}(\ln n)^2 \leq p$, $p = o((\ln n)^{-1})$ and for any constant $a < e^{1/2} - e^{-1/2}$, the following holds: $\mathcal{E}[X] \to \infty$ as $n \to \infty$.*

Hence, for p in the above range, an application of the first moment method to this random variable does not give an upper bound on $mas(D)$ which is less than $\frac{2(\ln w + x)}{\ln q}$ for some suitable positive constant x.

To keep the proof arguments simpler, we ignore floors and ceilings and also assume, without loss of generality, that k divides b. Without these assumptions, we only need to change the value of a to $a \pm o(1)$ and hence Lemma 6.3.5 still holds true. Focus on the term T of the expression for $\mathcal{E}[X]$ for which $b_1 = b_2 = \ldots = b_k = b/k$. We shall show that the term T goes to infinity as $n \to \infty$.

Proof Taking Stirling's approximation and using $(n)_b = n^b(1 - o(1))$ and using r to denote $\frac{p}{1-p}$, it follows that T is lower bounded by:

$$\begin{aligned}
T &\approx (n)_b \left(\frac{\sqrt{2\pi b/k}}{(b/ek)^{(b/k)}}\right)^k (1-p)^{\binom{b}{2}} (1-r)^{k.(b/k)(b/k-1)/2} \left(1 - (1-r)^{b/k}\right)^{b-b/k} \\
&\geq \left(\frac{n(1+o(1)).ek(1-p)^{(b-1)/2}(1-r)^{(b/k-1)/2}}{b}\right)^b \left(1 - (1-r)^{\frac{1}{\ln q}}\right)^b \\
&= \left(\frac{n(1+o(1)).e(\ln q)(1-r)^{(b/k-1)/2}}{npa}\right)^b \left(1 - (1-r)^{\frac{1}{\ln q}}\right)^b \\
&= \left(\left(\frac{e}{a}\right)(1+o(1)).(1-r)^{(b/k-1)/2}\left(1 - (1-r)^{\frac{1}{\ln q}}\right)\right)^b \\
&\geq \left(\left(\frac{e}{a}\right)(1+o(1))e^{((b/k-1)/2)\ln(1-r)}\left(1 - e^{\ln(1-r)\frac{1}{\ln q}}\right)\right)^b
\end{aligned}$$

Now, since $p = o((\ln n)^{-1})$, $\ln(1-r) = (1+o(1)).\ln(1-p)$, and so we get that the LHS

$$\begin{aligned}
&\geq \left(\left(\frac{e}{a}\right)(1+o(1))e^{-1/2+o(1)}\left(1 - e^{-(1+o(1))}\right)\right)^b \\
&\geq \left(\left(\frac{e^{1/2}}{a}\right)(1+o(1))e^{o(1)}\left(1 - e^{-(1+o(1))}\right)\right)^b \\
&\geq \left(\left(\frac{e^{1/2} - e^{-1/2-o(1)}}{a}\right)\right)^b \\
&\to \infty
\end{aligned}$$

for any positive $a < e^{1/2} - e^{-1/2}$.

□

6.4 An efficient heuristic with improved guarantee

It was shown in [63] that, for every fixed $\delta < 1$, with probability $1 - o(1)$, *every* maximal induced dag is of size is at least $\delta(\log_q np)$. A maximal solution can be obtained in linear time. It was also mentioned in [63] that one can possibly set $\delta = 1$. We further refine this analysis and show that the above statement holds for some $\delta = \delta(n) \to 1$. Precisely, we have the following strengthening of Theorem 3.1 of [63].

Theorem 6.4.1 *Let $p = p(n) \leq 0.5$ be such that $w = np \geq X$ for some sufficiently large constant $X > 0$. Then, for $D \in \mathcal{D}(n,p)$, with probability $1 - o(1)$,* every *maximal induced dag is of size at least $\delta(\log_q w)$ where $\delta = 1 - \frac{2(\ln\ln w)+10}{\ln w}$.*

The proof of this theorem follows by substituting the value of δ in the analysis given in [63].

We present below another efficient heuristic which will be analyzed and shown to have an additive improvement (for every fixed $p \leq 0.5$) of $\Theta\left(\sqrt{\log_q w}\right)$ over the guarantee given in [63] and in Theorem 6.4.1. It is similar to a heuristic presented in [44] for finding large independent sets in $G \in \mathcal{G}(n, 1/2)$. We show that, for every fixed $c > 0$, one can find in polynomial time an induced DAG of size at least $\lfloor \log_q w + c\sqrt{\log_q w} \rfloor$.

Let C be the set of those vertices which could be each individually added to the greedy solution. The idea is to construct greedily a solution A of size $g(n,p,c) = \lceil \log_q w - c\sqrt{\log_q w} \rceil$ and then add an optimal solution (found by an exhaustive search) in the subgraph induced by vertices in C. We will show that exhaustive search can be done in polynomial time and yields (a.a.s.) a solution of size $2c\sqrt{\log_q w}$. As a result, we finally get a solution of the stated size. The algorithm is described below.

MAXDAG$(D = (V,E), c)$

1. Choose and fix a linear ordering σ of V.
2. $c' := 1.2c$; $A := \emptyset$; $B := V$.
3. **while** $B > n/2$ **and** $|A| < g(n/2, p, c')$ **do**
4. Let u be the σ-smallest vertex in B.
5. **If** $D[A \cup \{u\}]$ induces an acyclic subgraph **then** add u to A.
6. remove u from B. **endwhile**
7. **if** $|A| < g(n/2, p, c')$, **then** Return FAIL and halt.

8. $C := \{u \in B : (u,v) \notin E,\ \forall v \in A\}$; $\mu = |B|(1-p)^{|A|}$.

9. **if** $|C| \notin [(0.9)\mu, (1.1)\mu]$ **then** Return FAIL.

10. **for each** $X \subset C$: $|X| = \left\lceil 2c'\sqrt{\log_q w/2} + 2\log_q 0.9 - W \right\rceil$ **do**

11. **if** $D[X]$ is acyclic **then** Return $D[A \cup X]$ and halt. **endfor**

12. Return FAIL.

We analyze the above algorithm and obtain the following result.

Theorem 6.4.2 *Let $p = p(n) \leq 0.5$ be such that $p \geq \tau$ for some fixed but arbitrary positive constant τ and let $D \in \mathcal{D}(n,p)$. Then, for every constant $c \geq 1$, with probability $1 - o(1)$, MAXDAG(D, c) will output an induced acyclic subgraph of size at least $b' = \lfloor (1+\epsilon')\log_q np \rfloor$, where $\epsilon' = c/\sqrt{\log_q np}$.*

Proof : Without loss of generality, we assume that c is sufficiently large.

Correctness : First, we prove the correctness. Note that $D[A]$ is always an induced acyclic subgraph. Also, each $u \in C$ is such that $D[A \cup \{u\}]$ is an acyclic subgraph with u as a sink vertex (having zero out-degree). Hence, any acyclic subgraph $D[X]$ present as a subgraph in $D[C]$ can be safely added to A so that $D[A \cup X]$ also induces an acyclic subgraph of D.

Time Complexity : It is easy to see that the running time is polynomial except for the **for** loop of lines 10 and 11. The maximum number of iterations of the **for** loop is at most

$$\begin{aligned}\binom{(1.1)\cdot|B|(1-p)^{|A|}}{\lfloor 2c'\sqrt{\log_q w/2}\rfloor} &\leq \binom{(2.2)\cdot p^{-1}q^{c'\sqrt{\log_q w/2}}}{\lfloor 2c'\sqrt{\log_q w/2}\rfloor} \leq q^{2c'^2(\log_q w/2)}\cdot((2.2)p^{-1})^{2c'\sqrt{\log_q w/2}} \\ &= O\left(n^{O(1)}\right),\end{aligned}$$

since p is a constant. Since each iteration takes polynomial time, the algorithm always finishes in polynomial time.

Analysis : Consider the following events defined as

$$\begin{aligned}\mathcal{E}_1 &: |A| < g(n/2,p,c')\ ; \\ \mathcal{E}_2 &: |C| \notin [(0.9)\mu, (1.1)\mu]\ ; \\ \mathcal{E}_3 &: mas(D[C]) < \left\lceil 2c'\sqrt{\log_q w/2} + 2\log_q 0.9 - W \right\rceil\ ;\end{aligned}$$

If none of these events holds, then the algorithm will succeed and output a solution whose size is

$$\begin{aligned}
|A \cup X| &\geq \log_q(w/2) - c'\sqrt{\log_q(w/2)} + 2c'\sqrt{\log_q w/2} + 2\log_q 0.9 - W \\
&\geq (1+\epsilon')(\log_q w) + (c'-c)\sqrt{\log_q w/2} + 2\log_q 0.9 - W - \log_q 2 \\
&\geq (1+\epsilon')(\log_q w) + (0.2c)\sqrt{\log_q w/2} + 2\log_q 0.9 - (W + \log_q 2) \\
&\geq (1+\epsilon')(\log_q w)
\end{aligned}$$

We have

$$\begin{aligned}
\mathbf{Pr}(\overline{\mathcal{E}_1}\ \overline{\mathcal{E}_2}\ \overline{\mathcal{E}_3}) &= \mathbf{Pr}(\overline{\mathcal{E}_1}) \cdot \mathbf{Pr}(\overline{\mathcal{E}_2} \mid \overline{\mathcal{E}_1}) \cdot \mathbf{Pr}(\overline{\mathcal{E}_3} \mid \overline{\mathcal{E}_1}\ \overline{\mathcal{E}_2}) \\
&\geq 1 - \sum_{i \leq 3} \mathbf{Pr}(\mathcal{E}_i \mid \wedge_{j<i}\, \overline{\mathcal{E}_j}) \qquad (6.2)
\end{aligned}$$

Let V_1 denote the set of first $n/2$ vertices of σ. Now using Theorem 6.4.1, the greedy algorithm run on the first $n/2$ vertices yields with probability $1 - o(1)$, an acyclic subgraph of size

$$\begin{aligned}
\delta(\log_q(w/2)) &\geq \log_q w/2 - 2(\log_q(\ln w/2)) - (10/\ln q) \ \geq\ g(n/2, p, c') \\
&= \lceil \log_q w/2 - c'\sqrt{\log_q w/2} \rceil,
\end{aligned}$$

with probability $1 - o(1)$. Here, δ is defined in Theorem 6.4.1. Hence, $\mathbf{Pr}(\mathcal{E}_1) = o(1)$.

For any fixed vertex $u \in B$,

$$\mathbf{Pr}(u \in C) \ =\ \mathbf{Pr}(\forall\, v \in A,\ (u,v) \notin E) \ =\ (1-p)^{|A|}.$$

Hence

$$\mu \ =\ E[|C|] \ =\ |B| \cdot (1-p)^{|A|}.$$

Since $|C|$ is the sum of $|B|$ identical and independent indicator random variables, by applying Chernoff-Hoeffding bounds (see [53, 9]), we get that

$$\mathbf{Pr}\left(|C| \notin [(0.9)\mu, (1.1)\mu]\right) \ \leq\ 2e^{-\mu/300}.$$

Since $|A| = g(n/2, p, c')$, we deduce that

$$\mu \ \approx\ |B| \cdot 2q^{c'\sqrt{\log_q w/2}}/w,$$

after justifiably ignoring the effect of the ceiling function used in the definition of $g(n/2, p, c')$. Given that $\overline{\mathcal{E}_1}$ holds and also since $|B| \geq n/2$, it is easy to verify that $\mu \to \infty$ as $n \to \infty$. Hence $\mathbf{Pr}(\mathcal{E}_2 \mid \overline{\mathcal{E}_1}) \ =\ o(1)$.

Given that neither of $\mathcal{E}_1$ and $\mathcal{E}_2$ holds, it follows that $|C| \geq (0.9)\mu \approx (0.9) \cdot p^{-1} \cdot q^{c'\sqrt{\log_q w/2}}$. Hence, using $q \leq 2$ and applying Theorem 5.1.4,

$$mas(D[C]) \geq \lfloor 2c'\sqrt{\log_q w/2} \rfloor + 2\log_q 0.9 - W \geq \lfloor 2c'\sqrt{\log_q w/2} \rfloor - 1$$

with probability $1 - o(1)$. This establishes that $\mathbf{Pr}(\mathcal{E}_3 \mid \overline{\mathcal{E}_1}\ \overline{\mathcal{E}_2}) = o(1)$. It then follows from (6.2) that MAXDAG(D, c) outputs a solution of required size with probability $1 - o(1)$. □

6.5 Bounds for the Non-simple Case

"What happens if the random digraph be allowed to have 2-cycles"? - This question is addressed in the current section. The model, which was introduced in [63], is as follows:

Model $D \in \mathcal{D}_2(n, p)$: Choose each directed edge $u \to v$ joining distinct elements of V independently with probability p.

Let $D \in \mathcal{D}_2(n, p)$. Using similar arguments as in the $\mathcal{D}(n, p)$ case, the following analogues of Theorems 5.1.2, 5.1.3, 5.1.4, 5.1.5 and 6.1.1 can be derived.

Theorem 6.5.1 *For $p \leq 1/2$, for any $\beta = \beta(n)$ such that $\beta \to \infty$ as $n \to \infty$, a.a.s*

$$|mas(D) - m| \leq \beta\sqrt{\log_q w}$$

where m is a median of $mas(D)$.

Theorem 6.5.2 *There is a large constant W such that if $n^{-1/2}(\ln^2 n)\beta(n) \leq p \leq 1/2$ (where $\beta(n) \to \infty$ as $n \to \infty$), then a.a.s.*

$$mas(D) \geq \left(\frac{1}{\ln q}\right)(2\ln np - W)$$

Theorem 6.5.3 *For every $\epsilon > 0$, the following is true : if $n^{-1/3+\epsilon} \leq p \leq 1/2$ then a.a.s.*

$$mas(D) \geq \left(\frac{2\ln np}{\ln q}\right) - 1$$

Theorem 6.5.4 *There exist suitable positive constants C and W such that if p satisfies $C/n \leq p \leq 1/2$, then a.a.s.*

$$mas(D) \geq \frac{2}{\ln q}(\ln np - W)$$

Theorem 6.5.5 *For every $\epsilon > 0$, if $n^{-1/2+\epsilon} \leq p \leq 1/2$ then a.a.s.*

$$mas(D) \leq \frac{2}{\ln q}(2\ln np + \ln(7e)) + 1$$

The Lower Bound: The analysis proceeds as in the $\mathcal{D}(n,p)$ case. The random variable $X = X_b$ is again the number of topologically ordered sets of size b, and the second moment method is used. The expression for $M = Cov(Y_i, Y_j)/E[Y]^2$ (where $Y = \sum_i Y_i$ and Y, Y_i are as in Section 5.4) is given by

$$M = \sum_{l=2}^{b} \binom{b}{l}^2 \binom{n-b}{b-l} (n)_b^{-1} (1-p)^{-\binom{l}{2}} \sum_{\sigma \in S_l} (1-p)^{i(\sigma)}$$

The proofs of theorems 6.5.2 and 6.5.3 can now be obtained using similar arguments as in Section 5.4.

The Upper Bound The upper bound given in Theorem 6.5.5 can be proved using essentially the same ideas as in Section 6.2. However, there are certain differences which need to be handled carefully. We use the notation of Section 6.2.

$$\begin{aligned} Pr[D[R] \text{ is acyclic }] &\leq \sum_{m=0}^{\binom{b}{2}} \frac{a(m)}{2^m} Pr[G[R] \text{ is simple and has } |E[R]| = m] \\ &\leq \sum_{m=0}^{\binom{b}{2}} \frac{a(m)}{2^m} \binom{\binom{b}{2}}{m} (2p(1-p))^m ((1-p)^2)^{\binom{b}{2}-m} \end{aligned} \tag{6.3}$$

Let Z be the number of edges in the random undirected graph $G[R]$, conditioned on $G[R]$ being simple. The average degree is therefore $d = 2Z/b$. Then, $E[Z] = \binom{b}{2} \frac{2p(1-p)}{1-p^2} = b(b-1)p/(1+p)$. As in Section 6.2, Chernoff bounds can be applied when $Z > 3E[Z]$, and the sum in (10) can be split into 2 parts - when $m \leq 3E[Z]$, and when $m > 3E[Z]$. Again, $a(m)$ is bounded simply by $(d+1)^b$ in the first case and $b!$ in the second case respectively. It is easy to verify that these bounds can be put together to yield the statement of Theorem 6.5.5.

Algorithmic Aspects Coming to the MAS(D,k) problem, the following theorems can easily be proved following the proofs in Section 6.4.

