On the Strongly Connected Components of Random Directed Graphs with Given Degree Sequences

by

Alessandra Graf

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Combinatorics and Optimization

Waterloo, Ontario, Canada, 2016
© Alessandra Graf 2016

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

A strongly connected component of a directed graph G is a maximal subgraph H of G such that for each pair of vertices u and v in H, there is a directed path from u to v and a directed path from v to u in H. A strongly connected component is said to be giant if it has linear size.

We determine the threshold at which a random directed graph with a well-behaved degree sequence asymptotically almost surely contains a giant strongly connected component. This is a new proof of a result by Cooper and Frieze in [12]. In addition, we predict the site percolation threshold for the presence of a giant strongly connected component in a graph with a well-behaved degree sequence.

Acknowledgements

I would like to thank the my supervisor, Jane Gao, for all her support, guidance, and patience. Without her, this work would not have been possible. I would also like to thank Peter Nelson for his advice and assistance as I prepared this thesis as well as Penny Haxell for her valuable comments.

I would also like to thank my family and friends for all their support and encouragement.

Table of Contents

List of Figures

Chapter 1

Introduction

One of the most studied phenomena in the theory of random graphs is the behavior of the size of the largest component of a random graph. The first such result, due to Erdős and Rényi in 1960 [14], showed that the size of the largest component in $G(n, p)$ undergoes a drastic change with respect to p. In particular, when p is below a certain threshold, the size of the largest component is $O(\log(n))$. When p is above that threshold, the size of the largest component is $\theta(n)$.

Since then, the presence of a component of size $\theta(n)$, called a *giant component*, has been investigated in other random graphs (see [23, 35]). For example, Molloy and Reed studied the presence of a giant component in the space of graphs with a given degree sequence. They found a threshold for when a giant component asymptotically almost surely exists and when such a component asymptotically almost surely does not exist in such graphs. This led to further work by Kang and Seierstad [27], Pittel [38], Janson and Luczak [23], Riordan [39], and Hatami and Molloy [21]. Recently, Bollobás and Riordan improved Molloy and Reed's result to allow for exponential bounds on the probabilities of large deviations [8].

However, few results about the presence of giant strongly connected components in random directed graphs are known. This is an area of great interest because many large, real-world networks are better modeled by directed graphs. The internet, cellular networks, the food web, and various metabolic and social networks have all been shown to follow directed graphs [11]. For example, in 1999 researchers from IBM found that the in- and out-degrees of a data set of 40 million webpages followed a *power law distribution* [26]. In a power law distribution, the number of vertices of degree k is proportional to k^{-c} for some fixed $c > 1$. Hence the researchers showed that within this data set, there were many

websites with large in- and out-degrees.

As a result, analyzing the behavior of properties of random directed graphs, particularly those with power law distributions, will provide more insight into the behavior of real-world systems. For example, if vertices or edges were removed from the graph at random, would a giant strongly connected component still remain in the graph? One process that removes vertices and edges in such a manner is known as percolation. Percolation has been well studied in undirected graphs (see [18, 7]) but not in directed graphs. This includes percolation on hypercubes [1], d-regular graphs with girth tending to infinity [2], and graphs with a given degree sequence [15, 22].

In this thesis, we discuss the size of the largest strongly connected component in random directed graphs with degree sequences that satisfy certain properties, which we call *well-behaved degree sequences*. We present a new proof of a result of Cooper and Frieze from [12], including a slight relaxation to the properties required for a degree sequence to be well-behaved. In addition, we present how this result can be applied to percolation on random directed graphs with these degree sequences. We provide some arguments that predict a threshold for the presence of a giant strongly connected component after site percolation on such graphs.

The main result of this thesis concerns the presence of a giant strongly connected component in random directed graphs with well-behaved degree sequences. This is proved in Chapter 4 and Chapters 2 and 3 provide the background and methods required by the proof. The application to percolation theory is discussed in Chapter 6, with Chapter 5 providing some results and techniques from the theory. We conclude with a discussion of the implications of our results and some further areas of study.

Specifically, Chapter 2 reviews some known results about the size of the largest component or strongly connected component in a random graph or directed graph. A greater emphasis is placed on results for random graphs since we adapt some of the techniques used for these results to directed graphs. We also formally define a degree sequence and the properties that make it well-behaved.

In Chapter 3, we describe the model used for the results in this thesis. This model, often attributed to Bollobás [5], is known as the *configuration model* and is a method of representing directed graphs to study their properties. The chapter also includes some results of branching processes and martingales necessary for the proof of Theorem 4.1.

Chapter 4 consists of the proof of Theorem 4.1, which states the threshold for the presence of a giant strongly connected component in a random directed graph. This result is a new proof of the result of [12] with a slight improvement to the restrictions on the

degree sequences. The proof is divided into two cases, *subcritical* and *supercritical*, which are discussed in Sections 4.1 and 4.2 respectively.

In Chapter 5, we define the two types of discrete percolation models that are studied in Chapter 6. Some techniques used to study these models, including a method developed by Janson in [22], are discussed in Section 5.2.

Chapter 6 presents a heuristic approach that predicts a certain site percolation threshold for directed graphs with well-behaved degree sequences. Specifically, we discuss how Theorem 4.1 and a technique of Janson in [22] can be applied to predict the site percolation threshold for the presence of a giant strongly connected component in random directed graphs with well-behaved degree sequences.

The final chapter will discuss some implications of the results in Chapters 4 and 6 as well as areas of further study.

Chapter 2

Properties of Random Graphs

In 1959, Erdős and Rényi introduced one of the first models for generating random graphs [13]. Today, this model is known as a probability space called $G(n, N)$. This space consists of the set of all labelled graphs on n vertices with N edges chosen randomly and independently from the set of $\binom{n}{2}$ possible edges. Every such graph has an equal probability of being chosen from this space, and so the space has uniform distribution.

A similar probability space, known as $G(n, p)$, was introduced by Edgar Gilbert in 1959 [16]. $G(n, p)$ is defined to be a probability space over the set of all graphs on vertex set $\{1, \ldots, n\}$ which include each of the possible $\binom{n}{2}$ edges with probability p independent of all other edges. Since $G(n, N)$ generally has similar properties to $G(n, p)$ when $p \sim N/\binom{n}{2}$, $G(n, p)$ is the space more commonly studied.

In [14], Erdős and Rényi were the first to study the probable structure of random graphs in terms of more than just their connectivity [14]. They investigated when $G(n, N)$ almost always satisfied a graph-theoretic property $\mathcal{P}$. Erdős and Rényi observed that many natural graph-theoretic properties change their behavior in $G(n, N)$ only over a small range of N. Specifically, $G(n, N)$ changes from almost always satisfying property $\mathcal{P}$ to almost always not satisfying $\mathcal{P}$ near a specific choice of N. This led to the following definition shown here in terms of $G(n, p)$. Note that $f(n) \ll g(n)$ if $\frac{f(n)}{g(n)} \to 0$ as $n \to \infty$.

Definition 2.1. *$t(n)$ is called a* threshold function *for a graph theoretic property $\mathcal{P}$ if*

1. *When $p \ll t(n)$, the probability $G(n, p)$ satisfies property $\mathcal{P}$ as $n \to \infty$ is 0,*
2. *When $p \gg t(n)$, the probability $G(n, p)$ satisfies property $\mathcal{P}$ as $n \to \infty$ is 1,*

or vice versa.

Erdős and Rényi showed that the presence of certain subgraphs, such as trees and cycles of a given order, as well as containing a given number of components in $G(n, N)$ have threshold functions. However, the size of the largest component of a graph is one of the most studied properties originally investigated by Erdős and Rényi that has a threshold function. In the next section, we provide a brief history of some of these results and techniques relating to these threshold functions in different random graphs.

For the remaining sections and chapters, we will use the following notation and abbreviations.

Definition 2.2. *Let n be an integer variable which tends to infinity and let g be a positive function. For any function $f(n)$,*

1. *$f(n) = O(g(n))$ if there exists a $c > 0$ such that $|f(n)| \leq cg(n)$ for all n,*
2. *$f(n) = o(g(n))$ if $f(n)/g(n) \to 0$,*
3. *$f(n) = \Omega(g(n))$ if there exists a $c > 0$ such that $f(n) \geq cg(n)$ for all sufficiently large n, and*
4. *$f(n) = \Theta(n)$ if $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.*

Definition 2.3. *A property holds* asymptotically almost surely*, denoted a.a.s., if the property holds with probability tending to 1 as $n \to \infty$.*

2.1 The Size of the Largest Component

We begin with a formal definition of a component of a multigraph.

Definition 2.4. *A* component *K of a multigraph G is a maximal connected subgraph of G, i.e. K is a maximal subgraph such that every pair of vertices in K are joined by a path in K.*

The behavior of the size of the largest component in a random graph has been studied by many authors over the past 50 years (see [14, 36, 6, 23, 21, 39, 8]). In many different random graphs, it is known that below the threshold function, the size of the largest

component is $O(\log(n))$. Similarly, above the threshold, the size of the largest component is known to be $\Omega(n)$. However, these functions differ for different random graphs.

The first result on the behavior of the size of the largest component in $G(n, N)$ is due to Erdős and Rényi in [14]. Here, we restate their result.

Proposition 2.5 ([14]). *Let $N = cn+o(n)$ and C denote the largest component in $G(n, N)$.*

(i) If $c < \frac{1}{2}$, then a.a.s. $|V(C)| = O(\log n)$.

(ii) If $c > \frac{1}{2}$, then a.a.s. $|V(C)| = \Omega(n)$.

The values of c near $\frac{1}{2}$ are far more difficult to study, which is true of most threshold functions. As a consequence, the behavior of random graphs is usually studied when the parameters in the threshold functions differ sufficiently in the graphs from the threshold (such as $p \ll t(n)$ or $p \gg t(n)$).