Theorem 6.5.6 *Let $p = p(n)$ be such that $w = np \geq X$ for some sufficiently large constant $X > 0$. Then, for $D \in \mathcal{D}_2(n,p)$, every maximal induced dag is of size at least $\delta(\log_q w)$ where $\delta = 1 - \frac{2\ln\ln w + 10}{\ln w}$.*

Theorem 6.5.7 *Let $p = p(n)$ be such that $p \geq \tau$ for some fixed but arbitrary constant τ and let $D \in \mathcal{D}_2(n,p)$. Then, for every constant $c \geq 1$, with probability $1 - o(1)$, MAXDAG(D,c) will output an induced acyclic subgraph of size at least $(1+\epsilon')\log_q np$, where $\epsilon' = c/\sqrt{\log_q np}$.*

Conclusions

We used an upper bound on the maximum number $ao(G)$ of acyclic orientations of an undirected graph $G = (V, E)$, to get an upper bound on $mas(D)$. This bound seems to be "nearly" the best possible, using the first moment method in the following sense. We also obtained an exact expression for the probability of an arbitrary b-set inducing a dag for any b. We also obtained a lower bound on this probability which, in turn, led to a lower bound on the expected number of induced induced dags of size b. This lower bound was shown to approach ∞ asymptotically for a value of b which is of the form $b = \frac{2(\ln np)}{\ln q} + X$ for some positive $X = \Theta((\ln q)^{-1})$.

Next, we analyse a polynomial time heuristic MAXDAG(D, c) for getting a large induced acyclic subgraph in a random digraph, and show that for fixed values of the arc-probability p, it gives an acyclic subgraph of size at least $\log_q w + c\sqrt{\log_q w}$ for any constant c, which is a slight improvement over the bound of the greedy heuristic **MaximalAcyclic**(D) given in [63]. Also, it was noted that the **MaximalAcyclic**(D) algorithm yields a subgraph of size at least $\epsilon(\log_q w)$ for some $\epsilon = \epsilon(n) \to 1$ as $n \to \infty$.

Finally, we obtained analogues of all the results obtained in this and the previous chapter, for random digraphs with 2-cycles allowed i.e. $\mathcal{D}_2(n, p)$. It is seen that the analysis does not differ significantly. A few differences were highlighted when they occurred.

Chapter 7

On Induced Paths, Holes and Trees in Random Graphs

7.1 Introduction

In this chapter, we study the sizes of largest induced subgraphs, such as paths, holes and trees in random graphs. We present and prove a 2-point concentration for largest induced paths and largest holes and also present a considerably improved concentration result on the size of the largest induced trees. The random model we use for random graphs is the $\mathcal{G}(n,p)$ model introduced in Chapter 3 with some assumption on $p = p(n)$. Throughout, we assume that $V = \{1, 2, \ldots, n\}$.

Notation : Given a natural number $n \in \mathbb{Z}^+$, we indicate the set $\{1, \ldots, n\}$ by $[n]$. Define $q := (1-p)^{-1}$. We ignore floors and ceilings wherever they are not crucial. We use $\mathcal{B}(n,\mu)$ to denote the sum of n identically and independently distributed indicator variables each having mean μ. For non-negative integers n and b, we use $(n)_b$ to denote the expression $\prod_{0\leq j\leq b-1} n-j$. We use $\ln n$ to denote the natural logarithm of n. For a set A and an integer $k \geq 0$, we use $\binom{A}{k}$ to denote the collection $\{B \subseteq A : |B| = k\}$.

7.1.1 Previous Work

The problem of finding large induced trees in the random graph $\mathcal{G}(n,p)$ was first studied by Erdős and Palka in [29]. Given a graph G, denote by $T(G)$ the size (= number of vertices) of any largest induced tree in G. Erdős and Palka showed that

Theorem 7.1.1 *For every $\epsilon > 0$, for every fixed $p : 0 < p < 1$, a.a.s. $G \in \mathcal{G}(n,p)$ has $T(G)$*

satisfying

$$(2-\epsilon)\log_q n \quad < T(G) \quad < (2+\epsilon)\log_q n$$

They also conjectured that for $p = c/n$ ($c > 1$ is any constant), $G \in \mathcal{G}(n,p)$ a.a.s. has an induced tree of size $\gamma(c)n$ where $\gamma(c)$ depends only on c. This was verified affirmatively and independently by de la Vega [69] and several others including Frieze and Jackson [35], Kučera and Rödl [46], and Łuczak and Palka [48] who showed that when $p = c/n$, $c \in \Re^+$, $T(G) \geq \gamma(c)n$, where $\gamma(c)$ depends only on the constant c. Later de la Vega [70] determined the constant $\gamma(c)$ to $\approx 2\ln c/c$.

Given a graph G, let $h(G)$ denote the size of a largest induced cycle (shortly *hole*) in G. Large holes in random graphs were first studied by Frieze and Jackson in [34], in which they showed that the random graph $\mathcal{G}(n,p)$, $p = c/n$, a.a.s. has a hole of size $\Omega(nc^{-3})$. They also proved that for any fixed $d \geq 3$, the random regular graph $G(n,d)$ a.a.s. has a hole of size $\Omega(nd^{-2})$. Later Suen [64] improved the lower bound for $h(G)$, for $G \in \mathcal{G}(n,c/n)$, for any fixed $c > 1$ and $\epsilon > 0$, to at least $(h(c)-\epsilon)n$ where $h(c)$ is defined below and approaches $(\ln c)/c$ for large enough c.

The question of studying the size of the largest induced *path* in $\mathcal{G}(n,p)$, was first studied by Frieze and Jackson in [34], in the course of their work on holes. Since a hole is just an induced path with an edge joining the endpoints, the existence of a large hole in $\mathcal{G}(n,p)$ is very likely if a large induced path is shown to exist a.a.s., and this was the idea used by Frieze and Jackson. On the other hand, large induced paths in $\mathcal{G}(n,p)$ are interesting in their own right, and Suen [64] studied this problem, showing that when $p = c/n$, for any fixed $c > 1$ and any $\epsilon > 0$, a.a.s. the random graph $\mathcal{G}(n,p)$ has an induced path of size at least $(1-\epsilon)h(c)n$, where

$$h(c) \quad = \quad c^{-1}\int_1^c \frac{(1-y(\zeta)}{\zeta}d\zeta$$

where $y(\zeta)$ is the smallest positive root of $y = e^{\zeta(y-1)}$. As $c \to \infty$, $h(c) \to (\ln c)/c$ and hence a.a.s., $mip(G) \geq (1-\epsilon)(n\ln c)/c$. Almost all of the above mentioned previous results are for sparse random graphs (that is, for $p = c/n$ for constant $c > 1$). But no previous work on tight concentration of these invariants is known to have been carried out. The dense case (corresponding to higher values of $p = p(n)$) has not been looked at in such great detail. In this chapter, we look at this case and obtain very tight concentration results.

7.1.2 Improved results on sizes of induced paths, trees and holes

Throughout the chapter, we assume that $G \in \mathcal{G}(n,p)$ for $p = p(n) \leq 1 - \epsilon$ where $\epsilon > 0$ is an arbitrary but fixed constant. We study induced subgraphs (paths, holes and trees) for dense random graphs. Our first result is a 2-point concentration for $mip(G)$, for $G \in \mathcal{G}(n,p)$ and $p \geq n^{-1/2}(\ln n)^2$:

Definition Let $b^* = b^*(n,p)$ be the maximum integer b such that $(n)_b p^{b-1}(1-p)^{\binom{b-1}{2}} \geq np/(\ln \ln n)$.

It can be verified (see Claim 7.2.1). using the given lower bound on p, that

$$\lfloor 2(\log_q np) + 2 \rfloor \leq b^* \leq \lceil 2\log_q np + 3 \rceil \quad \ldots\ldots\ldots \quad (A)$$

Theorem 7.1.2 *If $p \geq n^{-1/2}(\ln n)^2$, then $mip(G)$ lies in the set $\{b^*, b^*+1\}$ a.a.s.*

Since an induced path is also an induced tree, we get a significant additive improvement over Erdős and Palka's long-standing (30 year old) lower bound for $T(G)$ (which was for fixed $0 < p < 1$) as a corollary.

Corollary 7.1.3 *If $p \geq n^{-1/2}(\ln n)^2$, then there exists an induced tree of size b^* in G a.a.s.*

The above corollary, combined with a more careful analysis of Erdős and Palka's first moment bound gives

Corollary 7.1.4 *For $p \geq n^{-1/2}(\ln n)^2$ and $G \in \mathcal{G}(n,p)$,*

$$b^* \quad \leq T(G) \leq \quad 2\log_q np + O(1/\ln q)$$

a.a.s. As a result, $T(G) = 2(\log_q np) + O(1/\ln q)$ a.a.s.

Hence it is seen that the asymptotic upper bound on the range of concentration of $T(G)$ is improved, from the previously known bound of $O(\ln n/\ln q)$, to $O(1/\ln q) = O(1/p)$.

As for induced paths, we also obtain a similar 2-point concentration for the size of a longest hole:

Definition Let $h^* = h^*(n,p)$ be the maximum integer b such that $(n)_b p^b (1-p)^{\binom{b-1}{2}-1}/2b \geq np/(\ln \ln n)$.

Also, it can be verified (see Claim 7.3.1). using the given lower bound on p, that

$$\lfloor 2(\log_q np) + 2 \rfloor \leq h^* \leq \lceil 2\log_q np + 2 \rceil \quad \ldots\ldots\ldots \quad (B)$$

Theorem 7.1.5 *Let $G \in \mathcal{G}(n,p)$. Then, provided $p \geq n^{-1/2}(\ln n)^2$, a.a.s., $h(G) \in \{h^*, h^*+1\}$.*

The proofs of the above results involve just the well-known first and second moment methods.

7.2 Induced paths : Proof of Theorem 7.1.2

For $b \geq 1$, let $X(b) = X(n, b, p)$ be the number of induced paths on b vertices in $G \in \mathcal{G}(n, p)$. The following claim determines b^* upto constant additive factors.

Claim 7.2.1 *For $p \geq n^{-1/2}(\ln^2 n)$, $\lfloor 2\log_q np + 2 \rfloor \leq b^* \leq \lceil 2\log_q np + 3 \rceil$.*

Proof The upper bound follows from the proof of Claim 7.2.2 given below. Hence we establish only the lower bound. Suppose $b = \lfloor 2(\log_q np) + 2 \rfloor$. Let $0 \leq \delta < 1$ be such that $b = 2\log_q np + 2 - \delta$. Let X denote $X(b)$.

$$(b-2)/2 = \log_q np - (\delta/2)$$

Now, $(n)_b \geq (n-b)^b = n^b(1-b/n)^b$. From the assumed lower bound on p and the established upper bound on b^*, it follows that $b^* = O(n^{1/2}/(\ln n))$. Hence $(1-b/n)^b = e^{-o(1)} = 1 - o(1)$. Hence for all $p \geq n^{-1/2}(\ln^2 n)$, we get

$$\begin{aligned}
2\mathcal{E}[X] &\geq n[1-o(1)](np(1-p)^{(b-2)/2})^{b-1} \\
&\approx n(np(1-p)^{\log_q np}(1-p)^{-\delta/2})^{b-1} \\
&= n(1-p)^{-\delta\log_q np - \Theta(1)} \\
&= \left(\frac{n(1-p)^{-\Theta(1)}}{(np)^{\delta}}\right) \\
&= \Omega(n^{1/4}) \geq \ln n \quad \text{for large } n
\end{aligned}$$

This establishes that $b^* \geq \lfloor 2(\log_q np) + 2 \rfloor$. In fact, a more careful analysis shows that $b^* \geq \lceil 2(\log_q np) + 2 \rceil$ for $n^{-1/2}(\ln^2 n) \leq p \leq 1/5(\ln n)$. □

7.2.1 Proof of $mip(G) \leq b^* + 1$

Using Inequality (A) (stated before) and also the lower bound on p, we notice that $b^* = O(n^{1/2}/(\ln n))$.
The probability that a given ordered (orderings considered up to reversal) set A of b vertices induces a path is given by

$$\mathbf{Pr}[G[A] \text{ is an induced path }] = p^{b-1}(1-p)^{\binom{b-1}{2}}$$

Hence the expected number of b-length induced paths is

$$\mathcal{E}[X(b)] = \binom{n}{b}\frac{b!}{2}p^{b-1}(1-p)^{\binom{b-1}{2}} = \frac{(n)_b}{2}p^{b-1}(1-p)^{\binom{b-1}{2}}$$

Hence, for $b = b^* \pm O(1)$,

$$\frac{\mathcal{E}[X(b+1)]}{\mathcal{E}[X(b)]} \;=\; (n-b)p(1-p)^{b-1} \;\approx\; \frac{(1-p)^{\pm O(1)}}{np} \;=\; \Theta\left(\frac{1}{np}\right)$$

As a result,

$$\mathcal{E}[X(b^*+2)] \;=\; O(E[X(b^*+1)](np)^{-1}) \;=\; O((\ln\ln n)^{-1}) = o(1) \tag{7.1}$$

This establishes that $mip(G) \leq b^* + 1$ a.a.s. In fact, we can prove the following upper bound on $mip(G)$ which holds for any value of $p = p(n)$.

Claim 7.2.2 *For any $p = p(n)$, $mip(G) \leq \lceil 2\log_q np + 3 \rceil$ a.a.s.*

Proof Suppose $b = \lceil 2(\log_q np) + 4 \rceil$. Let X denote $X(b)$. Let $0 \leq \delta < 1$ be such that $b = 2\log_q np + 4 + \delta$.

$$(b-2)/2 = \log_q np + 1 + (\delta/2)$$

Now, $(n)_b = n(n-1)\ldots(n-b+1) \leq n^b e^{-b(b-1)/2n}$. Hence for all $p \geq 2/n$, we get

$$\begin{aligned}
2\mathcal{E}[X] &\leq ne^{-\binom{b}{2}/n}(np(1-p)^{(b-2)/2})^{b-1} \\
&= ne^{-b(b-1)/2n}(np(1-p)^{\log_q np}(1-p)^{1+\delta/2})^{b-1} \\
&= ne^{-\binom{b}{2}/n}(1-p)^{(2+\delta)\log_q np + \Theta(1)} \\
&= e^{-b(b-1)/2n}\left(\frac{n(1-p)^{\Theta(1)}}{(np)^{2+\delta}}\right) \\
&= A \cdot B
\end{aligned}$$

where $A \leq 1$ and $B \leq n$ always. Let $\omega = \omega(n)$ be a sufficiently slowly growing function. For $p \geq w/\sqrt{n}$, we have $B \to 0$. For p such that $p \leq \omega/\sqrt{n}$, we have $A = o(n^{-1})$. Hence, for $p \geq 2/n$, $E[X] \to 0$.
When $p < 2/n$, the largest component of $G(n,p)$ is a.a.s $O(\ln n)$ in size (see e.g. [14]), and hence much smaller than $2\log_q np$. □

7.2.2 Proof of $mip(G) \geq b^*$

Let $b = b^*$. Consider the variance and the expectation of the random variable $X = X(b)$ defined in the previous sub-subsection. Let X_i be the indicator variable for the i-th ordered b-set to induce a path, for a fixed enumeration of ordered b-sets. Therefore, $X = \sum_i X_i$.

Applying Chebyshev's Inequality and using standard simplifications (see e.g. [9], Chapter 4), it follows that

$$\mathbf{Pr}(X = 0) \;\leq\; \frac{Var(X)}{\mathcal{E}[X]^2} \leq \frac{1}{\mathcal{E}[X]} + M \tag{7.2}$$

where $M := \sum_i \sum_{j:2\leq|A_i\cap A_j|\leq b-1}(\mathcal{E}[X_iX_j] - \mathcal{E}[X_i]\mathcal{E}[X_j])/\mathcal{E}[X]^2$.
Since the random graph model $\mathcal{G}(n,p)$ is homogenous, the above expression for M simplifies to:

$$M \;=\; \sum_{j:2\leq|A_1\cap A_j|\leq b-1} \frac{\mathcal{E}[X_j|X_1 = 1] - \mathcal{E}[X_j]}{\mathcal{E}[X]} \tag{7.3}$$

By our choice of $b = b^*$, it follows that $\mathcal{E}[X] \to \infty$ and hence it suffices to prove that $M = o(1)$ in order to deduce that $mip(G) \geq b^*$ with probability $1 - o(1)$. That $M = o(1)$ follows from Claim 7.2.6 (established below) as follows : Using the previously observed fact $b^* = O(n^{1/2}/(\ln n))$, we infer that $M = O(b^4p/n^2) = o(1)$. This completes the proof of Theorem 7.1.2.

It remains to show that $M = o(1)$ and the following bound on M will be useful in that direction and it is established in Subsection 7.2.3.

For the remainder of this section, we use α denote any fixed and "sufficiently slowly" growing function $\alpha = \alpha(n)$. We will use this as a short notation to represent any $\omega(1)$ growth that arises in the proof arguments.