However, values of c near the threshold have been well studied in $G(n, N)$ and $G(n, p)$. Erdős and Rényi were able to show a lower bound on the size of the largest component when $c = \frac{1}{2}$. The first upper bound was not shown until 1984 by Bollobás [5]. The proof that the largest component of $G(n, p)$ has size $\Theta(n^{2/3})$ when $N = \frac{1}{2}n + o(n)$ was completed six years later [28]. Since then, more precise results have been presented for $G(n, p)$ in [31] and [37].

In addition to $G(n, N)$ and $G(n, p)$, several authors have studied the emergence of a giant component in a random graph with a given degree sequence.

Definition 2.6. *A* degree sequence D_n *is a sequence of n non-negative integers whose sum is even.*

We say that a graph G has degree sequence $D_n = (d_1, d_2, \ldots, d_n)$ if G is isomorphic to a graph with vertices $\{1, 2, \ldots, n\}$ such that $\deg(i) = d_i$. Define $\mathcal{G}(D_n)$ to be the set of all labelled graphs with degree sequence D_n. Then a graph with degree sequence D_n is a uniformly random member of $\mathcal{G}(D_n)$. Figure 2.1 provides an example of a graph with degree sequence $(4, 4, 3, 3, 4)$.

Many results for random graphs with a given degree sequence study asymptotic behavior as n tends to infinity. As the value of n remains constant in a degree sequence, this led to a generalization of a degree sequence, called a *degree array*.

Definition 2.7. *A* degree array *is an array of integer-valued functions $\mathcal{D} = d_0(n), d_1(n), \ldots$ such that*

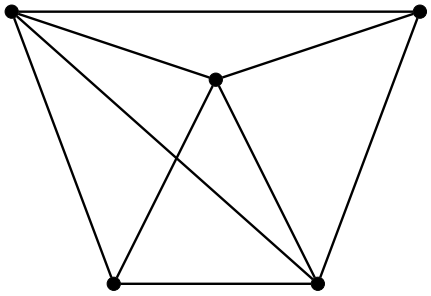


Figure 2.1: A graph with degree sequence $(4, 4, 3, 3, 4)$.

1. $d_i(n) = 0$ *for all* $i \geq n$ *and*
2. $\sum_{i \geq 0} d_i(n) = n$.

Unlike a degree sequence, $d_i(n)$ represents the number of vertices of degree i in a graph as a function of n. A degree sequence D_n can be obtained from $\mathcal{D}$ by defining $D_n = (d_1, d_2, \ldots, d_n)$ such that $|\{j \mid d_j = i\}| = d_i(n)$ for all $i \geq 0$. Degree arrays can therefore be used to study behaviors in graphs as the number of vertices increases.

Molloy and Reed were the first to study the presence of a giant component in random graphs with given degree sequences. They used degree arrays satisfying certain constraints to find the appropriate threshold function [35]. Instead of representing a probability or number of edges, this threshold function approximated the increase in the expected number of "unknown" neighbors when a vertex is exposed in the branching process. (The branching process is described in more detail in Section 3.1.) Molloy and Reed were later able to approximate the size of the giant component under these constraints [36].

The work of Molloy and Reed led to a series of papers studying the largest component of random graphs with specified degree sequences. Some of these focused on improving the error term in the result of [36] (see [27, 38, 23, 39, 21]). Others sought to relax some of the constraints on the degree sequences, such as the maximum degree (see [23, 8]).

Far less focus has been placed on random directed graphs and the size of their strongly connected components.

Definition 2.8. *A* strongly connected component *(or SCC)* K *of a directed multigraph* G *is a maximal subgraph such that every pair of vertices* (u, v) *in* K *are joined by both a directed* uv*-path and directed* vu*-path in* K*.*

Before discussing known results for the size of the largest SCC in a directed graph, note that the following properties hold for strongly connected components.

Lemma 2.9. *Let G be a directed multigraph and $u, v \in V(G)$.*

(i) If there is a directed uv-walk in a directed multigraph G, then there is a directed uv-path in G.

(ii) If there exists both a directed uv-path and directed vu-path in G, then u and v are in the same strongly connected component.

(iii) The strongly connected components of G partition $V(G)$.

Proof: (i) Let $W = w_0w_1w_2 \dots w_\ell$ be the directed uv-walk in G, i.e. $u = w_0$ and $v = w_\ell$. The proof is by induction on ℓ. If $\ell = 0$, then $u = v$ and the path of length 0 is a directed uv-path in G. For $\ell = 1$, two cases must be considered. If $u = v$, then the path of length 0 is a directed uv-path in G. If $u \neq v$, then W consists of the edge uv and so W is a directed uv-path of length 1.

Let $k \geq 1$ and assume that for all integers $0 \leq \ell \leq k$, if there exists a directed uv-walk of length ℓ in G, there exists a directed uv-path in G. Let $W = w_0w_1 \dots w_{k+1}$ be a directed uv-walk in G of length $k+1$. If W is a directed path, then clearly there exists a directed uv-path in G. Thus, suppose W is not a directed uv-path. Then there exists some $0 \leq i < j \leq k+1$ such that $w_i = w_j$. Then $W^* = w_0w_1 \dots w_iw_{j+1}w_{j+2} \dots w_{k+1}$ is a uv-walk in G of length $k^* \leq k$. Hence by the inductive hypothesis, there exists a directed uv-path in G.

(ii) Let $P = u_0u_1u_2 \dots u_\ell$ be a uv-path and $Q = v_0v_1v_2 \dots v_m$ be a vu-path in G (note $u = u_0 = v_m$ and $v = v_0 = u_\ell$). By (i), it suffices to show that for all $a, b \in V(P \cup Q)$, there exists a directed ab-walk and directed ba-walk in $P \cup Q$. For $a = b$, the path of length 0 is a directed ab-walk and directed ba-walk in $P \cup Q$. Thus, assume $a \neq b$.

Suppose $a = u_i$ and $b = v_j$ for some $1 \leq i \leq \ell$ and $1 \leq j \leq m$. Then $P_{i,j} = u_iu_{i+1} \dots u_\ell v_1v_2 \dots v_j$ is a directed u_iv_j-walk and $Q_{i,j} = v_jv_{j+1} \dots v_mu_1u_2 \dots u_i$ is a directed v_ju_i-walk in $P \cup Q$. Now suppose $a = u_i$ and $b = u_j$ for some $0 \leq i < j \leq \ell$. Clearly $u_iu_{i+1} \dots u_j$ is a directed u_iu_j-path in $P \cup Q$ and $P_{j,m}Q_{i,m}$ is a directed u_ju_i-walk in $P \cup Q$. Similarly, for $a = v_i$ and $b = v_j$, $v_iv_{i+1} \dots v_j$ is a directed v_iv_j-path and $Q_{\ell,j}Q_{\ell,i}$ is a directed u_ju_i-walk in $P \cup Q$.

Thus for every pair of vertices (a, b) in $P \cup Q$, there is a directed ab-path and directed ba-path in $P \cup Q$. By the maximality of an SCC, this implies $P \cup Q$ is a subgraph of a SCC of G. Hence u and v are in the same SCC.

(iii) Let K_1 and K_2 be two strongly connected components of G. For all $u, v \in V(K_1)$, let P_{uv} be a directed uv-path in K_1 and for all $u, v \in V(K_2)$, let Q_{uv} be a directed uv-path in K_2.

Suppose $V(K_1) \cap V(K_2) \neq \emptyset$ and let $v \in V(K_1) \cap V(K_2)$. Let $u \in V(K_1)$ and $w \in V(K_2)$ and consider the paths $P_{uv} = u_0 u_1 u_2 \dots u_\ell$ and $Q_{wv} = v_0 v_1 \dots v_m$. As $v = u_\ell = v_m \in V(K_1) \cap V(K_2)$, $P_{uv}Q_{vw}$ is a directed uw-walk and $Q_{wv}P_{vu}$ is a directed wu-walk in $K_1 \cup K_2$. Hence by (i), there exists a directed uw-path and directed wu-path in $K_1 \cup K_2$.

By (ii), this implies u and w are in the same SCC K^*. The maximality of K_1 and K_2 then implies $K^* = K_1$ and $K^* = K_2$, which is a contradiction. Hence $V(K_1) \cap V(K_2) = \emptyset$ and so the strongly connected components of G partition $V(G)$. □

The earliest results by Karp [25] and Łuczak [29] determined the threshold function for a giant SCC in $D(n,p)$. Here $D(n,p)$ is a probability space over the set of all graphs on vertex set $\{1, \dots, n\}$ which includes each of the possible $n(n-1)$ arcs in the directed graph with probability p independent of all other arcs. They independently proved the following proposition.

Proposition 2.10 ([25, 29])**.** *Let $\omega(n) \to \infty$ be a function and C denote the largest strongly connected component in $D(n,p)$.*

(i) If $np \to c < 1$, then a.a.s. $|V(C)| \leq \omega(n)$.

(ii) If $np \to c > 1$, then a.a.s. $|V(C)| \geq \alpha(c)n$, where $\alpha(c)$ is an explicitly defined constant.

Some improvements to the estimates of the size of the largest SCC were made in 2009 by Łuczak and Seierstad [32]. However, the few remaining results for the size of the strongly connected components in a random directed graph involve other probability spaces. For example, Łuczak and Cohen studied the threshold function for a giant SCC in a three-parameter random directed graph model [30]. The presence of a giant SCC in random directed graphs whose arcs are included with different probabilities was studied in 2012 by Bloznelis et al. [4].

One of the results that motivates this work was presented by Cooper and Frieze in [12]. In this paper, Cooper and Frieze studied the size of the largest SCC in random directed graphs with a given degree sequence. This requires a notion of degree sequences for directed graphs, which we define as follows.

Definition 2.11. *A* degree sequence D_n *of a directed graph is an ordered pair* (D_n^-, D_n^+) *of sequences* D_n^- *and* D_n^+ *such that* D_n^- *contains* n *non-negative integers,* D_n^+ *contains* n *non-negative integers, and the sum of the terms of* D_n^- *is also the sum of the terms of* D_n^+.

As before, a directed graph G has degree sequence $D_n = ((d_1^-, d_2^-, \ldots, d_n^-), (d_1^+, d_2^+, \ldots, d_n^+))$ if G is isomorphic to a graph with vertices $\{1, 2, \ldots, n\}$ such that the in-degree $\deg^-(i) = d_i^-$ and the out-degree $\deg^+(i) = d_i^+$. As $\mathcal{G}(D_n)$ is the set of all labelled directed graphs with degree sequence D_n, a directed graph with degree sequence D_n is a uniformly random member of $\mathcal{G}(D_n)$. Given a degree sequence D_n, we use $n_{i,j}$ to denote the number of vertices of in-degree i and out-degree j in a graph with degree sequence D_n.

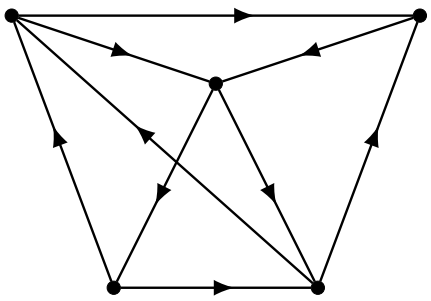


Figure 2.2: A directed graph with degree sequence $((2, 2, 2, 1, 2), (2, 2, 1, 2, 2))$.

Similar to the work of Molloy and Reed in [35], the threshold function for the presence of a giant SCC determined by Cooper and Frieze is related to the expected in- and out-degree of a vertex in the graph. This led Cooper and Frieze to study random directed graphs with *proper* degree sequences.

Definition 2.12. *A degree sequence* D_n *is* proper *if:*

1. *There is an absolute constant* A_1 *such that* $\theta = \frac{1}{n} \sum_{i \geq 0} \sum_{j \geq 0} i n_{i,j} = (1 + o(1))A_1$,
2. *Every term of the sequence is at most* Δ *and* $\Delta \leq n^{1/12}/\log(n)$,
3. $\max\left(\sum_{i \geq 0} \sum_{j \geq 0} i^2 \frac{n_{i,j}}{n}, \sum_{i \geq 0} \sum_{j \geq 0} j^2 \frac{n_{i,j}}{n} \right) \leq A_3$ *for some absolute constant* A_3,
4. $\sum_{i \geq 0} \sum_{j \geq 0} ij \frac{n_{i,j}}{\theta n} = (1 + o(1))A_2$ *where* $A_2 \leq 1 - \epsilon$ *or* $A_2 \geq 1 + \epsilon$ *for absolute constants* A_2 *and* ϵ, *and*

5. *For* $\rho = \max\left(\sum_{i\geq 0}\sum_{j\geq 0} i^2 j \frac{n_{i,j}}{\theta n}, \sum_{i\geq 0}\sum_{j\geq 0} ij^2 \frac{n_{i,j}}{\theta n}\right)$, *if* $\rho \to \infty$ *with* n *then* $\rho = o(\Delta)$.

The main result of [12] is restated below.

Theorem 2.13 ([12])**.** *Let* D_n *be a proper degree sequence and define* $\lambda = \sum_{i\geq 0}\sum_{j\geq 0} ij\frac{n_{i,j}}{\theta n}$.

(i) If $\lambda < 1$, *then a.a.s. every strongly connected component of* $G \in \mathcal{G}(D_n)$ *has size* $O(\Delta^2 \log(n))$.

(ii) If $\lambda > 1$, *then a.a.s. there exists in* $G \in \mathcal{G}(D_n)$ *a unique giant strongly connected component.*

In Chapter 4, we provide a new proof of Theorem 2.13 using well-behaved degree sequences, which we formally define in the next section. Well-behaved degree sequences are similar to proper degree sequences, but with slightly different constraints, such as the permitted maximum degree in the sequence. Thus, we prove a threshold function for the presence of a giant SCC in a random directed graph with a well-behaved degree sequence.

2.2 Well-behaved Degree Sequences

As discussed in Section 2.1, we require the degree sequence of a random directed graph to satisfy a few key properties in order to determine if the random graph asymptotically almost surely has a giant SCC. These properties are different from those of a proper degree sequence and some limitations still remain. We will discuss some of these limitations at the end of this section.

First, we define a degree array for directed graphs.

Definition 2.14. *A* degree array *is an array of integer-valued functions* $\mathcal{D} = \{n_{i,j}(n) \mid i, j \geq 0\}$ *such that*

1. $n_{i,j}(n) = 0$ *for all* $i \geq n$ *and* $j \geq n$, *and*
2. $\sum_{i\geq 0}\sum_{j\geq 0} n_{i,j}(n) = n$.

As with the definition for undirected graphs, $n_{i,j}(n)$ represents the number of vertices of in-degree i and out-degree j in a directed graph as a function of n. A degree sequence D_n can be obtained from $\mathcal{D}$ by defining $D_n = ((d_1^-, d_2^-, \ldots, d_n^-), (d_1^+, d_2^+, \ldots, d_n^+))$ such that $|\{m \mid d_m^- = i, d_m^+ = j\}| = n_{i,j}(n)$ for all $i, j \geq 0$. We will use degree arrays to study the presence of a giant SCC as the number of vertices increases.

We now define two important properties of degree arrays we require for a degree array to be well-behaved. Recall $\mathcal{G}(D_n)$ is the set of all labelled directed graphs with degree sequence D_n.

Definition 2.15. *A degree array D is* feasible *if for all $n \geq 1$, $\mathcal{G}(D_n) \neq \emptyset$.*

Definition 2.16. *A degree array is* smooth *if for all $i, j \geq 0$, there exist constants $\kappa_{i,j}$ such that $\lim_{n\to\infty} \frac{n_{i,j}(n)}{n} = \kappa_{i,j}$.*

We now define a well-behaved degree array. Note that we consider all degree sequences obtained from a well-behaved degree array to be well-behaved.

Definition 2.17. *A feasible degree array $\mathcal{D}$ is* well-behaved *if it is smooth and:*

1. *There is an absolute constant A_1 such that $\theta = \frac{1}{n}\sum_{i\geq 0}\sum_{j\geq 0} i n_{i,j}(n) = (1 + o(1))A_1$,*
2. *Every term of the degree sequence D_n is at most Δ and $\Delta = o(n^{1/4})$,*
3. $\max\left(\sum_{i\geq 0}\sum_{j\geq 0} i^2 \frac{n_{i,j}}{n}, \sum_{i\geq 0}\sum_{j\geq 0} j^2 \frac{n_{i,j}}{n}\right) \leq A_2$ *for some absolute constant A_2, and*
4. $\lambda = \lim_{n\to\infty}\sum_{i\geq 0}\sum_{j\geq 0} ij\frac{n_{i,j}(n)}{\theta n}$ *exists, is finite, and the sum approaches this limit uniformly, i.e. for all $\epsilon > 0$, there exists k and N such that for all $n > N$,*

$$\left|\sum_{i=0}^{k}\sum_{j=0}^{k} ij\frac{n_{i,j}(n)}{\theta n} - \lambda\right| < \epsilon.$$

A simple example of a well-behaved degree array is $\mathcal{D} = \{n_{d,d}(n)\}$ for some constant $d > 0$. This $\mathcal{D}$ is the degree array of a d-regular directed graph where each vertex has both in-degree d and out-degree d. It is feasible and smooth ($\kappa_{d,d} = 1$ and $\kappa_{i,j} = 0$ for all $i, j \neq d$). Furthermore, $\theta = d = A_1$, $d = o(n^{1/4})$ for sufficiently large n, and $\sum_{i\geq 0}\sum_{j\geq 0} i^2\frac{n_{i,j}}{n} = d^2 = A_2$.

Finally, $\lim_{n\to\infty} \sum_{i\geq 0}\sum_{j\geq 0} ij\frac{n_{i,j}(n)}{\theta n} = \frac{d^2}{d} = d$ is a finite limit such that for all $\epsilon > 0$ and all $n > N$ where N is an integer such that $d = o(N^{1/4})$, $\left|\sum_{i=0}^{d}\sum_{j=0}^{d} ij\frac{n_{i,j}(n)}{\theta n} - d\right| = |\frac{d^2}{d} - d| < \epsilon$.

Note that the first and third conditions of Definition 2.17 are equivalent to the first and third conditions of Definition 2.12. Also, the second condition of Definition 2.17 is less restrictive than the second condition of Definition 2.12. However, some stronger conditions than that of Definition 2.12 are assumed in a well-behaved degree array (see the fourth condition). Hence the class of proper degree sequences is not necessarily contained in the class of well-behaved degree sequences.

Furthermore, well-behaved degree sequences with $\Delta > \frac{n^{1/12}}{\log(n)}$ are trivially not proper degree sequences. However, the following lemma shows that some well-behaved degree sequences are also proper degree sequences.

Lemma 2.18. *Let D_n be a well-behaved degree sequence such that every term in D_n is at most $\Delta \leq n^{1/12}/\log(n)$ and $\lambda \neq 1$. Then D_n is also a proper degree sequence.*

Proof: Since D_n is well-behaved, D_n satisfies the first and third conditions of Definition 2.12. Furthermore, by assumption D_n satisfies the second condition of Definition 2.12. It therefore remains to verify the fourth and fifth conditions of Definition 2.12.