Lemma 7.2.3

$$M \;\leq\; \sum_{l=2}^{b-1} F_l$$

where

$$F_l \;=\; \frac{(n-b)_{b-l}}{(n)_b} \cdot p^{-l} \cdot (1-p)^{l-\binom{l}{2}} \cdot \sum_{k=1}^{\min\{l,b-l+1\}} f(k)\;;$$

$$f(k) \;=\; \left(\frac{(b-l+1)^{2k}}{(k!)^2}\right) \cdot \left(\frac{l^{k-1}}{(k-1)!}\right) \cdot 2^k \cdot k! \cdot \left[\left(\frac{p}{1-p}\right)^k - p^l(1-p)^{\binom{l}{2}-l}\right]$$

From the above expression,

$$M \;\leq\; \sum_{l=2}^{l\leq(b+1)/2} F_l + \sum_{l>(b+1)/2} F_l \;=\; M_1 + M_2.$$

Claim 7.2.4 *$f(k)$ is maximized at $k = k_{max} = \min\{l, b-l+1\}$. Further $\sum_k f(k)$ is $(1+o(1))f(k_{max})$.*

Proof of Claim 7.2.4 We prove the claim for all large values of n. We have the ratio

$$\begin{aligned}
\frac{f(k+1)}{f(k)} &= \frac{2(b-l+1)^2 l}{k(k+1)} \left(\frac{p^{k+1}(1-p)^{-k-1} - p^l (1-p)^{\binom{l}{2}-l}}{p^k (1-p)^{-k} - p^l (1-p)^{\binom{l}{2}-l}} \right) \\
&= \frac{2(b-l+1)^2 lp}{k(k+1)(1-p)} \left(\frac{1 - p^{l-k-1}(1-p)^{\binom{l}{2}-l+k+1}}{1 - p^{l-k}(1-p)^{\binom{l}{2}-l+k}} \right) \\
&= \frac{2(b-l+1)^2 lp}{k(k+1)(1-p)} \cdot S \text{ where} \\
S &= \left(\frac{1 - a\left(\frac{1-p}{p}\right)}{1-a} \right) \quad \text{where} \quad a = \left(\frac{p}{1-p} \right)^{l-k} \cdot (1-p)^{\binom{l}{2}}
\end{aligned}$$

We use Claim 7.2.5 stated and proved below. If $p \geq 1/2$, then the ratio is at least $(b+1)p/(1-p) \geq \alpha$. Suppose $p \leq 1/2$. When $l \leq L := \sqrt{\frac{2}{100(\ln q)}}$, $k_{max} = l$, hence, for all $k < k_{max}$, the ratio is at least $\frac{2(b-l+1)^2 l^3 p^2}{100k^2(1-p)} \geq (b+1)p \geq \alpha$ for all large n, for the assumed range of p. For l with $L < l \leq (b+1)/2$, the ratio is at least $(b+1)p/2 \geq \alpha$ again as $n \to \infty$. Therefore $f(k)$ achieves its maximum f_{max} at $k = l$ when $l \leq (b+1)/2$. When $l > (b+1)/2$. $k_{max} = b - l + 1$ and hence, for all $k < k_{max}$, the ratio is at least $2lp/(1-p) \geq (b+1)p \geq \alpha$. It follows that $\sum_k f(k)$ is upper bounded by the sum of a finite and increasing geometric series with a common ratio $\alpha = \omega(1)$. Hence $\sum_k f(k) = [1+o(1)]f(k_{max})$. □

Claim 7.2.5 *(i) $S \geq 1$ if $p \geq 1/2$.*
(ii) $S \geq \frac{l^2(\ln q)}{100}$ if $l \leq L := \sqrt{\frac{2}{100(\ln q)}}$; $S \geq 1 - e^{-1/200}$ if $l > L$.

Proof If $p \geq 1/2$, then $1 - p \leq p$, and hence $\left(\frac{1-p}{p}\right) \leq 1$. Therefore, $S = \left(\frac{1-a(1-p)p^{-1}}{1-a}\right) \geq 1$. Now assume $p < 1/2$. Write $l = \beta\sqrt{\frac{2}{\ln q}}$ and let $x = (1-p)^{\binom{l}{2}}$. Since $k < k_{max} \leq l$, we have $\frac{a(1-p)}{p} \leq x$. Hence, for each l, we have $S \geq 1 - x$. Now $\frac{\beta^2}{2(\ln q)} = \frac{l^2}{4} \leq \binom{l}{2} \leq \frac{l^2}{2} = \frac{\beta^2}{\ln q}$ and hence $x \leq e^{-\beta^2/2}$.

When $l \leq L$, we have $\beta \leq 1/10$ and hence $x = 1 - \beta^2/2 + O(\beta^4)$ and $S \geq \beta^2/2 - O(\beta^4) \geq \frac{l^2(\ln q)}{100}$.

For $l > L$, we have $\beta \geq 1/10$ and $x \leq e^{-1/200}$ and hence $S \geq 1 - e^{-1/200}$.

This completes the proof of Claim 7.2.5. □

Claim 7.2.6 $M \leq M_1 + M_2 = O(b^4 p/n^2)$.

Proof We consider two cases.
Case 1: $l \leq (b+1)/2$. By Claim 7.2.4,

$$\begin{aligned}\sum_{k=1}^{l} f(k) &= (1+o(1))f(l) \\ &\leq \left[2e^2\left(\frac{b-l+1}{l}\right)^2 l \cdot \frac{p}{1-p}\right]^l \cdot \left[1-(1-p)^{\binom{l}{2}}\right] \\ &= \left[2e^2\frac{(b-l+1)^2}{l} \cdot \frac{p}{1-p}\right]^l \cdot \left[1-(1-p)^{\binom{l}{2}}\right]\end{aligned}$$

Therefore

$$F_l \leq G_l := \frac{(1+o(1)) \cdot (n-b)_{b-l}}{(1-p)^{\binom{l}{2}} \cdot (n)_b} \cdot \left(\frac{2e^2(b-l+1)^2}{l}\right)^l \cdot \left[1-(1-p)^{\binom{l}{2}}\right]$$

By definition of G_l,

$$G_2 = (1+o(1)).\frac{(n-b)_{b-2}}{(1-p)(n)_b} \cdot \left(\frac{2e^2(b-1)^2}{2}\right)^2 \cdot [1-(1-p)] = O\left(\frac{b^4 p}{n^2}\right).$$

Therefore, the ratio G_l/G_2 is given by

$$\begin{aligned}\frac{G_l}{G_2} &= \frac{(n-b)_{b-l}}{(n-b)_{b-2}} \cdot (1-p)^{1-\binom{l}{2}} \cdot \frac{4(2e^2(b-l+1)^2)^l}{l^l\left(2e^2(b-1)^2\right)^2} \cdot \frac{[1-(1-p)^{\binom{l}{2}}]}{p} \\ &= \frac{O(1)}{n^{l-2}} \cdot (1-p)^{1-\binom{l}{2}} \cdot \frac{(2e^2(b-l+1)^2)^{l-2}}{l^l} \cdot \frac{[1-(1-p)^{\binom{l}{2}}]}{p} \\ &= O(1) \cdot \left(\frac{2e^2(b-l+1)^2(1-p)^{-(l+1)/2}}{nl}\right)^{l-2} \cdot \frac{[1-(1-p)^{\binom{l}{2}}]}{pl^2}\end{aligned}$$

Consider the base T_1 of the first term of the product, namely, $T_1 := \frac{2e^2(b-l+1)^2(1-p)^{-(l+1)/2}}{nl}$. When $l < \lfloor(\ln q)^{-1}\rfloor$, we have $(1-p)^{-(l+1)/2} < 1$. Therefore $T_1 \leq \frac{2e^2b^2}{n} = o(1)$ for each such l. For every l such that $\lfloor(\ln q)^{-1}\rfloor \leq l \leq (b+1)/2$, we have $(1-p)^{-(l+1)/2} \leq \sqrt{np}$ and hence $T_1 = O\left(\frac{(\ln np)^2}{\sqrt{np}}\right) = o(1)$, since $b = O(p^{-1}(\ln np))$.
Suppose $p \leq 1/2$ and hence $\frac{\ln q}{p} \leq 3/2$. Write $l = \beta\sqrt{\frac{2}{\ln q}}$ and $x = (1-p)^{\binom{l}{2}}$. For bounding the second term $T_2 := \frac{1-x}{pl^2}$ for the case $\beta \leq 1$, we apply, as explained before, $1-x \leq \beta^2 = \frac{l^2(\ln q)}{2}$ and hence $T_2 \leq 3/4$. For $\beta \geq 1$, we have $pl^2 \geq \frac{2p}{\ln q} \geq \frac{4}{3}$. Hence, $T_2 = \frac{1-x}{pl^2} \leq \frac{3}{4} < 1$. As a result, $T_1^{l_2} \cdot T_2 \leq (1/\alpha)^{l-2}$ for every $l \geq 3$.

Suppose $p \geq 1/2$. Then, $T_2 \leq 2/l^2 < 1$. Hence, $T_1^{l_2} \cdot T_2 \leq (1/\alpha)^{l-2}$ for every $l \geq 3$.

Therefore, the sum $\sum_{l=2}^{l \leq (b+1)/2} G_l$ is upper bounded by the sum of an infinite geometric progression, whose first term is G_2 and common ratio is $1/\alpha$. Hence we get,

$$M_1 \;=\; \sum_{l=2}^{l \leq (b+1)/2} F_l \;\leq\; \sum_{l=2}^{l \leq (b+1)/2} G_l \;\leq\; [1+o(1)] \cdot G_2 \;=\; O\left(\frac{b^4 p}{n^2}\right)$$

Case 2: $l > (b+1)/2$. Using Stirling's asymptotic estimate of factorials,

$$\sum_{k=1}^{b-l+1} f(k) \;=\; [1+o(1)] \cdot f(b-l+1) \;\leq\; \left(\frac{p}{1-p} \cdot (2e^2 l)\right)^{b-l+1}.$$

Therefore,

$$F_l \leq G_l := \frac{(n-b)_{b-l} \cdot (2e^2 l)^{b-l+1}}{(n)_b} \cdot \frac{p^{b-2l+1}}{(1-p)^{b-2l+\binom{l}{2}+1}}.$$

Using the definition of G_l and also that of $b = b^*$,

$$\begin{aligned} G_{b-1} &= \frac{n-b}{(n)_b} \cdot (2e^2(b-1))^2 \cdot \left(\frac{p}{1-p}\right)^{-b+3} \cdot \frac{1}{(1-p)^{\binom{b-1}{2}}} \\ &\leq \frac{np^2(2e^2(b-1))^2(1-p)^{b-3}}{(n)_b p^{b-1}(1-p)^{\binom{b-1}{2}}} \;=\; \frac{np^2(1-p)^{b-3}(2e^2(b-1))^2}{\mathcal{E}[X(b)]} \\ &= O\left(\frac{np^2(\ln\ln n)b^2}{(np)^3}\right) \;=\; O\left(\frac{b^2(\ln\ln n)}{n^2 p}\right) \end{aligned}$$

Now, the ratio G_l/G_{b-1} is given by

$$\begin{aligned} \frac{G_l}{G_{b-1}} &= (n-b-1)_{b-l-1} \cdot (2e^2 l)^{b-l-1} \cdot \left(\frac{p}{1-p}\right)^{2b-2l-2} \cdot (1-p)^{\binom{b-1}{2}-\binom{l}{2}} \cdot \left(\frac{l}{b-1}\right)^2 \\ &\leq \left(\frac{np^2(2e^2 l)}{(1-p)^2}\right)^{b-l-1} \cdot (1-p)^{\binom{b-1}{2}-\binom{l}{2}} \;=\; \left(\frac{np^2(2e^2 l)(1-p)^{(b+l)/2}}{(1-p)^3}\right)^{b-l-1} \\ &\leq \left(\frac{np^2 l \cdot O(1)}{(np)^{3/2}}\right)^{b-l-1} \;\leq\; \left(\frac{O(1) \cdot l\sqrt{p}}{\sqrt{n}}\right)^{b-l-1} \quad \text{since } (b+l)/2 \geq 3b/4 \text{ for } l > (b+1)/2 \\ &= \left(O\left(\frac{\ln np}{\sqrt{np}}\right)\right)^{b-l-1} \quad \text{for } l < b-1 \end{aligned}$$

Therefore

$$M_2 \;=\; \sum_{l > (b+1)/2} F_l \leq \sum_{l > (b+1)/2} G_l \;=\; (1+o(1))\dot{G}_{b-1} \;=\; O\left(\frac{b^2 \ln\ln n}{n^2 p}\right) \;=\; o\left(\frac{b^4 p}{n^2}\right).$$

Hence,

$$M \;=\; M_1 + M_2 \;=\; O\left(\frac{b^4 p}{n^2}\right).$$

□

7.2.3 Proof of Lemma 7.2.3

Proof As shown in Equation 7.3, fix the first ordered set A_i to be A_1. Let the number of vertices in the intersection be l.

Claim 7.2.7 *If $G[A_j]$ is an induced path, then the induced subgraph $G[A_1 \cap A_j]$ must be a union of path segments. The order of occurrence, as well as the alignment (i.e. which end of a segment comes first) of these segments in $G[A_1]$ and $G[A_j]$ however, can differ.*

Proof Notice that $G[A_1 \cap A_j]$ is an induced subgraph of $G[A_1]$ as well as $G[A_j]$. Since both $G[A_1]$ and $G[A_j]$ are paths of length b, the only possible induced subgraphs are disjoint unions of path segments. It is easy to see, however, that even for a fixed A_1, and $A_1 \cap A_j$, there is no constraint on the relative ordering and alignment of these segments in $G[A_j]$. □

We next make an observation on the conditional probability $\mathbf{Pr}[X_j = 1 | X_1 = 1]$ when there are l vertices in the intersection $A_1 \cap A_j$, divided into k path segments:

Claim 7.2.8 *For any ordered set A_j which has k intersection segments with A_1 and has intersection $|A_1 \cap A_j| = l$, the conditional probability is given by:*

$$\mathbf{Pr}[X_j = 1 | X_1 = 1] = p^{b-1-(l-k)}(1-p)^{\binom{b-1}{2}-\binom{l}{2}+l-k}$$

Proof The proof follows by counting the number of edges and non-edges lying in the intersection $A_1 \cap A_j$, when $|A_1 \cap A_j| = l$ and $A_1 \cap A_j$ induces k contiguous segments on A_1. □

Given the ordered b-set A_1, let $S(l,k)$ be the number of ways of choosing an ordered b-set A_j, such that $|A_1 \cap A_j| = l$, and $A_1 \cap A_j$ induces a forest on k vertex disjoint paths in G. Then from Claim 7.2.8 and Equation 7.1, it follows that

$$M \;\leq\; \sum_{l=2}^{b-1} \sum_{k=1}^{l} \frac{S(l,k)}{(n)_b} . [p^{-(l-k)}(1-p)^{-\binom{l}{2}+l-k} - 1] \tag{7.4}$$

We next define three sets of tuples which will help determine the value of $S(l,k)$:

Definition Let the set $\mathcal{T}_1$ denote all tuples $(a_0, b_1, a_1, b_2, \ldots, b_k, a_k)$, such that

(i) $\sum_{i=1}^{k} b_i = l$, $\forall i \;:\; b_i \geq 1,\; b_i \in \mathbb{Z}^+$.

(ii) $\sum_{i=0}^{k} a_i = b - l$, $a_i \in \mathbb{Z}^+$, $a_0, a_k \geq 0$, $\forall i \neq 0, k \ : \ a_i \geq 1$.

Definition Given a tuple $\tau \in \mathcal{T}_1$, let $\mathcal{T}_2(\tau)$ denote the set of all tuples $(\pi, c_1, \ldots, c_k)$, where π is an ordering of $\{1, \ldots, k\}$, and for $i = 1, \ldots, k$,

$$c_i \in \begin{cases} \{0\} & \text{if } b_i = 1 \\ \{0, 1\} & \text{otherwise} \end{cases}$$

Definition Let $\mathcal{T}_{12}$ denote the set of all ordered pairs (τ_1, τ_2) where $\tau_1 \in \mathcal{T}_1$ and $\tau_2 \in \mathcal{T}_2(\tau_1)$.

Definition Let $\mathcal{T}_3$ denote the set of tuples $(d_0, \ldots, d_k)$ such that $\sum_{i=0}^{k} d_i = b - l$, $d_i \in \mathbb{Z}^+$, $d_0, d_k \geq 0$ and $d_i \geq 1$ for all other i.

Finally, let $\mathcal{T} = \mathcal{T}_{12} \times \mathcal{T}_3$. To get an upper bound on $S(l, k)$, just observe the following:

Proposition 7.2.9 *Given A_1 and an ordered set B of $b - l$ vertices, each element of $\mathcal{T}$ specifies (in a bijective fashion) a unique ordered set A_j having b vertices, such that $A_j \setminus A_1 = B$, $|A_1 \cap A_j| = l$ and $A_1 \cap A_j$ induces a collection of k disjoint and separated sub-paths of A_1.*

Proof Given an element of $\mathcal{T} = \mathcal{T}_{12} \times \mathcal{T}_3$, say $u = ((a_0, b_1, \ldots, a_k), (\pi, c_1, \ldots, c_k), (d_0, \ldots, d_k))$, first, divide the ordered set A_1 into an ordered partition into paths having $a_0, b_1, a_1, \ldots, a_k$ vertices respectively. The paths corresponding to integers b_i are chosen to lie in the intersection $A_1 \cap A_j$. Let these intersection segments be called $P_1, \ldots, P_k$. By the definition of b_i's, the total number of vertices in the intersection is thus l. Order the P_i's using the ordering π. For each P_i, call the end-point occuring earlier in A_1 as the "head" if $c_i = 0$, otherwise the head is the end-point occuring later in A_1. Next, divide B into ordered parts $D_0, \ldots, D_k$, having sizes $d_0, \ldots, d_k$. Now, insert P_1 between D_0 and D_1, P_2 between D_1 and D_2, and so on, while ensuring that the head of each P_i succeeds the last vertex of D_{i-1}. By the definitions of P_is and D_is, we get a unique ordered b-set A_j such that $A_1 \cap A_j$ has l vertices, divided into k separated path segments. □

The value of $S(l, k)$ can now be ascertained from the following claim:

Claim 7.2.10 $S(l, k) = (n - b)_{b-l} . |\mathcal{T}|$

Proof Choose an ordered set B of size $b - l$, in $(n - b)_{b-l}$ ways, and apply Proposition 7.2.9. □

Hence, to upper-bound $S(l,k)$, we need to upper-bound the sizes of the sets $\mathcal{T}_i$, $i = 1, 2, 3$. Clearly, $|\mathcal{T}_2(\tau)|$ is at most $k!2^k$ for each $\tau \in \mathcal{T}_1$. To estimate the sizes of $\mathcal{T}_i$, $i = 1, 3$, we recall a basic combinatorial fact:

Proposition 7.2.11 (see e.g. [68], Chapter 13) *The number of integral solutions of $\sum_i x_i = a$, with integral constraints $x_i \geq c_i$; $a, c_i \in \mathbb{Z}$; $1 \leq i \leq r$, is $\binom{a-(\sum_i c_i)+r-1}{r-1}$.*

Hence, the set $\mathcal{T}_3$ consisting of all integral solutions of $\sum_i d_i = b - l$, such that $d_0, d_k \geq 0$ and $d_i \geq 1$, for $i = 1, \ldots, k$, has cardinality $\binom{b-l+1}{k}$. The size of $\mathcal{T}_1$ can be determined by counting all solutions, in non-negative integers of the following pairs of equations

$$\sum_{(a_i):a_0,a_k\geq 0,a_1,\ldots a_{k-1}\geq 1} a_i = b - l \tag{7.5}$$

$$\sum_{(b_j):b_j\geq 1} b_j = l \tag{7.6}$$

The number of solutions satisfying both of the above equations, by Proposition, 7.2.11 comes to $\binom{b-l+1}{k}\binom{l-1}{k-1}$. Therefore, $|\mathcal{T}_1| = \binom{b-l+1}{k}.\binom{l-1}{k-1}$. ¿From the above argument and Claim 7.2.10 we get

$$\begin{aligned} S(l,k) &\leq (n-b)_{b-l}\binom{b-l+1}{k}^2\binom{l-1}{k-1}2^k k! \\ &\leq (n-b)_{b-l}\frac{(b-l+1)^{2k}}{(k!)^2}\binom{l-1}{k-1}2^k k! \end{aligned}$$

Plugging the above bound on $S(l,k)$ in Equation (7.4) proves the Lemma. □

7.3 Holes – Proof of Theorem 7.1.5

Redefine, for this section, $X := X(n,b,p)$ to be the number of holes of size b in $G \in G(n,p)$. The following claim determines h^* upto constant additive factors.

Claim 7.3.1 *(i) For $p \geq n^{-1/2}(\ln n)^2$, $\lfloor 2\log_q np + 2 \rfloor \leq h^* \leq \lceil 2\log_q np + 2 \rceil$.*
(ii) For any $p = p(n)$, $h(G) \leq \lceil 2(\log_q np) + 2 \rceil$ a.a.s.

Proof Suppose $p \geq n^{-1/2}(\ln n)^2$. Write $b = 2(\log_q np) + 2 + \delta$ where δ is defined by the value we assign to b. Let X denote $X(b)$. We have, after employing simplifications similar

to the ones employed before,

$$\begin{aligned} E[X] &= \frac{(n)_b}{2b} \cdot p^b \cdot (1-p)^{\binom{b-1}{2}-1} \\ &\approx \frac{(np)^{1-\delta} \cdot (1-p)^{\Theta(1)}}{2b(1-p)} \\ &\to 0 \quad \text{for} \quad b = \lceil 2(\log_q np) + 3 \rceil \\ &= \Omega((\ln n)^2) \quad \text{for} \quad b = \lfloor 2(\log_q np) + 2 \rfloor \end{aligned}$$

This establishes Part (i) of the claim. For Part (ii) of the claim, the '$\approx$' in the second equation will be replaced by '$\leq$', thereby establishing the claim. □

Proof of Theorem 7.1.5 The proof of this theorem is along similar lines as the proof of Theorem 7.1.2. Consider the ratio $r(b)$ of the expected number of holes of size $b+1$ to the expected number of holes of size b, where $b = h^* \pm O(1)$.

$$\begin{aligned} r(b) &= \frac{\mathcal{E}[X(b+1)]}{\mathcal{E}[X(b)]} = (n-b)p(1-p)^{b-1}\frac{b}{b+1} \\ &= \Theta\left(np\frac{1}{(np)^2}\right) = \Theta\left(\frac{1}{np}\right) \end{aligned}$$

Hence, it follows from the definition of h^* that

$$\mathcal{E}[X(h^*+2)] = O(E[X(h^*+1)](np)^{-1}) = O\left(\frac{\ln n}{np}\right) = o(1) \tag{7.7}$$

This establishes that $h(G) \leq h^* + 1$ a.a.s.

For the lower bound, set $b = h^*$ and using a fixed but arbitrary enumeration of all cyclically ordered subsets of size b, write $X = X(b) = \sum_{1 \leq i \leq m} X_i$, where $m = (n)_b/(2b)$, note that $E[X] \to \infty$. Here, X_i denotes the indicator random variable for the i-th cyclically ordered set inducing a hole. We again use Chebyshev's inequality and arguments similar to those used for induced paths. As before, it suffices to prove that $M_h = o(1)$ where M_h is defined by equation (7.3). An upper bound for M_h is given by Lemma 7.3.2. Comparing this bound with the upper bound on M (for induced paths) stated in 7.2.3 and applying Claim 7.2.6, we deduce that $M_h \leq O\left(b^4 p/n^2\right) = o(1)$ for $p \geq n^{-1/2}(\ln n)^2$. Thus, we have shown that

$$\mathbf{Pr}[h(G) < h^*] = \mathbf{Pr}[X = 0] \leq (E[X])^{-1} + M_h = o(1) \quad \text{and}$$

Hence, it follows that $h(G) \in \{h^*, h^*+1\}$ a.a.s. □

It remains to prove the following lemma.

Lemma 7.3.2

$$\begin{aligned} M_h &\leq \sum_{l=2}^{b-1} \sum_{k=1}^{\min\{l,b-l\}} \frac{(n-b)_{b-l}}{(n)_b} \cdot \frac{b^2}{k} \cdot \binom{b-l-1}{k-1}^2 \binom{l-1}{k-1} 2^k k! \left(p^{-(l-k)} (1-p)^{-\binom{l}{2}+l-k} - 1 \right) \\ &= O\left(b^4 p/n^2\right) \end{aligned}$$

Proof An induced cycle can be considered to be an induced path with the endpoints joined. The argument for holes therefore, follows exactly the same lines as those of paths, with the following differences:

(i) We define $S(l,k)$ in an analogous way for holes. To upper bound $S(l,k)$, we note that in choosing the intersection $A_i \cap A_j$ and the difference set $A_j \backslash A_i$, the parts a_0 and a_k are now considered as one part - say a_0 - since the last and first vertex of A_i will be joined. Since a_0 must now differentiate between b_1 and b_k, there must be at least one vertex in the segment corresponding to a_0. This changes the number of solutions of equation (7.5) to $\binom{b-l-1}{k-1}$, for both choices of $A_i \backslash A_j$ and $A_j \backslash A_i$.

(ii) For $A_i \backslash A_j$, we also need to introduce a multiplicative factor of b to account for the choice of the starting vertex of the "first" non-intersecting segment of size a_0. Another multiplicative factor of b is introduced to account for the fact that number of potential holes is $(n)_b/(2b)$. For $A_j \backslash A_i$, we have already accounted for the starting vertex in the term $(n-b)_{b-l}$. We also need to divide by a factor of k since the number of ways in the k paths of $A_i \cap A_j$ can be cyclically ordered is $k!/k$.

To prove the bound in the lemma, we shall use the results of Section 7.2. Notice that expressions for M, M_h are both very similar. We'll define $F_h(l)$ and $f_h(k)$ similar to F_l and $f(k)$ respectively, from Section 7.2:

$$\begin{aligned} & \binom{b-l-1}{k-1}^2 \binom{l-1}{k-1} 2^k k! [p^k (1-p)^{-k} - p^l (1-p)^{\binom{l}{2}-l}] \\ = & \binom{b-l}{k}^2 \left(\frac{k}{b-l}\right)^2 \binom{l-1}{k-1} 2^k k! [p^k (1-p)^{-k} - p^l (1-p)^{\binom{l}{2}-l}] \\ \leq & \binom{b-l+1}{k}^2 \left(\frac{k}{b-l}\right)^2 \binom{l-1}{k-1} 2^k k! [p^k (1-p)^{-k} - p^l (1-p)^{\binom{l}{2}-l}] \end{aligned}$$

However, note that for $k > b-l$, $\binom{b-l}{k} = 0$. Hence, define $f_h(k)$ as follows:

$$f_h(k) = \begin{cases} 0 \text{ if either } k > b-l \text{ or } k > l \\ \binom{b-l+1}{k}^2 \cdot \left(\frac{k}{b-l}\right)^2 \cdot \binom{l-1}{k-1} \cdot 2^k k! \cdot [p^k (1-p)^{-k} - p^l (1-p)^{\binom{l}{2}-l}], \text{ otherwise} \end{cases}$$

Observe that $f_h(k) \leq f(k).\left(\frac{k}{b-l}\right)^2$. Define

$$F_h(l) \;:=\; \frac{(n-b)_{b-l}}{(n)_b} b^2 p^{-l} (1-p)^{l-\binom{l}{2}} \sum_{k}^{\min\{l,b-l\}} f_h(k)$$

Hence,

$$M_h \;\leq\; \sum_{l=2}^{b-1} F_h(l).$$

Notice that the ratio of successive terms

$$\frac{f_h(k+1)}{f_h(k)} \;=\; \frac{f(k+1)}{f(k)}.\left(\frac{k+1}{k}\right)^2$$

is greater than the ratio of $f(k)$'s, and hence we can bound the sum $\sum_k f_h(k)$ by $(1 + o(1)).f_h(k_{max})$, where $k_{max} = \min\{l, b-l\}$. Therefore, ignoring constant multiplicative factors, one can upper bound the terms $F_h(l)$ by $G_h(l)$, where $G_h(l)$ is defined as follows:

$$G_h(l) \;=\; \begin{cases} \frac{(n-b)_{b-l}}{(n)_b (1-p)^{\binom{l}{2}}} \cdot b^2 \cdot \left(\frac{l}{b-l}\right)^2 \cdot \left(\frac{2e^2(b-l+1)^2}{l}\right)^l \cdot [1-(1-p)^{\binom{l}{2}}], & l \leq b/2 \\ \frac{(n-b)_{b-l}}{(n)_b (1-p)^{\binom{l}{2}}} \cdot b^2 \cdot (2e^2 l)^{b-l} \cdot \left(\frac{p}{1-p}\right)^{b-2l}, & l > b/2 \end{cases}$$

Again, it is easy to see that the ratios $\frac{G_h(l)}{G_h(2)}$ (for $l \leq b/2$) and $G_h(l)/G_h(b-1)$ (for $l > b/2$) are the same as $\frac{G_l}{G_2}$ (or $\frac{G_l}{G_{b-1}}$) in the proof of Claim 7.2.6 except that for the former case, we need to multiply by a factor of $\left(\frac{l(b-2)}{2(b-l)}\right)^2$ which is at most l^2. Since $\frac{G_l}{G_2} \leq (1/\alpha)^{l-2}$, the l^2 multiplicative factor can also be easily absorbed.

Inspecting the terms for $l = 2, b-1$, we get that $G_h(2) = O(b^4 p/n^2)$, and

$$\begin{aligned} G_h(b-1) \;&=\; \frac{n-b}{(n)_b (1-p)^{\binom{b-1}{2}}} (2e^2 b^2 (b-1)).\left(\frac{1-p}{p}\right)^{b-2} \\ &\leq\; \frac{2e^2 b^3 n p^2}{(n)_b p^b (1-p)^{\binom{b-1}{2}-1}} (1-p)^{b-3} \;=\; O\left(\frac{b^2 p^2 n (1-p)^{b-3}}{\mathcal{E}[X(h^*)]}\right) \\ &=\; O\left(\frac{b^2 p^{-1} (\ln \ln np)}{n^2}\right) \;=\; O\left(\frac{b^4 p \cdot (\ln \ln n)}{n^2 (\ln np)^2}\right) \;=\; o\left(\frac{b^4 p}{n^2}\right). \end{aligned}$$

Therefore, the dominating term is $G_h(2)$, and hence the sum $\sum_{l=2}^{b-1} F_h(l)$ can be bound by $\Theta(1) \cdot G_h(2) = O\left(\frac{b^4 p}{n^2}\right)$, which is $o(1)$ for $p \geq n^{-1/2+o(1)}$.

□

7.4 Conclusion

We investigated certain non-monotone functions on graphs, such as the size of the largest induced path ($mip(G)$), induced cycle ($h(G)$) and induced tree ($T(G)$), in the random graph model $\mathcal{G}(n,p)$ and obtained a 2-point concentration of $mip(G)$ and $h(G)$, while improving the lower bound for $T(G)$. The 2-point concentration results for $mip(G)$ and $h(G)$ lead to the following question:

Question 1 : Given a fixed $p : 0 < p < 1$, do there exist 2 consecutive values $b(n), b(n)+1$, such that for $G \in \mathcal{G}(n,p)$, a.a.s. $T(G)$ is either $b(n)$ or $b(n)+1$?

Our results work for moderately-dense to dense random graphs:e.g. $mip(G)$ is 2-point concentrated for $p \geq n^{-1/2}(\ln n)^2$. What happens when the edge probability is smaller? Here an interesting observation is that several well-known techniques, such as sharp concentration inequalities (e.g. Azuma's, Talagrand's, etc.) employed for commonly-encountered monotone functions such as the independence number $\alpha(G)$ or the chromatic number $\chi(G)$, do not work for $mip(G)$, because the Lipschitz coefficient of $mip(G)$ cannot be bounded by any constant. Indeed, if G is a forest where every component (except one) is a path on $k+1$ vertices and the remaining component P is a path on $2k+2$ vertices, then $mip(G) = 2k+1$. But if the middle edge of P is removed to get G', then $mip(G') = k$. Even the application of polynomial concentration inequalities (see [71]) seems non-trivial due to the large degree of the polynomials involved. As such, it may be interesting to consider the following question:

Question 2 : Is there a function $f(n) = o(p^{-1}(\ln np))$, such that the random variable $mip(G)$, $G \in \mathcal{G}(n,p)$ with $p = p(n)$, $1/n \leq p \leq n^{-1/2}$, a.a.s. takes integer values from an interval of length at most $f(n)$?

Solving the last problem could give an easy proof of the following question investigated and asked by Suen [64], as well as de la Vega [70], which has remained open for more than two decades:

Question 3 : Does $\frac{mip(G)}{p^{-1}\ln np} \to 2$ when $n \to \infty$, for $p = \Theta(1/n)$?

Chapter 8

Independence Number of Locally Sparse Graphs and Hypergraphs

8.1 Introduction

For $k \geq 2$, a k-uniform hypergraph H is a pair $(V(H), E(H))$ where $E \subseteq \binom{V(H)}{k}$. A set $I \subset V(H)$ is an independent set of H if $e \not\subseteq I$ for every $e \in E(H)$, or equivalently, $\binom{I}{k} \cap E(H) = \emptyset$. The independence number of H, denoted by $\alpha(H)$, is the maximum size of an independent set in H. For $u \in V(H)$, its degree in H, denoted by $d_H(u)$, is defined to be $|\{e \in E(H) : u \in e\}|$ (we omit the subscript if it is obvious from the context). **Throughout** this chapter, we use t to denote $k-1$ except in some places where it stands for some real value (the correct meaning can be easily inferred from the context). Also, we use the term graph whenever k happens to be 2. A k-uniform hypergraph is *linear* if it has no 2-cycles where a 2-cycle is a set of 2 hyperedges containing at most $2t$ vertices. The dual of the above definition says that a linear hypergraph is one in which every pair of vertices is contained in at most one hyperedge.