It is clear that for $A_2 = \lambda$, the fourth condition of Definition 2.12 holds by uniform convergence and the assumption on λ. Furthermore, the uniform convergence implies that for all $\epsilon > 0$, there exists a k and N such that for all $n > N$,

$$\sum_{i>k}\sum_{j\geq 0} ij\frac{n_{i,j}(n)}{\theta n} + \sum_{i=0}^{k}\sum_{j>k} ij\frac{n_{i,j}(n)}{\theta n} < \epsilon.$$

Hence for $\rho = \sum_{i\geq 0}\sum_{j\geq 0} i^2 j\frac{n_{i,j}(n)}{\theta n}$,

$$\begin{aligned}
\rho &= \sum_{i=0}^{k}\sum_{j=0}^{k} i^2 j\frac{n_{i,j}(n)}{\theta n} + \sum_{i>k}\sum_{j\geq 0} ij\frac{n_{i,j}(n)}{\theta n} + \sum_{i=0}^{k}\sum_{j>k} i^2 j\frac{n_{i,j}(n)}{\theta n} \\
&\leq \left(\sum_{i=0}^{k}\sum_{j=0}^{k} i^2 j\frac{n_{i,j}(n)}{\theta n}\right) + \Delta\left(\sum_{i>k}\sum_{j\geq 0} ij\frac{n_{i,j}(n)}{\theta n} + \sum_{i=0}^{k}\sum_{j>k} ij\frac{n_{i,j}(n)}{\theta n}\right) \\
&< \left(\frac{k^3}{\theta}\right) + \epsilon\Delta.
\end{aligned}$$

Note that assuming $\rho = \sum_{i \geq 0} \sum_{j \geq 0} ij^2 \frac{n_{i,j}(n)}{\theta n}$ leads to the same upper bound.

For arbitrarily small ϵ, $\rho \to \infty$ as $n \to \infty$ implies $\Delta \to \infty$ as $n \to \infty$. Thus $\rho = o(\Delta)$ if $\rho \to \infty$. □

In the next chapter, we introduce the models we use to study random directed graphs with well-behaved degree sequences as well as some important properties of those models.

Chapter 3

Modeling Directed Graphs with Given Degree Sequences

This chapter introduces the probabilistic tools and models used to prove the results in Chapter 4. We begin with a discussion of the probabilistic tools used in Sections 4.1 and 4.2, such as Galton-Watson branching processes and the Azuma-Hoeffding inequality. We then explain the configuration model and how it can be used to study directed graphs. All of the results presented in this chapter are known in the literature and we will provide references for these results as they appear.

3.1 The Branching Process

The Galton-Watson branching process was developed by Henry Watson in response to a question posed by Francis Galton about the extinction of family surnames [41]. The premise of this process is as follows.

Suppose that we wish to model a population of individuals which changes at discrete time intervals in the following manner. First, the process begins with a single individual in time 0. Then, at every integral time $t > 0$, a chosen individual created before time t produces a random non-negative number of individuals, called *offspring*. The number of offspring produced is independent of the number of offspring produced by any other individual, including those created at earlier times. Also, after producing its offspring, the chosen individual dies and so cannot produce more offspring. The process terminates when every individual is dead.

Definition 3.1. *The probability that this process terminates is the* extinction probability*.*

The process described above is a version of a Galton-Watson branching process. It creates a probability space in which each element of the space is an infinite process determined by the number of offspring each individual creates. A more formal definition of a Galton-Watson branching process is as follows.

Definition 3.2. *Let Z be a probability distribution over the non-negative integers. The* Galton-Watson branching process *begins with a single individual at time $t = 0$. This individual dies after creating Z offspring at time $t = 1$ and these offspring are ordered in some way. In order, each of these offspring die after independently creating Z offspring and these offspring are also ordered in some way. This procedure continues by having each individual in order produce an independent number Z of offspring.*

This branching process has a simple recursive structure that makes it easy to determine the number of individuals capable of producing offspring at some time t. To do so, let Z_t, $t = 1, 2, \ldots$, be a countable sequence of independent identically distributed variables with distribution Z. Label the first individual of the process 1 and its offspring $2, \ldots, Z_1 + 1$ in some way. Then, label individual 2's offspring $Z_1 + 2, \ldots, Z_1 + Z_2 + 2$ and individual 3's offspring $Z_1 + Z_2 + 3, \ldots, Z_1 + Z_2 + Z_3 + 3$ and so on. This procedure assigns each individual a distinct positive integer such that Z_t is the number of offspring of the t^{th} individual. Furthermore, since the Z_t are independent and have distribution Z, this procedure corresponds to a Galton-Watson branching process.

Suppose that at time t, individual t dies after creating its Z_t offspring. Let Y_t be the number of living (i.e. created but not dead) individuals at time t after individual t's death. Hence we have $Y_0 = 1$ and

$$Y_t = Y_{t-1} + Z_t - 1.$$

This leads to two possibilities. First, $Y_t > 0$ for all $t \geq 0$, in which case the process is infinite. Otherwise, $Y_t = 0$ for some $t \geq 0$ and so the total number of individuals in the process is T where $T = \min(\{t \mid Y_t = 0\})$. As this event is a termination of the branching process, we refer to it as *extinction* and can calculate the extinction probability using $\mathbf{E}(Z)$. The following is a well-known result for the extinction probability that can be found in [20].

Proposition 3.3. *Let ρ be the probability of extinction in a Galton-Watson branching process defined by the distribution of a random variable Z.*

(i) If $\mathbf{E}(Z) < 1$ then $\rho = 1$.

(ii) If $\mathbf{E}(Z) > 1$ *then* $0 < \rho < 1$.

We will present a proof of Proposition 3.3.i. and 3.3.ii. separately. First, consider the expectation of Y_t throughout the process.

Lemma 3.4. $\mathbf{E}(Y_t) = (\mathbf{E}(Z))^t$ *for all* $t \geq 1$.

Proof: The proof is by induction. For the base case, consider $\mathbf{E}(Y_1)$.

$$Y_1 = Y_0 + Z_1 - 1 = 1 + Z_1 - 1 = Z_1$$

and so $\mathbf{E}(Y_1) = \mathbf{E}(Z)$.

Suppose $\mathbf{E}(Y_t) = (\mathbf{E}(Z))^t$ for some $t \geq 1$. We wish to show that $\mathbf{E}(Y_{t+1}) = (\mathbf{E}(Z))^{t+1}$. Recall that $Z_1, Z_2, \ldots$ are independent copies of Z. Hence,

$$\begin{aligned}
\mathbf{E}(Y_{t+1}) &= \sum_i \mathbf{Pr}(Y_t = i)\mathbf{E}(Z_1 + Z_2 + \ldots + Z_i) \\
&= \sum_i \mathbf{Pr}(Y_t = i) i \mathbf{E}(Z) \\
&= \mathbf{E}(Z) \sum_i i \mathbf{Pr}(Y_t = i) \\
&= \mathbf{E}(Z)\mathbf{E}(Y_t) \\
&= \mathbf{E}(Z)(\mathbf{E}(Z))^t.
\end{aligned}$$

Thus $\mathbf{E}(Y_t) = (\mathbf{E}(Z))^t$ for all $t \geq 1$. □

We will also need Markov's inequality from probability theory. A different proof of this inequality can be found in [3].

Lemma 3.5 (Markov's Inequality)**.** *For any positive random variable* X *and* $\alpha > 0$, $\mathbf{Pr}(X \geq \alpha) \leq \frac{\mathbf{E}(X)}{\alpha}$.

Proof: Let $Y = \alpha \mathbb{1}_{X \geq \alpha}$ where $\mathbb{1}_{X \geq \alpha}$ is an indicator variable for the event $X \geq \alpha$. It is clear that $Y \leq X$ and so $\mathbf{E}(Y) \leq \mathbf{E}(X)$.

$$\begin{aligned}
\mathbf{E}(X) &\geq \mathbf{E}(Y) \\
&\geq \alpha \mathbf{Pr}(X \geq \alpha).
\end{aligned}$$

Hence $\mathbf{Pr}(X \geq \alpha) \leq \frac{\mathbf{E}(X)}{\alpha}$. □

We may now present a proof of Proposition 3.3.i.

Proof of 3.3.i.: By Markov's inequality, $\mathbf{Pr}(Y_t \geq 1) \leq \mathbf{E}(Y_t)$ for any $t \geq 1$. As Y_t takes on only non-negative integer values, $\mathbf{Pr}(Y_t > 0) \leq (\mathbf{E}(Z))^t$ for all $t \geq 1$ by Lemma 3.4. Thus, $\mathbf{E}(Z) < 1$ implies $\mathbf{Pr}(Y_t > 0)$ tends to 0 as $t \to \infty$. Hence $\mathbf{Pr}(Y_t = 0)$ tends to 1 as $t \to \infty$ and the statement holds. □

The proof of Proposition 3.3.ii. requires a different approach. Consider a single individual and suppose the subprocess that consists of the individual's descendants undergoes extinction, i.e. the individual has a finite number of descendants. We say such an individual "fails." Note that an individual fails if and only if all of its offspring fail and these latter events are independent.

Due to the independence of the creation of individuals, the probability an individual fails is ρ. Thus, by the law of total probability,

$$\rho = \sum_{i \geq 0} \mathbf{Pr}(Z = i)\rho^i.$$

With this fact, we present the following proof of Proposition 3.3.ii.

Proof of 3.3.ii.: Suppose $\mathbf{Pr}(Z = 0) = 0$. Then every individual has a positive number of offspring and so the branching process never terminates. Hence we may assume $\mathbf{Pr}(Z = 0) > 0$ and so $\rho > 0$.

Let f be the probability generating function for Z, so $f(x) = \sum_{i \geq 0} \mathbf{Pr}(Z = i)x^i$. Clearly $f(1) = 1$ and $f(0) > 0$. Also,

$$f'(x) = \sum_{i \geq 1} i\mathbf{Pr}(Z = i)x^{i-1}$$

and

$$f''(x) = \sum_{i \geq 2} i(i-1)\mathbf{Pr}(Z = i)x^{i-2}.$$

Thus f is increasing and convex on $[0, 1]$ as well as $\lim_{x \to 1^-} f'(x) = \mathbf{E}(Z)$.