A graph is said to be *K_r-free* if it does not contain any set of r vertices which form a clique. In [67], Turán proved a theorem giving a tight bound on the maximum number of edges that a K_r-free graph can have, which has since become the cornerstone theorem of extremal graph theory. Turán's theorem, when applied to the complement $\overline{G}$ of a graph G (i.e., the graph obtained by retaining the vertex set of G, and replacing the edges of G by non-edges and vice-versa), yields a lower bound $\alpha(G) \geq \frac{n}{d+1}$ where d denotes the average degree in G of its vertices.

Caro [20] and Wei [72] independently proved that $\alpha(G) \geq \sum_v \frac{1}{d(v)+1}$ which is at least $\frac{n}{d+1}$. The probabilistic proof of their result later appeared in Alon and Spencer's book [9]. [1] One natural extension of Turán's theorem to k-uniform hypergraphs H is the bound $\alpha(H) > c_k \frac{n}{d^{1/t}}$, and this was shown via an easy probabilistic argument by Spencer [60]. Caro and Tuza [21] improved this bound for arbitrary k-uniform hypergraphs. In order to state their lower bound, we need the following definition (of fractional binomial coefficients) from [37].

Definition For $t > 0$, $a \geq 0$, $d \in \mathbb{N}$

$$\binom{d+1/t}{a} := \frac{(td+1)(t(d-1)+1)...(t(d-a+1)+1)}{a!t^a}$$

What Caro and Tuza [21] showed was that

$$\alpha(H) \geq \sum_{v \in V(H)} \frac{1}{\binom{d(v)+1/t}{d(v)}}. \tag{8.1}$$

Indeed, an easy consequence of (8.1) is the following result.

Theorem 8.1.1 (Caro-Tuza [21]) *For every $k \geq 3$, there exists $d_k > 0$ such that every k-uniform hypergraph H has*

$$\alpha(H) \geq d_k \sum_{v \in V(H)} \frac{1}{(d(v)+1)^{1/t}}.$$

As a corollary, one infers the bound of Spencer above. Later, Thiele [66] provided a lower bound on the independence number of non-uniform hypergraphs, based on the degree rank (a generalization of degree sequence).

In this chapter, we prove new lower bounds for the independence number of locally sparse graphs and linear k-uniform hypergraphs. The starting point of our approach is the probabilistic proof of Boppana-Caro-Wei. This approach, together with some additional simple ideas, quickly yields a new short proof of the asymptotic version of Theorem 8.1.1 (see Section 8.2 for the detailed proof). Later, in section 8.4 we shall also give a probabilistic proof of the exact Caro-Tuza expression.

[1] According to R. Bopanna [19], the probabilistic argument in [9] was obtained by him, although it is possible that it was known earlier.

8.1.1 K_r-free graphs

For certain classes of sparse graphs, improvements of the Caro-Wei bound (in terms of average degree d) are known. Ajtai, Komlós and Szemerédi [5] proved a lower bound of $\Omega\left(\frac{n \log d}{d}\right)$ for the independence number of triangle-free graphs. An elegant and simpler proof was later given by Shearer [56], who also improved the constant involved. Ajtai, Erdős, Komlós and Szemerédi [3] showed that for K_r-free graphs ($r > 3$), the independence number is lower-bounded by $c_r(n/d)\log(\frac{(\log d)}{r})$, where $c_r \in \Re^+$ depends only on r. They also conjectured that the optimal bound is $c_r \frac{n \log d}{d}$. Shearer [58] improved their bound to $\Omega\left(\frac{n \log d}{d \log \log d}\right)$.

Caro and Tuza [21] raised the following question in their 1991 chapter :

(i) Can the lower bounds of Ajtai *et al* [5] and Shearer ([56], [58]) be generalized in terms of degree sequences?

We answer this question via the following two theorems.

Theorem 8.1.2 *For every $\epsilon \in (0,1)$ there exists $c > 0$ such that the following holds: Every triangle-free graph G with average degree D has independence number at least*

$$c(\log D) \sum_{v \in V(G)} \frac{1}{\max\{D^\epsilon, d(v)\}}.$$

Theorem 8.1.3 *For every $\epsilon \in (0,1)$ and $r \geq 4$, there exists $c > 0$ such that the following holds: Every K_r-free graph G with average degree D has independence number at least*

$$c\frac{\log D}{\log \log D} \sum_{v \in V(G)} \frac{1}{\max\{D^\epsilon, d(v)\}}.$$

8.1.2 Linear Hypergraphs

As mentioned earlier, a lower bound of $\Omega\left(n/d^{1/t}\right)$ for an n vertex k-uniform hypergraph with average degree d can be inferred from Theorem 8.1.1. Caro and Tuza [21] also raised the following question:

(ii) How can one extend the lower bounds of Ajtai *et al* [5] and Shearer ([56], [58]) to hypergraphs?

As it turns out, such extensions were known for the class of linear k-uniform hypergraphs. Indeed, the lower bound

$$\alpha(H) = \Omega\left(n\left(\frac{\log d}{d}\right)^{1/t}\right), \tag{8.2}$$

where H is a linear k-uniform hypergraph with average degree d was proved by Duke-Lefmann-Rödl [22], using the results of [4]. Our final result generalizes (8.2) in terms of the degree sequence of the hypergraph.

Theorem 8.1.4 *For every $k \geq 3$ and $\epsilon \in (0,1)$, there exists $c > 0$ such that the following holds: Every linear k-uniform hypergraph H with average degree D has independence number at least*

$$c(\log D)^{1/t} \sum_{v \in V(H)} \frac{1}{\max\{D^{\epsilon/t}, (d(v))^{1/t}\}}.$$

We also describe an infinite family of k-uniform linear hypergraphs to illustrate that the ratio between the bounds of Theorem 8.1.4 and (8.2) can be unbounded in terms of the number of vertices.

The remainder of this chapter is organized as follows. In Section 8.2, we give a new short proof of Theorem 8.1.1. In Section 8.3, we apply the analysis in Section 8.2 to the special case of linear hypergraphs, and obtain a "warm-up" result - Theorem 8.3.1, which will be helpful in proving the main technical result, Theorem 8.5.1, proved in Section 8.5. The expression obtained in Theorem 8.5.1 plays a crucial role in the proofs of Theorems 8.1.2, 8.1.3 and 8.1.4; these are provided in Section 8.6. In Section 8.7, we give infinite families of K_r-free graphs and k-uniform linear hypergraphs which illustrate that the bounds in Theorems 8.1.2, 8.1.3 and 8.1.4 can be bigger than the corresponding bounds in [5, 4, 22, 56, 58] by arbitrarily large multiplicative factors. Finally, in section 8.9, we state several combinatorial identities which follow as simple corollaries of Theorem 8.5.1.

8.2 A new proof of Theorem 8.1.1

In this section we obtain a new short proof of the asymptotic version of Theorem 8.1.1. First we obtain the following theorem which is later used to prove Theorem 8.1.1.

Theorem 8.2.1 *For every $k \geq 2$, there exists a constant $c = c_k$ such that any k-uniform hypergraph H on n vertices and $m \geq 1$ hyperedges satisfies*

$$\sum_{J \subset V(H)} \frac{1}{\binom{n}{|J|}} > c\frac{n}{m^{1/k}} \quad \ldots\ldots \quad (A)$$

where we sum over all independent sets J.

Proof Let $t_k(n,m)$ denote the LHS of (A). Consider any edge $e \in E(H)$. The edge e can belong to at most $\binom{n-k}{j-k}$ non-independent sets of size j. Since there are m edges there are at most $m\binom{n-k}{j-k}$ sets of size j that are not independent. Thus, at least $\binom{n}{j} - m\binom{n-k}{j-k}$ sets of size j are independent. Hence we have

$$\begin{aligned} t_k(n,m) &\geq \sum_{j=1}^{n} \left(1 - m\frac{\binom{n-k}{j-k}}{\binom{n}{j}}\right) = \sum_{j=1}^{n} \left(1 - m\frac{(j)_k}{(n)_k}\right) \\ &> \sum_{j=1}^{\lfloor n/(2m)^{1/k} \rfloor} \left(1 - m\frac{j^k}{n^k}\right) \geq \sum_{j=1}^{\lfloor n/(2m)^{1/k} \rfloor} \left(1 - m\frac{1}{2m}\right) \\ &\geq \frac{1}{2}\left\lfloor \frac{n}{(2m)^{1/k}} \right\rfloor \geq c_k \frac{n}{m^{1/k}} \end{aligned}$$

for some suitably chosen c_k which is close to $2^{-(k+1)/k}$. □

Let $H = (V,E)$ be a k-uniform hypergraph. For $k \geq 3$ and for $u \in V$ with $d_H(u) \geq 1$, the link graph associated with u in H is the t-uniform hypergraph $L_u = (U,F)$ where $U := \{v \neq u : \exists e \in E : \{u,v\} \subseteq e\}$ and $F = \{e \setminus u : u \in e \in E\}$. Let $\mathcal{I}(H)$ denote the collection of independent sets of H.

Proof of Theorem 8.1.1. As mentioned in the Introduction, the proof is an extension of the technique used in Alon and Spencer's book [9]. Let $H = (V,E)$ be an arbitrary k-uniform hypergraph. Choose uniformly at random a total ordering $\prec$ on V. Define an edge $e \in E$ to be *backward* for a vertex $v \in e$ if $u \prec v$ for every $u \in e \setminus \{v\}$. Define a random subset I to be the set of those vertices v such that no edge e incident at v is backward for v with respect to $\prec$. Clearly, I is independent in H. We have $E[|I|] = \sum_v Pr(v \in I)$. If $d_v = 0$, then $v \in I$ with probability 1. Hence, we assume that $d(v) \geq 1$. ¿From the definition of I, it follows that $v \in I$ if and only if for every e incident at v, $e \setminus \{v\} \not\subseteq S_v = \{u \in V(L_v) : u \prec v\}$. In other words, S_v is an independent set in L_v. Let $l_v = |V(L_v)|$. Then

$$Pr[v \in I] = \sum_{J \in \mathcal{I}(L_v)} \frac{|J|!(l_v - |J|)!}{(l_v+1)!} = \frac{1}{l_v+1} \sum_{J \in \mathcal{I}(L_v)} \frac{1}{\binom{l_v}{|J|}}$$

Applying Theorem 8.2.1 to the t-uniform link graph L_v (with $c = c_{k-1}$), we get

$$Pr[v \in I] \geq \frac{c}{l_v+1}\left(\frac{l_v}{d(v)^{1/(k-1)}}\right) \geq \frac{cl_v}{l_v+1}\left(\frac{1}{(d(v)+1)^{1/(k-1)}}\right).$$

Since $l_v \geq k-1$, we get $Pr[v \in I] \geq ((k-1)c/k)\frac{1}{(d(v)+1)^{1/(k-1)}}$. By choosing $d_k = (k-1)c/k$, we get the lower bound of the theorem. □

8.3 Linearity : Probability of having no backward edges

In this section, we state and prove a warm-up result on the probability of having no backward edges incident at a vertex for a randomly chosen linear ordering (Theorem 8.3.1 below). The problem is the same as in the previous section, only, now the hypergraph under consideration is assumed to be linear and we get an explicit closed-form expression for this probability. This result will be helpful for the proof of the main technical theorem, given in the next section.

Theorem 8.3.1 *Let H be a linear k-uniform hypergraph and let v be an arbitrary vertex having degree d. For a uniformly chosen total ordering $\prec$ on V, the probability $P_v(0)$ that v has no backward edge incident at it, is given by*

$$P_v(0) = \frac{1}{\binom{d+1/t}{d}}$$

Remark. It is interesting to note that the above expression when summed over all vertices, is the same bound which Caro and Tuza obtain in [21] (using very different methods), although their bound holds for independent sets in *general* k-uniform hypergraphs. This apparent coincidence is explained (and proved) in section 8.4, thus giving a probabilistic proof of the exact Caro-Tuza expression.

We prove the theorem using the well-known Principle of Inclusion and Exclusion (PIE). First we state an identity involving binomial coefficients.

Lemma 8.3.2 *Given non-negative integers d and t,*

$$\sum_{r=0}^{d} (-1)^r \binom{d}{r} \frac{1}{tr+1} \quad = \quad \frac{1}{\binom{d+1/t}{d}}$$

For proof see [37], Equation 5.41. Alternatively, it can be proved using the Chu-Vandermonde identity (see e.g. [37], Equation 5.93), as shown below:

Proof of Lemma 8.3.2 Write the LHS as $\sum_{r\geq 0} t_r$, since $\binom{d}{r} = 0$ for $r > d$. Now,

$$\begin{aligned} \frac{t_{r+1}}{t_r} &= \frac{(-1)(d-r)(tr+1)}{(r+1)(tr+t+1)} \\ &= \frac{(r-d)(r+1/t)}{(r+1)(r+1+1/t)} \end{aligned}$$

Also, notice that $t_0 = 1$. Therefore, the LHS can be written as the generalised hypergeometric function $F(1/t, -d; 1 + 1/t; 1)$, where the generalised hypergeometric function $F(a_1, ..., a_m; b_1, ..., b_n; z)$ is given by

$$F(a_1, ..., a_m; b_1, ..., b_n; z) \quad = \quad \sum_{r=0}^{\infty} \frac{(a_1)^{(r)}(a_2)^{(r)}...(a_m)^{(r)}}{(b_1)^{(r)}...(b_n)^{(r)}} \frac{z^r}{r!}$$

where $p^{(q)} = p(p+1)...(p+q-1)$ is the *rising factorial*. Next, we use the general version of Vandermonde convolution - also known as Chu-Vandermonde identity (a special case of Gauss's Hypergeometric Theorem, see e.g. [37], Chapter 5, equation 5.93, also [8, 10, 13, 73])

$$F(a, -n; c; 1) = \frac{(c-a)^{(n)}}{c^{(n)}}$$

The above is true whenever a, c are complex numbers and n is a natural number, such that $\mathbb{R}(a) - n < \mathbb{R}(c)$. In our case, $a = 1/t$, $n = d$ and $c = 1 + 1/t$. Hence we get $(c-a)^{(n)} = d!$, and $c^{(n)} = (1+1/t)^{(d)} = (1+1/t)(2+1/t)...(d+1/t)$. Therefore, the LHS of (8.3) becomes

$$F(1/t, -d; 1 + 1/t; 1) = \frac{d!}{(1+1/t)(2+1/t)...(d+1/t)} = \frac{1}{\binom{d+1/t}{d}} \qquad \square$$

Proof of Theorem 8.3.1 Firstly, observe that since H is linear, the number of vertices that are neighbors of v is exactly $(k-1)d = td$. Next, notice that since the random ordering is uniformly chosen, only the relative arrangement of these td neighbors and the vertex v, i.e. $td+1$ vertices in all, will determine the required probability. Hence the total number of orderings under consideration is $(td+1)!$.

Label the hyperedges incident at v with $1, ..., d$ arbitrarily. For a permutation π, we say that π has the property $T_{\geq S}$ if the edges with labels in S, $S \subseteq [d]$ are backward. Also, say π has the property $T_{=S}$ if the edges with labels in S are backward and no other edges are backward. For a set S of hyperedges incident at v, let $N(T_{\geq S})$ denote the number of orderings having the property $T_{\geq S}$, that is, the number of permutations such that the hyperedges in S will *all* be backward edges. $N(T_{=S})$ is similarly defined. $N(T_{\geq S})$ is determined as follows : Suppose S has r hyperedges incident at v. For a *fixed* arrangement of the vertices belonging to edges in S, the number of permutations of the remaining vertices is $(td+1)!/(tr+1)!$. In each allowed permutation, the vertex v must occur only after the vertices of S (i.e. the rightmost position). However the remaining tr vertices can be arranged among themselves in $(tr)!$ ways. Thus we have

$$N(T_{\geq S}) = (td+1)! \frac{(tr)!}{(tr+1)!} = \frac{(td+1)!}{(tr+1)}.$$

Clearly, if a permutation has the property $T_{\geq S}$, it has the property $T_{=S'}$ for some $S' \supseteq S$. Hence for every $S \subset [d]$,

$$N(T_{\geq S}) = \sum_{S' \supseteq S} N(T_{=S'}).$$

Therefore, by PIE (see [62], Chapter 2),

$$N(T_{=\emptyset}) = \sum_{S} (-1)^{|S|} N(T_{\geq S}).$$

$$\sum_{|S|=r} N(T_{\geq S}) \;=\; \binom{d}{r} N(T_{\geq [r]}) \;=\; \binom{d}{r} \frac{(td+1)!}{tr+1}.$$

Hence we get the required probability to be

$$\begin{aligned} P_v(0) &= \left(\sum_{r=0}^{d} \binom{d}{r} (-1)^r \frac{(td+1)!}{tr+1} \right) \times \frac{1}{(td+1)!} \\ &= \sum_{r=0}^{d} \binom{d}{r} (-1)^r \frac{1}{tr+1}. \end{aligned}$$

By Lemma 8.3.2,

$$P_v(0) = \frac{1}{\binom{d+1/t}{d}},$$

and this completes the proof. □

8.4 Probabilistic proof of the Caro-Tuza lower bound expression

In this section, we extend the result of the previous section (for linear k-uniform hypergraphs) to general k-uniform hypergraphs, thus obtaining an alternate proof of Caro and Tuza's [21] Theorem 8.1.1. First we state a generalized version of the problem:

Problem Given an s-uniform hypergraph S over a universe $\mathcal{U}$ of n vertices $\{a_1, \ldots, a_n\}$ and d hyperedges $C_1, \ldots, C_d$ such that $C_1 \cup \ldots \cup C_d = \mathcal{U} \setminus \{a_n\}$, what fraction $fr(S)$ of orderings of the vertices of S will have no edge occurring entirely before a_n?

The following lemma shows that the exact expression for this fraction obtained in the previous section (assuming linearity) is actually a lower bound for the more general case.

Lemma 8.4.1 *With the parameters defined as above:* $fr(S) \geq f_s(d) = \binom{d+1/s}{d}^{-1}$.

To get some idea of why this might be true, let us first look at an example:

Example : Take $r = 4, s = 2, m = 0$, such that $S = \{\{1,2\},\{2,3\},\{3,4\},\{4,1\}\}$. Then $fr(S) = 56/120$. On the other hand, $f_s(d) = 128/315 < 56/120 = fr(S)$.

Lemma 8.4.1 implies Equation 8.1 as follows: Let the hypergraph $S = S_u$ in Lemma 8.4.1 be the link hypergraph L_u corresponding to a vertex u (as defined in Section 8.2), with $a_n := u$, with $s := k-1$, and $d := d(u)$. As in the proof of Theorem 8.1.1, let I be the independent set formed by choosing a uniform random ordering of $V(H)$ and studying the size of the set I consisting of those vertices which have no backward edges with respect to the random ordering. Then the probability that $u \in I$ is exactly the fraction $fr(S)$ of orderings having no edge occuring entirely before a_n. Let $X = |I|$. Then by linearity of expectation, the expected size of I is given by

$$E[X] \;=\; \sum_{v \in V(H)} \mathbf{Pr}[v \in I] = \sum_{v \in V(H)} fr(S_v) \geq \sum_{v \in V(H)} f_s(d(v))$$

thereby yielding an alternate probabilistic proof of 8.1.

Proof of Lemma 8.4.1 Define $l = l(S) = \sum_{i \neq j} |C_i \cap C_j|$. The proof proceeds by induction on l. The base case is $l = 0$ corresponding to the case when any two sets C_i, C_j, $j \neq i$ are disjoint and this case has already been handled in Section 8.3.

Assume the Lemma for $l \leq r \in \mathbb{N}$, and consider S such that $l(S) = r+1$. The idea is to 'disconnect', in some sense, the edges, vertex-by-vertex, and show that at each step the fraction $fr(S)$ does not increase. Choose a vertex a which belongs to more than one edge, say C_i and some other edge C_j. In order to decrease the intersection of C_i with C_j by at least one, we add a new vertex a' to the universe $\mathcal{U}$ which will belong only to the edge C_i, and remove a from C_i, to get a new s-uniform hypergraph S' which satisfies $l(S') \leq l(S) - 1$. Since the number of elements in the universe has gone up by 1, the total number of orderings increases by a factor of $n+1$. So it only remains to show that the number of favorable orderings, that is, orderings with no backward edge increases by a factor of at most $n+1$. Formally, apply the following procedure:

Procedure: *Linearize*$(S, \mathcal{U})$

(i) If $l(S) = 0$, then return LINEAR.

(ii) Choose some $a \in C_i \cap C_j$ for some $i \neq j$;

(iii) $\mathcal{U}' := \mathcal{U} \cup \{a'\}$; $C_i' := (C_i \cup \{a'\}) \setminus \{a\}$; $S' := (S \cup \{C_i'\}) \setminus \{C_i\}$.

(iv) Return $(S', \mathcal{U}')$.

Claim 8.4.2 *With $(S,\mathcal{U})$ as defined above and $(S',\mathcal{U}') = Linearize(S,\mathcal{U})$, we have $fr(S) \geq fr(S')$.*

Lemma 8.4.1 now follows from Claim 8.4.2 (whose proof is provided below), the fact that $l(S') \leq l(S) - 1$ and by applying induction on l. □

Proof of Claim 8.4.2 Let $Perm(S)$ be the set of all linear orderings of the vertices of S, that is, bijections $\prec : S \to [|S|]$. We use the shorthand $a \prec b$ to indicate $\prec(a) < \prec(b)$. Given an ordering $\prec$ and a subset $C \subset \mathcal{U} \setminus \{a_n\}$, call C *backward* if for all $a_j \in C$, $a_j \prec a_n$, otherwise call C *forward*. Similarly, a vertex $a \in C$ is a a *backward* vertex if $a \prec a_n$ and is a *forward* vertex otherwise. Call $\prec \in Perm(S)$ *good* if the number of backward edges in $\prec$ is zero. Let

$$\begin{aligned} Good(S) &:= \{\prec \in Perm(S) : \prec \text{ is good } \} \\ Bad(S) &:= Perm(S) \setminus Good(S) \end{aligned}$$

Then by definition, $fr(S) = |Good(S)|/|Perm(S)|$. Clearly, $|Perm(S')| = (n+1)|Perm(S)|$. So, it suffices to prove that $|Good(S')| \leq (n+1)|Good(S)|$ to prove the claim. Let $\prec' \in Perm(S')$. Let $\prec$ be the projection of $\prec'$ into $Perm(S)$, obtained by removing the extra vertex a'. Notice that the only edge of S' whose status (forward or backward) can depend on the position of the new element is the edge C'_i; all other edges are unaffected by this addition. Consider the following cases of an arbitrary $\prec \in Perm(S)$:

(i) $\prec$ has no backward edges and at least one vertex of C_i other than a is a forward vertex. Such an ordering is the projection of exactly $n+1$ good orderings of S', since a' could be placed in any position in $\prec$ to get a good ordering of S'.

(ii) $\prec$ has no backward edges with a as the only forward vertex of C_i. Such an ordering is the projection of a bad ordering of S' if and only if a' is placed as a backward vertex. Call the collection of all such orderings $\prec' \in Bad(S')$ whose projection is an ordering of this Type (ii) as $Loss(S')$.

(iii) $\prec$ has C_i as the only backward edge. Such an ordering is in $Bad(S)$, but it can be the projection of a good ordering of S' if and only if a' is placed as a forward vertex. Call the collection of all such good orderings of S' whose projections are Type (iii) orderings of S as as $Gain(S')$.

(iv) $\prec$ is none of the three types above. Such an ordering is the projection of only bad orderings of S'.

Types (i) and (ii) are all the orderings belonging to $Good(S)$, and these together are the projections of $(n+1)|Good(S)| - |Loss(S')|$-many good orderings of S'. Clearly,

$$|Good(S')| = (n+1).|Good(S)| - |Loss(S')| + |Gain(S')| \tag{8.3}$$

To complete the proof, it suffices to show that $|Loss(S')| \geq |Gain(S')|$. But notice that an ordering in $Gain(S')$ can be converted into a unique ordering in $Loss(S')$, simply by exchanging the positions of the vertices a and a'. Thus there exists an injective map from $Gain(S')$ to $Loss(S')$ and hence $|Gain(S')| \leq |Loss(S')|$. So we get that $fr(S') \leq fr(S)$. $\square$

8.5 Linearity : Probability of having few backward edges

Now, we consider the more general case when at most $A-1$ backward edges are allowed. In this section, we get an exact expression for the corresponding probabiity. This estimate plays an important role later in getting new and improved lower bounds on $\alpha(H)$ for locally sparse graphs and linear hypergraphs. Our goal in this section is to prove the following result.

Theorem 8.5.1 *For a k-uniform linear hypergraph H, a vertex v having degree d, a uniformly chosen permutation π induces at most $A-1$ backward edges with probability $P_v(A-1)$ given by*

$$P_v(A-1) = \begin{cases} 1 & \text{if } d \leq A-1; \\ \frac{tA}{tA+1}\frac{\binom{d}{A}}{\binom{d+1/t}{d-A}} & \text{if } d \geq A. \end{cases}$$

Corollary 8.5.2 *As $d \to \infty$, the asymptotic expression for the probability $P_v(A-1)$ is given by*

$$P_v(A-1) \;\sim\; \frac{1}{1+(1/(tA))}\left(\frac{A}{d}\right)^{1/t} \;=\; \Omega((A/d)^{1/t})$$

Proof The asymptotics are w.r.t. $d \to \infty$, $d \geq A$. The expression for having at most $A-1$ backward edges is

$$\begin{aligned} P_v(A-1) &= \frac{1}{1+(tA)^{-1}}\frac{d(d-1)...(A+1)}{(d-A)!}\frac{(d-A)!}{(d+1/t)(d-1+1/t)...(A+1+1/t)} \\ &= \frac{1}{1+(tA)^{-1}}\frac{1}{(1+1/td)(1+(t(d-1))^{-1})...(1+(t(A+1))^{-1})} \end{aligned}$$

Now, for $0 < x$, we have $(1+x)^{-1} > e^{-x}$. So we get

$$\begin{aligned} P_v(A-1) &> (1+(tA)^{-1})^{-1} e^{(-1/t)\sum_{r=A+1}^{d}(1/r)} \\ &= (1+(tA)^{-1})^{-1} e^{(-1/t)[\sum_{r=1}^{d}(1/r)-\sum_{r=1}^{A}(1/r)]} \\ &= (1+(tA)^{-1})^{-1} e^{(-1/t)[\ln d - \ln A]+O((d-A)/(tdA))} \\ &= (1+(tA)^{-1})^{-1} e^{(-1/t)\ln(d/A)-O((d-A)/(tdA))} \\ &= (1+(tA)^{-1})^{-1}(A/d)^{1/t}\Omega(1) \\ &= \Omega((A/d)^{1/t}) \end{aligned}$$

The above expression therefore becomes $\Omega((A/d)^{1/t})$. □

The version of PIE used most commonly deals with $N(T_{=\emptyset})$, i.e. the number of elements in the set of interest - in this case, permutations of $[td+1]$ which do not have *any* of the properties under consideration (in this case, backward edges with respect to v). However we need something slightly different - an expression for the number of permutations which have *at least* A backward edges. Clearly, the remaining permutations are those which have *at most* $A-1$ backward edges.

Therefore, we use a slightly modified version of PIE, which is stated below in Theorem 8.5.5. This form is well-known (see e.g. [62], Chapter 2, Exercise 1), although it seems to be used less frequently. For the sake of completeness, we provide a simple proof. First we state and prove two identities involving binomial coefficients:

Lemma 8.5.3 *For a, b nonnegative integers,*

$$\sum_{i=0}^{b}(-1)^i\binom{a+b}{a+i}\binom{a+i-1}{i} = 1$$

Proof of Lemma 8.5.3 The proof is by induction on b. For $b=0$, the LHS reduces to

$$\sum_{i=0}^{0}(-1)^i\binom{a+0}{a+i}\binom{a+i-1}{i}$$

which is clearly 1. Assume the lemma to be true for $b=c$ and consider the LHS when $b=c+1$:

$$\sum_{i=0}^{c+1}(-1)^i\binom{a+1+c}{a+i}\binom{a+i-1}{i}$$

$$
\begin{aligned}
&= \sum_{i=0}^{c+1}(-1)^i\left[\binom{a+c}{a+i}+\binom{a+c}{a+i-1}\right]\binom{a+i-1}{i}\\
&= \sum_{i=0}^{c+1}(-1)^i\left[\binom{a+c}{a+i}\binom{a+i-1}{i}+\binom{a+c}{a+i-1}\binom{a+i-1}{i}\right]\\
&= 1+\sum_{i=0}^{c+1}(-1)^i\binom{a+c}{a+i-1}\binom{a+i-1}{i}
\end{aligned}
$$

by the induction hypothesis, since $\binom{a+c}{a+c+1}=0$. Now, the second sum is

$$
\begin{aligned}
&\sum_{i=0}^{c+1}(-1)^i\binom{a+c}{a+i-1}\binom{a+i-1}{i}\\
&= \frac{(a+c)!}{(a-1)!(c+1)!}\sum_{i=0}^{c+1}(-1)^i\binom{c+1}{i}\\
&= 0
\end{aligned}
$$

□

Lemma 8.5.4 *Given non-negative integers d, A, $d \geq A$ and a positive integer t,*

$$
\sum_{r=0}^{d-A}(-1)^r\binom{d}{r+A}\binom{A+r-1}{r}\frac{1}{t(r+A)+1} = 1-\left(\frac{At}{tA+1}\right)\frac{\binom{d}{A}}{\binom{d+1/t}{d-A}}
$$

Proof of Lemma 8.5.4 Let the LHS be denoted by S_d. Then, using the identity $\binom{n}{r} = \binom{n-1}{r}+\binom{n-1}{r-1}$, we have

$$
\begin{aligned}
S_d &= \sum_{r=0}^{d-A}(-1)^r\left[\binom{d-1}{r+A}+\binom{d-1}{r+A-1}\right]\binom{A+r-1}{r}\frac{1}{tr+tA+1}\\
&= S_{d-1}+\sum_{r=0}^{d-A}(-1)^r\binom{d-1}{r+A-1}\binom{A+r-1}{r}\frac{1}{tr+tA+1}
\end{aligned}
$$

since $\binom{d-1}{d}=0$. Now the second sum can be simplified as

$$
\begin{aligned}
T_d &= \sum_{r=0}^{d-A}(-1)^r\binom{d-1}{r+A-1}\binom{A+r-1}{r}\frac{1}{tr+tA+1}\\
&= \left(\frac{(d-1)!}{(d-A)!(A-1)!}\right)\sum_{r=0}^{d-A}(-1)^r\binom{d-A}{r}\frac{1}{tr+tA+1}\\
&= \binom{d-1}{A-1}\frac{1}{tA+1}\sum_{r=0}^{d-A}(-1)^r\binom{d-A}{r}\frac{1}{(t/(tA+1))r+1}
\end{aligned}
$$

By Lemma 8.3.2, we get

$$T_d = \frac{1}{tA+1}\frac{\binom{d-1}{A-1}}{\binom{d-A+(tA+1)/t}{d-A}}$$

Therefore,

$$S_d = S_{d-1} + \frac{1}{tA+1}\frac{\binom{d-1}{A-1}}{\binom{d+1/t}{d-A}}$$

Unraveling the recursion and noticing that $S_A = 1/(tA+1)$, we get that

$$\begin{aligned} S_d &= (1/(tA+1))\sum_{r=0}^{d-A}\frac{\binom{d-r-1}{A-1}}{\binom{d+1/t-r}{d-A-r}} \\ &= (1/(tA+1))\sum_{r=0}^{d-A}\frac{\binom{A-1+r}{A-1}}{\binom{A+1/t+r}{r}} \end{aligned}$$

by reversing the order of summation. Finally, the following claim completes the proof.

Claim. For $d \geq A, t \geq 0$,

$$\frac{1}{tA+1}\sum_{r=0}^{d-A}\frac{\binom{A-1+r}{A-1}}{\binom{A+1/t+r}{r}} = 1 - \frac{tA}{tA+1}\frac{\binom{d}{A}}{\binom{d+1/t}{d-A}}$$

Proof of Claim. We use induction on d. When $d = A$, the LHS is $(tA+1)^{-1}$, while the RHS is $1 - \frac{At}{tA+1}$, so we have equality. Now assume equality for d and consider the LHS for $d+1$:

$$\begin{aligned} &\frac{1}{tA+1}\sum_{r=0}^{d-A+1}\frac{\binom{A-1+r}{r}}{\binom{A+1/t+r}{r}} \\ =\ & 1 - \frac{At}{tA+1}\frac{\binom{d}{A}}{\binom{d+1/t}{d-A}} + (At+1)^{-1}\frac{\binom{d}{d-A+1}}{\binom{d+1+1/t}{d-A+1}} \\ =\ & 1 - \frac{At}{(tA+1)\binom{d+1+1/t}{d-A+1}}\left[\binom{d}{A}\frac{d+1+1/t}{d-A+1} - (At)^{-1}\binom{d}{d-A+1}\right] \\ =\ & 1 - \frac{At}{(tA+1)\binom{d+1+1/t}{d-A+1}}\left[\binom{d+1}{A} + \binom{d+1}{d-A+1}(t(d+1))^{-1} - (At)^{-1}\binom{d}{A-1}\right] \\ =\ & 1 - \frac{At}{(tA+1)\binom{d+1+1/t}{d-A+1}}\binom{d+1}{A} \end{aligned}$$

which is the required expression on the RHS. □

We now present the generalized PIE and its well-known proof.