As $\mathbf{E}(Z) > 1$, $f'(1) > 1$ and so for small $\epsilon > 0$, $f(1 - \epsilon) < 1 - \epsilon$. As $f(0) - 0 > 0$ and $f(1 - \epsilon) - (1 - \epsilon) < 0$, by the Intermediate Value Theorem there exists $s \in (0, 1 - \epsilon)$ such that $f(s) - s = 0$.

Note that $\mathbf{E}(Z) > 1$ and $\mathbf{Pr}(Z = 0) > 0$ imply $\mathbf{Pr}(Z = i) > 0$ for some $i > 1$. Thus $f''(x) > 0$ for $x > 0$ and so f is strictly convex. Hence for the smallest fixed point s of f, $f(x) < x$ for all $x \in (s, 1)$. Therefore s is the unique fixed point in $[0, 1)$ of f. As $\rho = \sum_{i \geq 0} \mathbf{Pr}(Z = i)\rho^i$, ρ is a fixed point of f. Thus $s = \rho$ and so the statement holds. □

Proposition 3.3.ii. will be used in Section 4.2.3 to prove an important lemma. This concludes our discussion of the Galton-Watson branching process.

3.2 Martingales and the Azuma-Hoeffding Inequality

A key theorem used to prove several lemmas in Sections 4.1 and 4.2 is the Azuma-Hoeffding inequality. This theorem is an important concentration result for a certain type of sequence of random variables known as a *martingale* and is well-known in the literature.

Definition 3.6. *A* martingale *is a sequence $X_0, \ldots, X_n$ of random variables so that for $0 \leq i < n$, $\mathbf{E}(X_{i+1} \mid X_0, X_1, \ldots, X_i) = X_i$.*

Note that for martingales, the conditional expected value of the next value in the sequence given all past values is equal to the present value in the sequence. Hence martingales can be used to model fair games where knowledge of past events does not help predict the expected value of the future winnings. Such games include a gambler's fortune in betting games that are fair as well as unbiased random walks in any number of dimensions.

However, it is also possible to model betting games that are biased in some manner. This can be accomplished using the following two generalizations of a martingale.

Definition 3.7. *Let $X_0, \ldots, X_n$ be a sequence of random variables.*

1. *The sequence is a* sub-martingale *if for every $0 \leq i < n$,* $\mathbf{E}(X_{i+1} \mid X_0, X_1, \ldots, X_i) \geq X_i$.
2. *The sequence is a* super-martingale *if for every $0 \leq i < n$,* $\mathbf{E}(X_{i+1} \mid X_0, X_1, \ldots, X_i) \leq X_i$.

Thus, if a sub-martingale and a martingale have equivalent expectations for a given time, the history of the sub-martingale tends to be bounded above by the history of the martingale. Similarly, a super-martingale tends to have its history be bounded below

by the history of a martingale whose expectations are equivalent (for a given time) to the expectations of the super-martingale. Hence sub-martingales model betting games with positive expected winnings and super-martingales model betting games with negative expected winnings.

We may now discuss the Azuma-Hoeffding inequality. A different proof of this result can be found in [3]. First, note the following lemma.

Lemma 3.8. *Let Y be a random variable such that $Y \in [-1, 1]$ and $\mathbf{E}(Y) = 0$. Then for any $t \geq 0$, $\mathbf{E}\left(e^{tY}\right) \leq e^{t^2/2}$.*

Proof: For any $x \in [-1, 1]$, $e^{tx} \leq \frac{1}{2}(1 + x)e^t + \frac{1}{2}(1 - x)e^{-t}$ by convexity. Taking the expectations,

$$\begin{aligned}
\mathbf{E}\left(e^{tY}\right) &\leq \frac{1}{2}e^t + \frac{1}{2}e^{-t} \\
&= \frac{1}{2}\left[\left(1 + t + \frac{t^2}{2} + \frac{t^3}{6} + \dots\right) + \left(1 - t + \frac{t^2}{2} - \frac{t^3}{6} + \dots\right)\right] \\
&= 1 + \frac{t^2}{2} + \frac{t^4}{4!} + \dots \\
&= \sum_{n \geq 0} \frac{t^{2n}}{(2n)!} \\
&\leq \sum_{n \geq 0} \frac{(t^2/2)^n}{n!} \\
&= e^{t^2/2}.
\end{aligned}$$

□

Theorem 3.9 (Azuma-Hoeffding Inequality)**.** *Let $X_0, X_1, \dots, X_n$ be a martingale. If there exist $c_i > 0$ such that $|X_i - X_{i-1}| \leq c_i$ for all $1 \leq i \leq n$, then for all positive reals λ,*

$$\mathbf{Pr}(X_n - X_0 \geq \lambda) \leq exp\left(\frac{-\lambda^2}{2\sum_{i=1}^{n} c_i^2}\right).$$

Proof: Let $Y_i = X_i - X_{i-1}$ for all $i \geq 1$. Note that for any $t > 0$,

$\mathbf{Pr}(X_n - X_0 \geq \lambda) = \mathbf{Pr}\left(e^{t(X_n - X_0)} \geq e^{t\lambda}\right)$. By Markov's inequality,

$$\begin{aligned}\mathbf{Pr}\left(e^{t(X_n-X_0)} \geq e^{t\lambda}\right) &\leq e^{-t\lambda}\mathbf{E}\left(e^{t(X_n-X_0)}\right) \\ &= e^{-t\lambda}\mathbf{E}\left(e^{t(Y_n+X_{n-1}-X_0)}\right) \\ &= e^{-t\lambda}\mathbf{E}\left(\mathbf{E}\left(e^{t(Y_n+X_{n-1}-X_0)} \mid X_0, X_1, \ldots, X_{n-1}\right)\right).\end{aligned}$$

Given $X_0, X_1, \ldots, X_{n-1}$, $e^{t(X_{n-1}-X_0)}$ is constant. Furthermore, $\frac{Y_n}{c_n}$ is a random variable with mean 0 and takes only values in $[-1, 1]$. Hence by Lemma 3.8,

$$\begin{aligned}\mathbf{E}\left(e^{t(Y_n+X_{n-1}-X_0)} \mid X_0, X_1, \ldots, X_{n-1}\right) &= e^{t(X_{n-1}-X_0)}\mathbf{E}\left(e^{tY_n} \mid X_0, X_1, \ldots, X_{n-1}\right) \\ &\leq e^{t(X_{n-1}-X_0)}e^{t^2c_n^2/2}.\end{aligned}$$

Thus,

$$\mathbf{Pr}(X_n - X_0 \geq \lambda) \leq e^{-t\lambda}e^{t^2c_n^2/2}\mathbf{E}\left(e^{t(X_{n-1}-X_0)}\right).$$

By handling $\mathbf{E}\left(e^{t(X_{n-1}-X_0)}\right)$ inductively in the same fashion as above,

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

As this holds for any $t > 0$, for $t = \frac{\lambda}{\sum_{i=1}^{n} c_i^2}$,

$$\mathbf{Pr}(X_n - X_0 \geq \lambda) \leq \exp\left(\frac{-\lambda^2}{2\sum_{i=1}^{n} c_i^2}\right).$$

□

Two important corollaries of the Azuma-Hoeffding inequality are the following.

Corollary 3.10. *If $X_0, X_1, \ldots, X_n$ is a martingale and there exist $c_i > 0$ such that $|X_i - X_{i-1}| \leq c_i$ for all $1 \leq i \leq n$, then for all positive reals λ,*

$$\mathbf{Pr}(X_n - X_0 \leq -\lambda) \leq exp\left(\frac{-\lambda^2}{2\sum_{i=1}^{n} c_i^2}\right).$$

Corollary 3.11. *Let X be a random variable determined by n trials $Z_1, Z_2, \ldots, Z_n$ and satisfying for each i*

$$\max |\mathbf{E}(X \mid Z_1, \ldots, Z_{i+1}) - \mathbf{E}(X \mid Z_1, \ldots Z_i)| \leq c_i,$$

where this maximum is taken over all possible outcomes of $Z_1, \ldots, Z_{i+1}$. Then

$$\mathbf{Pr}(|X - \mathbf{E}(X)| > \lambda) \leq 2exp\left(\frac{-\lambda^2}{2\sum_{i=1}^{n} c_i^2}\right).$$

Note that Theorem 3.9 can also be applied to super-martingales satisfying the condition that there exist $c_i > 0$ such that $|X_i - X_{i-1}| \leq c_i$ for all $1 \leq i \leq n$. Similarly, Corollary 3.10 can also be applied to sub-martingales that satisfy this condition.

In the remaining chapters, when using the result of Corollary 3.11, we will say "by the Azuma-Hoeffding inequality." Note that to apply Corollary 3.11, it suffices to show that a sequence of random variables is a martingale and that changing the present value of the sequence affects the next value by at most some constant $c > 0$.

The next section begins the discussion of the models used for our results.

3.3 The Configuration Model

Generating random graphs with a given degree sequence is rather difficult. Thus, to study properties of random graphs with given degree sequences, we do not directly analyze such graphs. Instead, we study random *configurations* with these degree sequences.

Definition 3.12. *Consider $2cn$ points partitioned into n bins where $n \in \mathbb{N}$ and $c > 0$. A* configuration *is a perfect matching of the points into cn pairs.*

Configurations correspond to multigraphs in which the bins are regarded as vertices and the pairs as edges. We refer to such a multigraph as the "underlying" multigraph of the configuration.