Theorem 8.5.5 *Let S be an n-set and $E_1, E_2, ...E_d$ not necessarily distinct subsets of S. For any subset M of $[d]$, define $N(M)$ to be the number of elements of S in $\cap_{i\in M} E_i$ and for $0 \le j \le d$, define $N_j := \sum_{|M|=j} N(M)$. Then the number $N_{\ge a}$ of elements of S in* at least *a, $0 \le a \le d$ of the sets E_i, $1 \le i \le d$, is*

$$N_{\ge a} = \sum_{i=0}^{d-a} (-1)^i \binom{a+i-1}{i} N_{i+a} \quad ... \quad (MPIE)$$

Proof Take an element $e \in S$.

(i) Suppose e is in no intersection of at least a E_i's. Then e does not contribute to any of the summands in the RHS of the equation (MPIE), and hence, its net contribution to the RHS is zero.

(ii) Suppose e belongs to exactly $a+j$ of the E_i's, $0 \le j \le d-a$. Then its contribution to the RHS of (MPIE) is

$$\sum_{l=0}^{j} (-1)^l \binom{a+j}{a+l} \binom{a+l-1}{l}$$

and by Lemma 8.5.3 this is equal to 1. □

Proof of Theorem 8.5.1 If $d \le A-1$, then $P_v(A-1) = 1$ obviously. The proof is similar to the proof of Theorem 8.3.1, except that in place of the PIE, we use Theorem 8.5.5. The set under consideration is the set of permutations of $[td+1]$, the subsets E_i correspond to the permutations for which the i-th edge is backward. It is easy to see that $N(M) = N(T_{\ge M})$ under the notation used in Theorem 8.3.1 and hence $N(M) = \frac{(td+1)!}{t|M|+1}$. Therefore we have $N_j = \binom{d}{j} \frac{(td+1)!}{tj+1}$ as before. Hence the expression for the probability $Q_v(A)$ that *at least* A edges *are* backward under a uniformly random permutation π, becomes:

$$Q_v(A) \;=\; \sum_{i=0}^{d-A} (-1)^i \binom{d}{i+A} \binom{A+i-1}{i} \frac{1}{t(i+A)+1}.$$

By Lemma 8.5.4 the RHS of the above expression is

$$Q_v(A) \;=\; 1 - \left(\frac{1}{1+(tA)^{-1}}\right) \frac{\binom{d}{A}}{\binom{d+1/t}{d-A}}.$$

Hence the probability of having at most $A-1$ backward edges is given by

$$P_v(A-1) \;=\; \frac{1}{1+(tA)^{-1}}\frac{\binom{d}{A}}{\binom{d+1/t}{d-A}}$$

and the proof is complete. □

8.6 Lower bounds for linear hypergraphs and K_r-free graphs

In this section we prove Theorems 8.1.2, 8.1.3, and 8.1.4. These follow by a simple application of Corollary 8.5.2. Since the proofs follow the same outline, we prove them simultaneously, highlighting only the differences as and when they occur.

Proofs of Theorems 8.1.2, 8.1.3 and 8.1.4. Consider a uniformly chosen random permutation of the vertices of the graph/hypergraph under consideration. Let D be the average degree of the graph or hypergraph and $A = D^{\epsilon}$. Let I be the set of those vertices each having at most $A-1$ backward edges incident on it. Clearly, the expected size of I is

$$E[|I|] \;=\; \sum_{v\in V} P_v(A-1) \;\geq\; c\sum_{v\in V}\left(\frac{A}{\max\{A,d(v)\}}\right)^{1/t} \;=\; cA^{1/t}\sum_{v\in V}\left(\frac{1}{\max\{A,d(v)\}}\right)^{1/t}$$

for some constant $c = c(k,\epsilon)$. (For a graph, $k=2$ and hence $t=1$). Also, by construction, the average degree of the sub(hyper)graph induced by I is at most $k(A-1)$. Therefore, there exists an independent set I' of size at least as follows

(i) Case $t=1$, graph is K_3-free: By [56], $\alpha(G)$ is at least

$$\Omega\left(\log(2(A-1))\frac{|I|}{2(A-1)}\right) \;=\; \Omega\left(\log D\sum_{v\in V}\frac{1}{\max\{A,d(v)\}}\right)$$

(ii) Case $t=1$, graph is K_r-free ($r>3$): By [58], $\alpha(G)$ is at least

$$\Omega\left(\frac{\log(2(A-1))}{\log\log(2(A-1))}\frac{|I|}{2(A-1)}\right) \;=\; \Omega\left(\frac{\log D}{\log\log D}\sum_{v\in V}\frac{1}{\max\{A,d(v)\}}\right)$$

(iii) Case $t>1$, hypergraph is linear: By [22], $\alpha(H)$ is at least

$$\Omega\left((\log k(A-1))^{1/t}\frac{|I|}{(k(A-1))^{1/t}}\right) \;=\; \Omega\left((\log D)^{1/t}\sum_{v\in V}\frac{1}{(\max\{A,d(v)\})^{1/t}}\right)$$

The above three cases prove Theorems 8.1.2, 8.1.3 and 8.1.4 respectively.

Note: An inspection of the proofs above show why we need ϵ to be a fixed constant. It is because all three expressions above essentially have $\log A$ i.e. $\epsilon \log D$ in the numerator. So, if $\epsilon = o(1)$, then $\log A = o(\log D)$, and we would get asymptotically weaker results. □

8.7 Construction comparing average degree vs. degree sequence based bounds

A degree sequence-based bound obviously reduces to a bound based on average degree, when the (hyper)graph is regular. However, the convexity of the function $x^{-1/t}$, $x \geq 1$ and $t \in \mathbb{N}$, shows that the bounds in Theorems 8.1.2, 8.1.3 and 8.1.4 are better than the corresponding average degree-based bounds proved in [4], [56] and [58] respectively provided the minimum degree is at least A, although it is not clear *a priori* if the improvement can become significantly larger. Also, at least half the vertices will have degree at most $2D$, so even in the general case (no restriction on the minimum degree) our bounds are no worse than the average degree based bounds (ignoring the constant factors). In fact, they can be much larger than the latter bounds. We now give infinite families of K_r-free graphs and linear k-uniform hypergraphs which show that

(i) The bounds given by Theorem 8.1.2, 8.1.3 can be better than the bounds in [5, 56, 58] respectively by a multiplicative factor of $\log(|V(G)|)$.

(ii) The bound in Theorem 8.1.4 can be better than the bound in [4] by a multiplicative factor of $((\log|V(H)|)/(\log\log|V(H)|))^{(1-\epsilon)/t}$, where ϵ is the constant mentioned in Theorem 8.1.4.

Case (i) Take a set of n disjoint graphs, $K_{1,1}$, $K_{2,2}$, $K_{4,4}$, ... , $K_{2^{n-1},2^{n-1}}$. For each $i \in [n]$, join one of the parts of the component $K_{2^i,2^i}$ to one of the parts in $K_{2^{i-1},2^{i-1}}$, by introducing a complete bipartite graph between them. (Use the other part of $K_{2^i,2^i}$ for joining to $K_{2^{i+1},2^{i+1}}$). Let G denote the resulting connected triangle-free graph.
The total number of vertices is $2^{n+1} - 2$, whereas the average degree is

$d_{av} = 2|E(G)|/|V(G)| = (2^n + 1)/2 - o(1)$. Hence, the average degree based bound gives $\Theta(|V(G)| \log d_{av}/d_{av}) = \Theta(\log d_{av})$. Denote by l the maximum j such that $3.2^j \leq A < 3.2^{j+1}$, where $A := d_{av}^{\epsilon}$. For every fixed $\epsilon \in (0,1)$, we have $n - l = \Theta(n)$. Theorem 8.1.2 gives

$$c \log d_{av} \sum_{v \in V} \frac{1}{\max\{d(v), A\}} \quad = \quad c \log d_{av} \left[\frac{1}{A} + \sum_{j=1}^{l} \frac{3.2^j}{A} + \sum_{j=l+1}^{n-2} \left(\frac{3.2^j}{3.2^j} \right) + \frac{2^{n-1}}{2^{n-1}} \right]$$

$$= c(\log d_{av})\,[\Theta(1) + \Theta(n)]$$
$$= c(\log d_{av})\Theta(\log(|V(G)|))$$

The same example works for Theorem 8.1.3 also, since triangle-free graphs are obviously K_r-free, for $r \geq 3$.

Case (ii) Fix some $m = m(n) = k^{2^n}$. For each $i \in \{0, \ldots, n-1\}$, first create a connected linear hypergraph as follows: Take the vertex set as $[k]^{2^i}$, i.e. the set of 2^i-dimensional vectors with each co-ordinate of a vector taking values in $\{1, 2, \ldots, k\}$. Let each hyperedge consists of the k vertices which have all but one co-ordinate fixed. Call this hypergraph an *i-unit*. It can be verified easily that each i-unit is k-uniform and 2^i-regular. Now for each i, create an *i-component* as follows:

(i) Take $m_i = \lceil \frac{m}{k^{2^i}} \rceil$ disjoint unions of i-units and linearly order them, say $i_1, \ldots, i_{m_i}$.

(ii) Consider the sets of vertices of size k formed by choosing at most one vertex from each i-unit. Add such edges greedily, ensuring the following:

  (i) No vertex belongs to more than one such edge;

  (ii) Choose the first edge from i-units $i_1, \ldots, i_k$, the second one from $i_2, \ldots, i_{k+1}$, etc. - in general the j-th such edge has one vertex from each of the i-units $i_{j \pmod{m_i}}, i_{j+1 \pmod{m_i}}, \ldots i_{j+k-1 \pmod{m_i}}$.

Each i-component is a connected, linear k-uniform graph and the degree of every vertex is either 2^i or $2^i + 1$. Take the disjoint unions of n such i-components, one for every $i \in \{0, \ldots, n-1\}$, to get the hypergraph $H = H(n) = (V, E)$. For each $j \in \{0, \ldots, n-2\}$, greedily add a maximal matching between components j and $j+1$, with each edge taking only one vertex from component j (and remaining $k-1$ from the component $j+1$), and no vertex belonging to more than one such edge. Let G be the resulting connected, linear k-uniform graph. The total number of vertices in the j-th component is $m_j \cdot k^{2^j} = m(1 + o(1))$, and hence $|V| = nm(1 + o(1))$. Also, the average degree is $d_{av} \sim (2^n - 1)/n \sim 2^n/n$. Let l denote the greatest integer j such that $2^j \leq (d_{av})^\epsilon \sim 2^{\epsilon n}/n^\epsilon$. Therefore the average degree based bounds in [4, 22] give a lower bound of

$$\alpha(H) = \Omega(mn^{1+1/t}(\log d_{av})^{1/t}/2^{n/t}) \quad \ldots \quad (A)$$

Notice that the degree of any vertex in the i-th component (after G has been constructed) is always between 2^i and $2^i + 3$. For a vertex v such that $d(v) < d_{av}^\epsilon$, the actual degree does not play a role in the expression in Theorem 8.1.4. For vertices v

such that $d(v) \geq d_{av}^{\epsilon}$, this increase is negligible $(2^{\epsilon n}/n^{\epsilon} + 3) \sim 2^{\epsilon n}/n^{\epsilon}$. Therefore, the bound in Theorem 8.1.4 gives

$$\begin{aligned}
\alpha(H) &= \Omega\left((\log d_{av})^{1/t}\left[\sum_{j=0}^{l}\frac{mn^{\epsilon/t}}{2^{\epsilon n/t}} + \sum_{j=l+1}^{n-1}\frac{m}{2^{j/t}}\right]\right) \\
&= \Omega\left(m(\log d_{av})^{1/t}\left[\epsilon 2^{-\epsilon n/t}n^{1+\epsilon/t} + 2^{-\epsilon n/t}n^{\epsilon/t}\frac{(1-2^{-(n-l-1)/t})}{1-2^{-1/t}}\right]\right) \\
&= \Omega\left(m(\log d_{av})^{1/t} \times 2^{-\epsilon n/t}\left[\epsilon n^{1+\epsilon/t} + n^{\epsilon/t}\frac{(1-2^{-(n-l-1)/t})}{1-2^{-1/t}}\right]\right) \\
&= \Omega\left(m(\log d_{av})^{1/t} \times 2^{-\epsilon n/t}(\epsilon n^{1+\epsilon/t} + \Theta(n^{\epsilon/t}))\right) \quad \ldots \quad (B)
\end{aligned}$$

The ratio of the bound in (B) to the one in (A) can be seen to be $\Omega((2^n/n)^{(1-\epsilon)/t})$, which is $\Omega((\log|V|/\log\log|V|))^{(1-\epsilon)/t})$.

8.8 Binomial Identities

In the course of this chapter, certain non-trivial binomial identities were also obtained, with semi-combinatorial proofs. Some of the identities are new, to the best of our knowledge, and may be of independent interest. These are described below:

$$\sum_{a=0}^{A}\sum_{i=0}^{d-a}\binom{d}{a+i}\binom{a+i}{i}2^i(2d-2a-i)!(2a+i)! = (d!)^2 4^{d-A}(A+1)\binom{2A+1}{A} \quad (8.4)$$

The LHS (when divided by $(2d+1)!$) amounts to the expression for $P_v(A)$ when $k = 3$: choose $a + i$ hyperedges from the d hyperedges incident on v, of these a hyperedges are backward, while i hyperedges each have one vertex occurring prior to v in the random permutation. These i vertices can be chosen from i pairs in 2^i ways. The $(2a + i)$ vertices before v can be arranged in $(2a + i)!$ ways amongst themselves. The remaining $(2d - 2a - i)$ vertices occur after v and can be arranged amongst themselves in $(2d - 2a - i)!$ ways. The RHS is easily obtained from Theorem 8.5.1 by taking $t = 2$.

Even the $A = 0$ case of the above identity gives (after some rearrangements):

$$\sum_{i=0}^{d}\binom{d+i}{d}2^{-i} = 2^d$$

The above identity merits discussion in some detail in [37] (Chapter 5, eqs. 5.20, 5.135-8); a nice combinatorial proof of it is provided in [65].

The next identity (for the more general case $k \geq 3$) is much more complicated. Given $i \in \mathbb{Z}^+$, let C_i^{t-1} denote the set of all solutions in non-negative integers $\mathbf{j} = (j_1, j_2, \ldots, j_{t-1})$ of the equation $\mathbf{j} \cdot \mathbf{1} = i$ i.e. $C_i^{t-1} := \{(j_1, \ldots, j_{t-1}) : \sum_{l=1}^{t-1} j_l = i \ ; \ \forall l \ j_l \geq 0\}$. Then

$$\sum_{a=0}^{A} \sum_{i=0}^{d-a} \sum_{\mathbf{j} \in C_i^{t-1}} \left[\binom{d}{a+i} \binom{a+i}{a, j_1, \ldots, j_{t-1}} (ta + \sum_{s=1}^{t-1} s.j_s)!(td - (ta + \sum_{s=1}^{t-1} s.j_s))! \prod_{r=1}^{t-1} \binom{t}{r}^{j_r} \right]$$
$$= (td+1)!(1 + (tA+t)^{-1})^{-1} \left[\binom{d}{A+1} \Big/ \binom{d+1/t}{d-A-1} \right] \quad (8.5)$$

The LHS again follows by similar arguments as for (8.4), this time for general t. There are a backward edges, i_1 edges which have one vertex before v, i_2 edges with 2 vertices before v, and so on. The RHS follows from Theorem 8.5.1.

Our proof techniques for identities (8.4, 8.5) involving PIE, are non-standard. It may be an interesting problem in Enumerative Combinatorics to come up with combinatorial proofs of the identities (8.4, 8.5). In particular, for (8.5), it would be interesting to come up with proofs using *any* standard technique such as induction, generating functions, the WZ method etc.

8.9 Concluding Remarks

As the constructions of Section 8.7 show, our degree-sequence-based lower bounds can be asymptotically better than the previous average-degree-based bounds. This is in spite of using the previous bounds in the proof. The power of the random permutation method lies in that it allows us to obtain a relatively large *sparse* induced subgraph, over which the application of the average-degree bound yields a much better result than a straightforward application over the entire graph would have.