For the underlying multigraph to be directed, the points of the configuration must be divided into cn blue points and cn red points which represent the heads and tails of the arcs of the graph respectively. Figure 3.1 is an example of such a configuration.

This leads to a slightly revised definition of a configuration when modeling directed multigraphs.

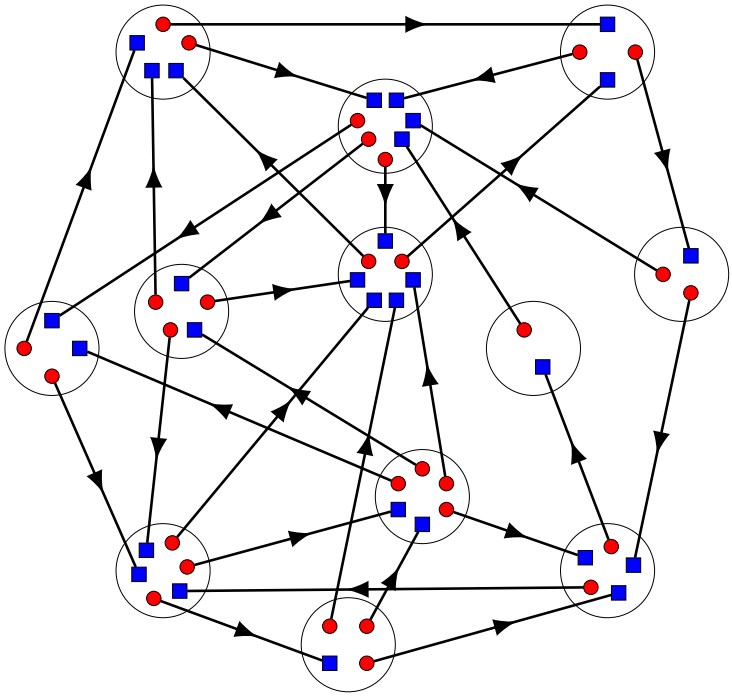


Figure 3.1: A configuration. Blue squares are heads of arcs and red circles are the tails.

Definition 3.13. *Consider a set of cn red points and cn blue points for some $n \in \mathbb{N}$ and $c > 0$. All $2cn$ points are partitioned into n bins. A* configuration *is a bipartite perfect matching of the points, i.e. no pair in the matching is monochromatic.*

When discussing configurations, we refer to Definition 3.12 when the underlying multigraph is undirected and Definition 3.13 when the underlying multigraph is directed. This allows us to use the appropriate type of configuration to model random multigraphs based on the provided degree sequence.

Consider a random graph with degree sequence $((d_1^-, d_2^-, \ldots, d_n^-), (d_1^+, d_2^+, \ldots, d_n^+))$. As this degree sequence is for a directed graph, a configuration with this degree sequence is created as follows:

1. Create n bins and order them 1 through n.
2. For each m, place d_m^- blue points and d_m^+ red points in bin m.
3. Choose a bipartite perfect matching of the red and blue points.

Figure 3.1 is an example of a configuration for a random directed multigraph with degree sequence $((1,1,2,2,1,3,3,2,3,2,5,4),(1,2,2,2,3,2,2,3,3,4,2,3))$.

Although the underlying graph of Figure 3.1 is simple, this is not true for all configurations produced by the above method. However, it is possible to calculate the probability the configuration model generates a simple directed graph for a well-behaved degree sequence. To do so, we apply a result by McKay in [33] for bipartite graphs.

Results for bipartite multigraphs can apply to directed multigraphs because there exists an isomorphism Φ between directed multigraphs and bipartite multigraphs. Specifically, let G be a directed multigraph on n vertices. For each vertex $v_i \in V(G)$, $\Phi(v_i)$ is a pair of vertices (v_i^+, v_i^-) in a graph B such that $d^+(v_i^+) = d^+(v)$, $d^-(v_i^-) = d^-(v)$, and $d^-(v_i^+) = d^+(v_i^-) = 0$. Hence $V(B) = \{v_i^-, v_i^+ \mid 1 \le i \le n\}$ and
$E(B) = \{v_i^+ v_j^- \mid v_i v_j \in E(G), 1 \le i \le n, 1 \le j \le n\}$. By letting $(\{v_1^-, \ldots, v_n^-\}, \{v_1^+, \ldots, v_n^+\})$ be the bipartition of the vertices, it is clear B is bipartite.

It is important to note that for this Φ, $\Phi^{-1}(B)$ could contain a loop when B is simple. Thus, a slight modification is typically needed to transfer results for simple bipartite graphs to results for simple directed graph. However, the result of McKay, stated below as Theorem 3.14, needs no modification due to a particular choice of X.

Before stating the theorem, we introduce some notation. For any integers x and $k > 0$, let $[x]_k = x(x-1)\ldots(x-k+1)$. For a degree sequence $D_n = ((d_1^-, \ldots, d_n^-), (d_1^+, \ldots, d_n^+))$, let $\Delta(D_n) = \max\{\max\{d_i^+ \mid 1 \le i \le n\}, \max\{d_i^- \mid 1 \le i \le n\}\}$. We define $C(D_n)$ to be a random configuration with degree sequence D_n and $\mathcal{G}(C(D_n))$ to be the underlying bipartite multigraph of $C(D_n)$. Finally, we define $\mathbf{Pr}(D_n, X)$ to be the probability $\mathcal{G}(C(D_n))$ is simple and has no edges in common with X, where X is a simple bipartite graph with the same bipartition of vertices as $\mathcal{G}(C(D_n))$.

Theorem 3.14 ([33]). *Let $D_n = ((d_1^-, \ldots, d_n^-), (d_1^+, \ldots, d_n^+))$ be a degree sequence of a bipartite graph and $S = \sum_{i=1}^{n} d_i^-$. Let X be a simple bipartite graph with the same vertex bipartition as $\mathcal{G}(C(D_n))$ and $\Delta(X)$ the maximum degree of a vertex in X. Suppose $\Delta(D_n) \ge 1$ and $\widehat{\Delta} = 3 + 2\Delta(D_n)[2\Delta(D_n) + \Delta(X) + 2] \le \delta S$ for some constant $\delta < \frac{2}{3}$. Then*

$$\mathbf{Pr}(D_n, X) = \exp\left[-\frac{\left(\sum_{i=1}^{n} [d_i^-]_2\right)\left(\sum_{j=1}^{n} [d_j^+]_2\right)}{2S^2} - \frac{\sum_{i=1}^{n} d_i^- d_i^+}{S} + O\left(\frac{\widehat{\Delta}^2}{S}\right) \right].$$

Suppose X is the bipartite graph with bipartition $(\{v_1^-, \ldots, v_n^-\}, \{v_1^+, \ldots, v_n^+\})$ and edge set $E(X) = \{v_i^+ v_i^- \mid 1 \leq i \leq n\}$. Then $\mathbf{Pr}(D_n, X)$ is the probability $\mathcal{G}(C(D_n))$ is simple and $\Phi^{-1}(\mathcal{G}(C(D_n)))$ contains no loops. Hence $\mathbf{Pr}(D_n, X)$ is the probability the configuration model generates a simple directed graph with degree sequence D_n. This leads to the following result for well-behaved degree sequences.

Proposition 3.15. *Let D_n be a well-behaved degree sequence, $(\{v_1^-, \ldots, v_n^-\}, \{v_1^+, \ldots, v_n^+\})$ the vertex partition of $\Phi(\mathcal{G}(C(D_n)))$, and $\Phi(C(D_n))$ the configuration of $\Phi(\mathcal{G}(C(D_n)))$. Let X be a graph with vertex bipartition $(\{v_1^-, \ldots, v_n^-\}, \{v_1^+, \ldots, v_n^+\})$ and $E(X) = \{v_i^+ v_i^- \mid 1 \leq i \leq n\}$. Then,*

$$\mathbf{Pr}(D_n, X) > e^{-A_2^2/2-\lambda}.$$

Proof: Let $D_n = ((d_1^-, \ldots, d_n^-), (d_1^+, \ldots, d_n^+))$. Since D_n is well-behaved, $\Delta(D_n) \geq 1$ and $\widehat{\Delta} = O(\Delta(D_n)^2) < \frac{1}{2}\theta n$, where $\widehat{\Delta}$ is defined as in Theorem 3.14. Hence we may apply Theorem 3.14. Note that $\Delta(D_n) = o(n^{1/4})$ implies $O\left(\frac{\widehat{\Delta}^2}{\theta n}\right) = o(1)$.

$$\begin{aligned}
\mathbf{Pr}(D_n, X) &= \exp\left[-\frac{\left(\sum_{i=1}^{n}[d_i^-]_2\right)\left(\sum_{j=1}^{n}[d_j^+]_2\right)}{2(\theta n)^2} - \frac{\sum_{i=1}^{n} d_i^- d_i^+}{\theta n} + O\left(\frac{\widehat{\Delta}^2}{\theta n}\right)\right] \\
&= \exp\left[-\frac{\left(\sum_{i\geq 0}\sum_{j\geq 0} i(i-1)n_{i,j}\right)\left(\sum_{i\geq 0}\sum_{j\geq 0} j(j-1)n_{i,j}\right)}{2(\theta n)^2} - \frac{\sum_{i\geq 0}\sum_{j\geq 0} ijn_{i,j}}{\theta n} + o(1)\right] \\
&> \exp\left[-\frac{1}{2}\left(\sum_{i\geq 0}\sum_{j\geq 0} i^2\frac{n_{i,j}}{\theta n}\right)\left(\sum_{i\geq 0}\sum_{j\geq 0} j^2\frac{n_{i,j}}{\theta n}\right) - \sum_{i\geq 0}\sum_{j\geq 0} ij\frac{n_{i,j}}{\theta n} + o(1)\right] \\
&\geq e^{\left[-\frac{1}{2}A_2^2 - \lambda + o(1)\right]} \\
&> e^{-A_2^2/2-\lambda}.
\end{aligned}$$

□

Since A_2 and λ are both finite, Proposition 3.15 bounds $\mathbf{Pr}(D_n, X)$ away from 0 for all well-behaved degree sequences D_n. We will use this fact to prove an important relationship between properties in configurations and properties in their underlying graphs. Specifically,

for a well-behaved degree sequence D_n, if $C(D_n)$ a.a.s. has a property $\mathcal{P}$, then a random directed graph with degree sequence D_n a.a.s. has property $\mathcal{P}$. The proof of this statement also requires the following lemma concerning the number of configurations that have simple underlying directed graphs. This lemma is used by Cooper and Frieze in [12].

Lemma 3.16. *Let G be a directed graph with degree sequence $D_n = ((d_1^-, d_2^-, \ldots, d_n^-), (d_1^+, \ldots, d_n^+))$. Then there are exactly $\left(\prod_{m=1}^{n} d_m^-!\right)\left(\prod_{m=1}^{n} d_m^+!\right)$ configurations whose underlying graph is G.*

Proof: First, note that G is provided and so the arcs of G are known. Let v_i, v_j, and v_k be distinct vertices in G and define $S = \sum_{m=1}^{n} d_m^-$. Label the red and blue points of the configuration with integers 1 to $2S$.

Suppose v_jv_i and v_jv_k are arcs of G. There are $d_j^+ d_i^-$ choices for the pair of red and blue points in the configuration that could represent v_jv_i in the underlying multigraph. However, once this pair is chosen, the arc v_jv_k has only $(d_j^+ - 1)d_k^-$ possible pairs to represent it in the configuration. Thus, as the pairs of points are chosen, the remaining arcs have fewer possible pairs that could represent them.

By choosing pairs one at a time, it is clear there are $\left(\prod_{m=1}^{n} d_m^-!\right)\left(\prod_{m=1}^{n} d_m^+!\right)$ configurations whose underlying graph is G. Furthermore, since G is simple, this number is unaffected by the actual choice of arcs of G. Hence every simple directed graph with degree sequence D_n has $\left(\prod_{m=1}^{n} d_m^-!\right)\left(\prod_{m=1}^{n} d_m^+!\right)$ configurations representing it. □

From Lemma 3.16, it is clear that all simple graphs appear uniformly from the configuration model. We now state and prove Proposition 3.17, which is an application of a well-known result stated in [42]. It is an important proposition because it allows us to use the configuration model to prove our results. Note that we say a configuration has a graph property $\mathcal{P}$ if its underlying multigraph has property $\mathcal{P}$.

Proposition 3.17. *If a random configuration with a well-behaved degree sequence a.a.s. has property $\mathcal{P}$, then a random graph with the same degree sequence a.a.s. has property $\mathcal{P}$.*

Proof: Let D_n be a well-behaved degree sequence. Define $\mathcal{G}(D_n)$ to be the uniform probability space of directed graphs with degree sequence D_n. Similarly, define $\mathcal{C}(D_n)$ to be the probability space of configurations with degree sequence D_n. Let $\mathbf{Pr}(\text{Simple})$ denote the

probability in $\mathcal{C}(D_n)$ that the underlying multigraph has no loops or multiple edges. We have the following result found in [42].

Claim 1. *Let $\mathcal{S}$ be a set of graphs in $\mathcal{G}(D_n)$ and $\mathcal{S}'$ the set of configurations in $\mathcal{C}(D_n)$ that correspond to graphs in $\mathcal{S}$. Then, $\mathbf{Pr}_{\mathcal{G}(D_n)}(S) = \frac{\mathbf{Pr}_{\mathcal{C}(D_n)}(S')}{\mathbf{Pr}(Simple)}$.*

Proof: The equation follows immediately from Lemma 3.16 and the uniformity of $\mathcal{G}(D_n)$ and $\mathcal{C}(D_n)$. □

Let $\mathcal{P}$ be a graph property that a.a.s. a random configuration in $\mathcal{C}(D_n)$ has. We choose $\mathcal{S}$ to be the set of graphs in $\mathcal{G}(D_n)$ that do not have property $\mathcal{P}$ and $\mathcal{S}'$ to be the set of configurations in $\mathcal{C}(D_n)$ whose underlying graph is in $\mathcal{S}$. As a configuration a.a.s. has property $\mathcal{P}$, $\mathbf{Pr}_{\mathcal{C}(D_n)}(S') \to 0$. Hence $\mathbf{Pr}_{\mathcal{G}(D_n)}(S) \to 0$ since $\mathbf{Pr}(\text{Simple})$ is bounded away from 0 by Proposition 3.15. Thus $\mathbf{Pr}_{\mathcal{G}(D_n)}(\overline{S}) \to 1$ and so a.a.s. a random graph in $\mathcal{G}(D_n)$ has property $\mathcal{P}$. □

By Proposition 3.17, we can determine the presence of a giant strongly connected component in a random directed graph by studying random configurations with the same degree sequence. We will use this to prove our main result in Chapter 4.

Chapter 4

The Presence of a Giant Strongly Connected Component

As stated in Theorem 2.13, Cooper and Frieze determined a threshold function for the presence of a giant strongly connected component in random graphs with proper degree sequences. This chapter presents a new proof of this function for a slightly larger class of degree sequences, called well-behaved degree sequences. These sequences are defined in Section 2.2, Definition 2.17.

Before stating our result, we define some notation. For a degree sequence $D_n = ((d_1^-, \ldots, d_n^-), (d_1^+, \ldots, d_n^+))$, let $\mathcal{G}(D_n)$ be the set of all directed graphs with degree sequence D_n. Define $\Delta^-(D_n) = \max\{d_i^- \mid 1 \leq i \leq n\}$, $\Delta^+(D_n) = \max\{d_j^+ \mid 1 \leq j \leq n\}$, and $\Delta(D_n) = \min\{\Delta^+(D_n), \Delta^-(D_n)\}$. Also, let $n_{i,j} = |\{1 \leq \ell \leq n \mid d_\ell^- = i, d_\ell^+ = j\}|$ and $\lambda(D_n) = \sum_{i\geq 0}\sum_{j\geq 0} ij\frac{n_{i,j}}{\theta n}$, where θ is as defined in Definition 2.17 (i.e. $\theta = \frac{1}{n}\sum_{i\geq 0}\sum_{j\geq 0} in_{i,j}$). Our main result is the following.

Theorem 4.1. *Let D_n be a well-behaved degree sequence.*

(i) If $\lambda(D_n) < 1$, then a.a.s. every SCC of $G \in \mathcal{G}(D_n)$ has size $O([\Delta(D_n)]^2 \log(n))$.

(ii) If $\lambda(D_n) > 1$, then a.a.s. there exists a SCC in $G \in \mathcal{G}(D_n)$ of size $\Theta(n)$.

To prove Theorem 4.1, we analyze the components of configurations with degree sequence D_n and apply Proposition 3.17. Thus, let D_n be well-behaved and define $\mathcal{C}(D_n)$ to be the set of configurations generated by the configuration model with degree sequence

D_n. For each $C \in \mathcal{C}(D_n)$, let $\mathcal{K}(C)$ be the set of all strongly connected components in C. We analyze the size of the elements of $\mathcal{K}(C)$ using different exploration processes defined in Sections 4.1.1 and 4.2.2.

We will discuss the proof of Theorem 4.1 in two sections. Section 4.1 contains the proof of Theorem 4.1.(i). This proof uses a different method than the proof presented by Cooper and Frieze in [12] for Theorem 2.13.(i). However, it uses the same approach as Molloy and Reed in [35]. We will then use a coupling technique in Section 4.2 to prove Theorem 4.1.(ii).

4.1 The Subcritical Case: $\lambda(D_n) < 1$

We begin this section with a description of an exploration process that exposes pairs of points in $C \in \mathcal{C}(D_n)$. In Section 4.1.2, we will use this process to prove that all elements of $\mathcal{K}(C)$ have size $O([\Delta(D_n)]^2 \log(n))$.

4.1.1 The Exploration Process

Let $V(C)$ be the set of bins in C and $v \in V(C)$. Define $K(v)$ to be the SCC in $\mathcal{K}(C)$ that contains v. To study $|V(K(v))|$, we will analyze the *fan-in* and the *fan-out* of v.

Definition 4.2. *Define the* fan-in *of v to be*

$$\mathcal{F}^-(v) = \{u \in V(C) \mid \textit{there is a directed } uv\textit{-path in } C\}.$$

Similarly, define the fan-out *of v to be*

$$\mathcal{F}^+(v) = \{u \in V(C) \mid \textit{there is a directed } vu\textit{-path in } C\}.$$

Let $F^-(v) = |\mathcal{F}^-(v)|$ and $F^+(v) = |\mathcal{F}^+(v)|$. Note that $|V(K(v))| \leq \min\{F^-(v), F^+(v)\}$. Without loss of generality, assume $\Delta^+(D_n) \leq \Delta^-(D_n)$. We now define an exploration process that will start at v and expose the entire fan-out of v. We will then see that, with high probability, the size of this fan will be small and so the SCC containing v will also be small.