With regard to the tightness of our results and the weakening parameter A, firstly, from the proof of Theorems 8.1.2-8.1.4, it is clear that $\epsilon = \log A / \log D$ has to be at least a constant. Ideally, we may want to have $\epsilon = 0$ in the bounds of Theorems 8.1.2, 8.1.3 and 8.1.4. The following example, however, shows that it is possible to construct a triangle-free graph for which the bound in say, Theorem 8.1.2 would give a value more than the number of vertices: Take a disjoint union of $A = K_{n/3,n/3}$ and $B = \overline{K}_{n/3}$, and introduce a perfect matching between B and one of the parts of A. Now, $|V| = n$, $D \sim 2n/9$, and hence if $\epsilon = 0$, Theorem 8.1.2 would give a lower bound of $\Omega(n \log n)$, which is asymptotically larger than $|V|$. Similar examples can be constructed with linear hypergraphs also.

Chapter 9

Conclusion and Future Directions

9.1 Summary

In this thesis we studied some (di)-graph invariants over random (di)-graphs,and obtained results on their concentration, for a range of edge/arc probabilities.

For a random directed graph $D \in \mathcal{D}(n,p)$, we first considered the size of the largest induced acyclic tournament, $mat(D)$. We showed that for all $p = p(n)$, $mat(D)$ is two-point concentrated at $\{b^*, b^*+1\}$, where $b^* = \lceil 2\log_{p^{-1}} n + 0.5 \rceil$. Further, for every *fixed* p, $mat(D)$ is one-point concentrated for *most* n, i.e. for a dense subset of integers. Moreover, we showed some non-1-point concentration results for the case when $b^* = b^*(n)$ is always close to some integer. We also obtained threshold results for the digraph properties $mat(D) = k$, for each $k \in \mathbb{Z}^+$. Next, we analysed some heuristic algorithms for obtaining a large induced acyclic tournament in a given random digraph. We showed that a greedy heuristic always yields a maximal acyclic tournament, and hence, a.a.s. every maximal acyclic tournament is of size at least $\log_{p^{-1}} n$, and further that an improved algorithm based on Matula's "expose-and-merge" paradigm yields a slight improvement of $\log_{p^{-1}} n + c\sqrt{\log_{p^{-1}} n}$, where $c \in \Re^+$ is *any* real number.

We also considered the problem of the size $mas(D)$ of the largest induced acyclic subgraph in a random digraph $D \in \mathcal{D}(n,p)$. We improved upon the earlier bound of

$$\frac{2}{\ln q}\left(\ln np - \ln\ln np - O(1)\right),$$

(where $q = (1-p)^{-1}$), given by Subramanian [63] and Spencer and Subramanian [61], showing that for $p \geq C/n$, where C is a suitably large constant,

$$mas(D) \;\geq\; \frac{2}{\ln q}\left(\ln np - O(1)\right).$$

Besides, we gave a small additive improvement on the upper bound for $mas(D)$, using the idea of counting the number of acyclic orientations. As a result, the gap between the upper and the lower bounds on $mas(D)$ was improved from $O(p^{-1}\ln\ln np)$ to $O(p^{-1})$. Using Talagrand's inequality, we showed that the *actual* gap is $\sim \sqrt{p^{-1}\ln np}$. Further, as in the case of $mat(D)$, we also analyzed some heuristic algorithms and showed that a.a.s. an algorithm based on Matula's "expose-and-merge" paradigm yields an induced acyclic subgraph of $\log_q n + c\sqrt{\log_q n}$, where $c \in \Re^+$ is *any* real number.

Next, we studied the size of a largest induced path, cycle or tree in the random graph $\mathcal{G}(n,p)$, indicated by $mip(G)$, $h(G)$ and $T(G)$ respectively. Using the first and second moment methods, together with some innovative ways of computing the second moment, we were able to show that

$$T(G) \;\geq\; mip(G) \;\geq\; \lceil 2\log_q np + 2\rceil.$$

This improves the old lower bound given by Erdős and Palka [29] for $T(G)$. We also obtained 2-point concentration results for $mip(G)$, and $h(G)$ when $n^{-1/2}\ln^2 n \leq p \leq 1-\epsilon$, for any real $\epsilon > 0$.

Finally, we looked into the problem of finding a degree-sequence based lower bound on the independence number of general and linear k-uniform hypergraphs. Extending the Bopanna-Caro-Wei technique for independence numbers of graphs, we gave a probabilistic proof of a result of Caro and Tuza [21], showing that for any k-uniform hypergraph $H = (V,E)$

$$\alpha(H) \;\geq\; \sum_{v\in V(H)} \frac{1}{\binom{d(v)+1/t}{d(v)}}.$$

Further, we used this method, together with a special formulation of the Principle of Inclusion and Exclusion, to show that for any linear k-uniform hypergraph H, there exists $c > 0$ such that the following holds:

$$\alpha(H) \;\geq\; c(\log D)^{\frac{1}{k-1}} \sum_{v\in V(H)} \frac{1}{\max\{D^{\frac{\epsilon}{k-1}}, (d(v))^{\frac{1}{k-1}}\}}.$$

This answered some questions asked by Caro and Tuza in their 1991 paper [21].

Future Directions

In the final section of this thesis, we shall discuss some future directions that the research done in this thesis can lead to:

(i) In Chapters 5, 6, we've seen the size of the largest induced directed acyclic subgraph (DAG) of a random digraph $\mathcal{D}(n,p)$. What would such a largest DAG look like? For example, does it have a directed hamiltonian path(s)? Through an easy first moment calculation, it can be shown that the answer is "no". Still other questions remain interesting, e.g. the number of sources and sinks, the number of layered orderings and their general structure, i.e. the number of vertices in each layer.

(ii) Is it possible to obtain a lower bound of the order of $(2\log_q np)(1-o(1))$ on $mip(G)$, $G \in \mathcal{G}(n,p)$, when $p = O(n^{-1/2})$? This might require newer techniques for proving concentration of a random variable.

(iii) In the reverse direction of the previous question, is it possible to prove that $mas(D)$ and/or $mip(G)$ are *not* concentrated in an interval of constant length, for $p = o(n^{-1/2})$?

(iv) Recently, the Caro-Wei lower bound on graphs (see Chapter 8) was improved by Angel, Campigotto and Laforest [2] using the *Bhatia-Davis* inequality, essentially by looking at the variance of a certain random variable, instead of just the expected value as the Bopanna-Caro-Wei method does. Is it possible to have a similar improved lower bound for the independence number of hypergraphs also?

Index

Bibliography

[1] D. Achlioptas and A. Naor, "The two possible values of the chromatic number of a random graph", *Annals of Mathematics*, 162(3), (2005), 1333-1349.

[2] Eric Angel, Romain Campigotto and Christian Laforest: A New Lower Bound on the Independence Number of Graphs. *Discrete Applied Mathematics*, To appear.

[3] M. Ajtai, P. Erdős, J. Komlós and E. Szemerédi: On Turán's Theorem for Sparse Graphs. *Combinatorica* **1**(4) (1981), 313-317.

[4] M. Ajtai, J. Komlós, J. Pintz, J.H. Spencer and E. Szemerédi: Extremal uncrowded hypergraphs, *J. Combinatorial Theory Ser. A* **32**, 321-335 (1982).

[5] M. Ajtai, J. Komlós and E. Szemerédi: A Note on Ramsey Numbers. *J. Comb. Theory Ser. A* **29** (1980) 354-360.

[6] N. Alon, J.H. Kim, and J. Spencer, Nearly perfect matchings in regular simple hypergraphs, *Israel J. Math.* **100** (1997), 171-187.

[7] N. Alon and M. Krivelevich, "The concentration of the chromatic number of random graphs", *Combinatorica*, 17, 303-313.

[8] George E. Andrews, Richard Askey and Ranjan Roy: Special Functions. Encyclopedia of Mathematics and its Applications (1999) **71** Cambridge University Press.

[9] N. Alon and J.H. Spencer, *The Probabilistic Method*, Wiley International, 2001.

[10] W. N. Bailey: Generalized Hypergeometric Series (1935) Cambridge.

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

[12] Barabási, A-L., Albert, R. and Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web, *Physica* **A281**, 69-77

[13] Frits Beukers: Gauss' Hypergeometric Function (2002)
http://www.math.uu.nl/people/beukers/MRIcourse93.ps

[14] B. Bollobás, Random Graphs (2nd Edition) Camb. Univ. Press (2001).

[15] B. Bollobás, The chromatic number of random graphs. *Combinatorica* **8(1)**: 49-55 (1988).

[16] B. Bollobás, The structure of hereditary properties and Colourings of Random Graphs. *Combinatorica* **20(2)**: 173-202 (2000).

[17] B. Bollobás and P. Erdös, "Cliques in random graphs", *Math. Proc. Camb. Phil. Soc.* **80**, 419-427, 1988.

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

[19] R. Bopanna: [Comment on Lance Fortnow's blog, article by Bill Gasarch] http://blog.computationalcomplexity.org/2010/08/today-is-paul-turans-100th-birthday.html, comment #1 (2010).

[20] Y. Caro: New results on the independence number. Technical Report, Tel Aviv University, 1979.

[21] Yair Caro and Zsolt Tuza: Improved Lower Bounds on k- Independence, *J. Graph Theory* 1991, Vol. **15**, 99-107.

[22] R. Duke, H. Lefmann and V. Rödl: On uncrowded hypergraphs, *Random Structures and Algorithms*, **6**, 209-212, 1995.

[23] K. Dutta and C.R. Subramanian, "Largest induced acyclic tournament in random digraphs : A 2-point concentration", *Proceedings of LATIN-2010 (9th Latin American Theoretical Informatics Symposium)*, Oaxaca, Mexico, April 2010.

[24] K. Dutta and C.R. Subramanian, "Induced acyclic subgraphs in random digraphs : Improved bounds", *Proceedings of AofA-2010 (21st International Meeting on Probabilistic and Asymptotic Methods for the Analysis of Algorithms)*, Vienna, Austria, June 2010.

[25] Erdős, P. (1947). Some remarks on the theory of graphs, *Bull. Amer. Math. Soc.* **53**, 292-294.

[26] Erdős, P. (1959). Graph theory and Probability, *Canad. J. Math.* **11**, 34-38.

[27] Erdős, P. (1961). Graph theory and probability II, *Canad. J. Math.* **13**, 346-352.

[28] P. Erdős and L. Lovász: Problems and results on 3-chromatic hypergraphs and related questions, *Finite and Infinite Sets* A. Hajnal, R. Rado and V.T. Sós eds. North-Holland, Amsterdam, pp. 609-628.

[29] P. Erdős, Z. Palka, Trees in Random Graphs. *Discrete Mathematics* **46** (1983) pp. 145-150.

[30] P. Erdős, A. Rényi, On random graphs I. *Publicationes Mathematicae* **6** (1959) pp. 290-297.

[31] P. Erdős, A. Rényi, On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** (1960) pp. 17-61.

[32] Floyd, R.W., (1967). "Assigning meaning to programs" (1967) *Proc. Symp. Appl. Math.* 19 19-32.

[33] Frieze, A.M. (1990). On the independence number of random graphs, *Discrete Mathematics* **81**, 171-176.

[34] A. M. Frieze, B. Jackson, Large holes in sparse random graphs. *Combinatorica* **7** pp. (1987) 265-274.

[35] A. M. Frieze, B. Jackson, Large induced trees in sparse random graphs. *Journal of Combinatorial Theory, Series B* **42** pp. (1987) 181-195.

[36] M.R. Garey and D.S. Johnson, (1978). Computers and Intractability: A Guide To The Theory of NP-Completeness, W.H.Freeman, San Francisco.

[37] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik: Concrete Mathematics (Reading, Massachusetts: Addison-Wesley, 1994)

[38] J. Hastad, "Clique is hard to approximate within $n^{1-\varepsilon}$", *Acta Mathematica*, 1999, 182, 105-142.

[39] S. Janson, T. Łuczak and A. Ruciński, An exponential bound for the probability of nonexistence of a specified subgraph in a random graph. *Random Graphs '87 (Proceedings, Poznan 1987)*, John Wiley & Sons (1990), pp. 73-87.

[40] Janson, S., Łuczak, T. and Ruciński, A. (2000). Random Graphs, Wiley Interscience Series in Mathematics and Optimization.

[41] Kahale, N., and Schulman, L.J., (1996). "Bounds on the Chromatic Polynomial and on the number of Acyclic Orientations of a graph", *Combinatorica*, **16** (3) 383-397.

[42] S. Khot, "Improved Inapproximability Results for MaxClique, Chromatic Number and Approximate Graph Coloring", (2001) *Proc. 42nd IEEE Symp. Foundations of Computer Science (FOCS 2001)*, 600-609.

[43] J. Komlós, J. Pintz and E. Szemerédi: A lower bound for Heilbronn's problem, *J. London Math. Soc.* (**2**) 25 (1982), no. 1, 13-24.

[44] Krivelevich, M. and Sudakov, B. (1998). "Coloring random graphs", *Information Processing Letters*, 67, 71-74.

[45] M. Krivelevich, B. Sudakov, V. Vu and N. Wormald, On the probability of indpendent sets in random graphs, *Rand. Struct. Alg.* **22** (2003) 1-14.

[46] L. Kučera and V. Rödl, Large trees in random graphs, *Comment. Math. Univ. Carolina* **28**, 7-14, (1987).

[47] T. Łuczak, "A note on the sharp concentration of the chromatic number of random graphs", *Combinatorica*, 11, 1991, 295-297.

[48] T. Łuczak and Z. Palka, Maximal induced trees in sparse random graphs, *Discrete Math.* **72**, 257-265 (1988).

[49] Lund, C. and Yannakakis, M. (1993). "The Approximation of Maximum Subgraph Problems", *Proceedings of the 20th International Colloquium on Automata, Languages and Programming (ICALP'93)*, LNCS **700** 40-51.

[50] Manber, U. and Tompa M. (1984). "The Effect of Number of Hamiltonian Paths on the Complexity of a Vertex-Coloring Problem", SIAM J. Comp., 13, 109-115.

[51] "The Transitive Closure of a Random Digraph," Random Structures and Algorithms, Vol. 1, No. 1 (1990) pp. 73-93.

[52] Matula, D.W., (1987). "Expose-and-Merge exploration and the chromatic number of a random graph", *Combinatorica* 7(3), (1987), 275-284.

[53] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, 1995.

[54] C.H. Papadimitriou and M. Yannakakis (1991), "Optimization, Approximation, and Complexity Classes", *J. Computer and System Sciences* (special issue for 20th ACM Symposium on Theory of Computing) 43(3), 425-440.

[55] Rosen B. K., (1982). "Robust linear algorithms for cutsets", *J. Algorithms* 205-217 3.

[56] J.B. Shearer: A note on the independence number of triangle-free graphs. *Discrete Math.* **46** (1983) 83-87.

[57] J.B. Shearer: A note on the independence number of triangle-free graphs II. *J. Combintorial Theory Series B* **2** 300-307.

[58] J.B. Shearer: On the independence number of sparse graphs. *Random Structures Algorithms* **7** (1995) 269-271.

[59] Speckenmeyer, E. (1989). "On feedback problems in digraphs", *Proceedings of 15th International Workshop on Graph Theoretic Concepts in Computer Science (WG'89)*, Springer-Verlag, LNCS 411, 218-231.

[60] J. Spencer, Turán's theorem for k-graphs, Discrete Mathematics 2 (1972) 183–186.

[61] J.H. Spencer and C.R. Subramanian, (2008). "On the size of induced acyclic subgraphs in random digraphs", *Disc. Math. and Theoret. Comp. Sci.*, **10:2**, 47-54.

[62] R.P. Stanley: Enumerative Combinatorics, 1997, Cambridge University Press.

[63] C.R. Subramanian, (2003). "Finding induced acyclic subgraphs in random digraphs", *The Electronic Journal of Combinatorics*, **10**, #R46.

[64] W. C. Suen, On large induced trees and long induced paths in sparse random graphs, *Journal of Combinatorial Theory, Series B*, **56**(2) (1992) pp. 250-262.

[65] Tamás Lengyel: A combinatorial identity and the world series. *SIAM Review* Vol. 35, No. 2 (Jun. 1993), pp. 294-297.

[66] Torsten Thiele: A lower bound on the independence number of arbitrary hypergraphs. *J. Graph Theory* Vol. 30, **3**, March 1999.

[67] P. Turán: On an extremal problem in graph theory [in Hungarian]. *Math Fiz. Lapok* **48** (1941) 436-452.

[68] J. H. Van Lint, R. M. Wilson, A Course in Combinatorics, *Cambridge University Press*, (1991).

[69] W. Fernandez de la Vega, Induced trees in sparse random graphs. *Graphs and Combinatorics*, **2** 1 (1986), pp. 227-231.

[70] W. Fernandez de la Vega, The largest induced tree in a sparse random graph. *Random Struct. Alg.* **9** 1-2, (1996), pp. 93-97.

[71] V. Vu, Concentration of non-Lipschitz functions and applications, *Random Structures and Algorithms*, **20** (3) (2002), pp. 262-316.

[72] V.K. Wei, A lower bound on the stability number of a simple graph. Technical Memorandum TM 81-11217-9, Bell Laboratories, 1981.

[73] http://en.wikipedia.org/wiki/Hypergeometric_function