Let $\mathcal{R}$ be the set of all red points and $\mathcal{B}$ be the set of all blue points in C. Each point in C is assigned one of three states: *active*, *used*, or *asleep*. Define $\mathcal{A}_t^+$ to be the set of

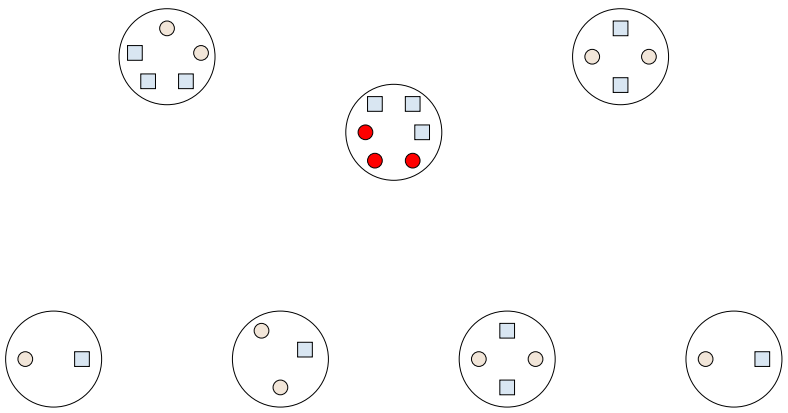
active red points at the end of iteration t and $\mathcal{U}_t^-$ to be the set of used blue points at the end of iteration t. Let $A_t^+ = |\mathcal{A}_t^+|$ and $U_t^- = |\mathcal{U}_t^-|$.

The exploration process Γ_0 begins with all red points in v being active and all other points in C being asleep. Thus $\mathcal{A}_0^+$ is the set of red points in v and $\mathcal{U}_0^- = \emptyset$.

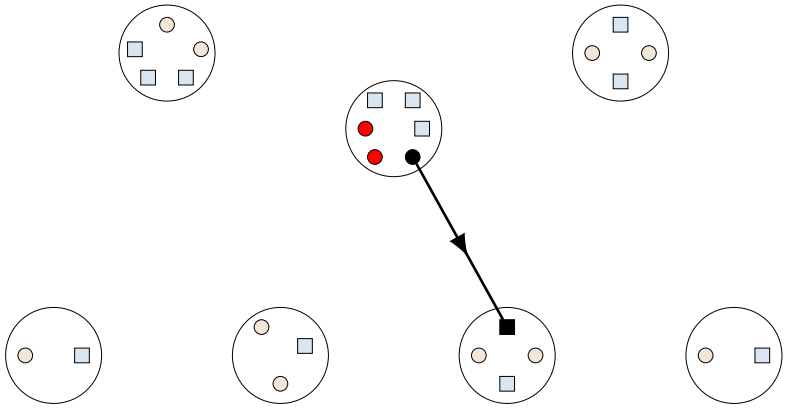
For each integer $t > 0$, the following procedure is performed.

1. If $\mathcal{A}_{t-1}^+ \neq \emptyset$, proceed to Step 2. Otherwise, terminate the procedure.
2. Choose a red point r_t uniformly at random from $\mathcal{A}_{t-1}^+$. Select a blue point b_t uniformly at random from $\mathcal{B}\backslash\mathcal{U}_{t-1}^-$. Let (r_t, b_t) be a pair in C and change the states of r_t and b_t to used.
3. All sleeping red points in the same bin as b_t change their state to active.

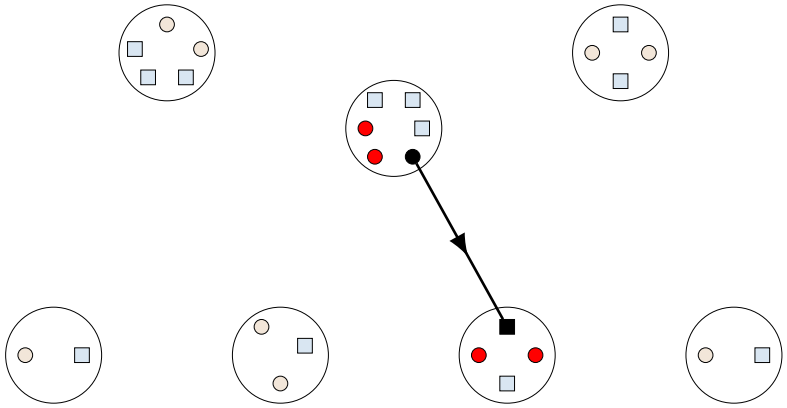
Figure 4.1 outlines one iteration of this procedure.



(a) The start of the exploration process.



(b) Exposing the pair of one active red point (Step 2).



(c) Changing the states of the appropriate points (Step 3).

Figure 4.2: One iteration of the subcritical exploration procedure.

Note that at the start of Γ_0, every pair is present in C with uniform probability. The process then exposes a pair in C during each iteration of its procedure. The remaining pairs in C are distributed as a uniformly random matching from the remaining red points to the remaining blue points. We will use the stopping time of Γ_0 to bound $|V(K(v))|$.

Define τ to be the iteration in which Γ_0 terminates. Then, by definition $\mathcal{A}_{\tau-1}^+ = \emptyset$ and $\mathcal{A}_t^+ \neq \emptyset$ for all $0 \leq t \leq \tau - 2$. Furthermore, recall that exactly one active red point of $\mathcal{A}_t^+$

is not in $\mathcal{A}_{t+1}^+$ for all $0 \le t \le \tau - 2$. Hence $\left|\bigcup_{t=0}^{\tau-1} \mathcal{A}_t^+\right| = \tau$. Thus $F^+(v) \le \tau$ and so

$$|V(K(v))| \le \tau. \tag{4.1}$$

It therefore suffices to bound τ for the proof of Theorem 4.1(i).

4.1.2 Proof of Theorem 4.1(i)

In this section we prove the subcritical portion of Theorem 4.1.

Recall that we assumed D_n is well-behaved, $\Delta(D_n) = \Delta^+(D_n)$, $\lambda(D_n) = 1 - \epsilon$ for some $\epsilon > 0$, and $v \in V(C)$. For convenience, let $\Delta = \Delta(D_n)$ and $\lambda = \lambda(D_n)$. Define $\mathcal{R}(v)$ and $\mathcal{B}(v)$ to be the sets of red and blue points respectively in v and let $R(v) = |\mathcal{R}(v)|$ and $B(v) = |\mathcal{B}(v)|$. Furthermore, define

$$\mathcal{N}_{i,j} = \{v \in V(C) \mid B(v) = i, R(v) = j\}$$

and note $|\mathcal{N}_{i,j}| = n_{i,j}$.

Consider the exploration process Γ_0 defined in Section 4.1.1. Let $\mathcal{H}_t = \bigcup_{i=1}^{t} (r_i, b_i)$, which is the set of pairs exposed by Γ_0 by the end of iteration t. Also, let $\mathcal{U}_t^+$ be the set of all used red points at the end of iteration t. We define

$$\mathcal{N}_{i,j}(t) = \{v \in \mathcal{N}_{i,j} \mid \mathcal{B}(v) \cap \mathcal{U}_t^- = \emptyset, \mathcal{R}(v) \cap \mathcal{U}_t^+ = \emptyset\}$$

and let $|\mathcal{N}_{i,j}(t)| = n_{i,j}(t)$.

Let τ be as in (4.1), i.e. $\tau = \min\{t \ge 1 \mid \mathcal{A}_{t-1}^+ = \emptyset\}$. We then have the following result.

Lemma 4.3. *For any iteration $t < \min\{\tau - 1, \frac{\epsilon\theta}{2+\epsilon} n\}$, $\mathbf{E}(A_{t+1}^+ - A_t^+ \mid \mathcal{H}_t) < -\frac{\epsilon}{2}$.*

Proof: Note that $n_{i,j}(t) \leq n_{i,j}$ for all i and j and $t < \frac{\epsilon\theta}{2+\epsilon}n$ implies $\frac{\epsilon}{2} > \frac{t}{\theta n - t}$. Thus,

$$\begin{aligned}
\mathbf{E}(A^+_{t+1} - A^+_t \mid \mathcal{H}_t) &= \left(A^+_t - 1 + \sum_{j\geq 0}\sum_{i\geq 0} j\frac{in_{i,j}(t)}{\theta n - t}\right) - A^+_t \\
&= -1 + \sum_{j\geq 0}\sum_{i\geq 0} j\frac{in_{i,j}(t)}{\theta n - t} \\
&\leq -1 + \sum_{j\geq 0}\sum_{i\geq 0} j\frac{in_{i,j}}{\theta n - t} \\
&= -1 + \frac{\theta n}{\theta n - t}\sum_{j\geq 0}\sum_{i\geq 0} ij\frac{n_{i,j}}{\theta n} \\
&= -1 + \left(1 + \frac{t}{\theta n - t}\right)\lambda \\
&< -1 + (1-\epsilon) + \frac{t}{\theta n - t} \\
&< -\frac{\epsilon}{2}.
\end{aligned}$$

□

This allows us to study the number of iterations that Γ_0 performs by using super-martingale inequalities.

Lemma 4.4. *For any $t' < \min\{\tau - 1, \frac{\epsilon\theta}{2+\epsilon}n\}$, $\{A^+_t + \frac{\epsilon}{2}t\}_{0\leq t\leq t'}$ is a super-martingale.*

Proof: For any iteration $t \leq t'$, $0 \leq A^+_t \leq \theta n$ and so $\mathbf{E}\left(\left|A^+_t + \frac{\epsilon}{2}t\right|\right) < \infty$. Furthermore, $A^+_{t^*} > 0$ for all $0 \leq t^* \leq t$. Hence,

$$\begin{aligned}
\mathbf{E}\left[\left[A^+_{t+1} + \frac{\epsilon}{2}(t+1)\right] - \left(A^+_t + \frac{\epsilon}{2}t\right) \mid \mathcal{H}_t\right] &= \mathbf{E}\left(A^+_{t+1} - A^+_t + \frac{\epsilon}{2} \mid \mathcal{H}_t\right) \\
&= \mathbf{E}(A^+_{t+1} - A^+_t \mid \mathcal{H}_t) + \frac{\epsilon}{2} \\
&< -\frac{\epsilon}{2} + \frac{\epsilon}{2} \\
&= 0.
\end{aligned}$$

Thus by Definition 3.7, $\{A^+_t + \frac{\epsilon}{2}t\}_{0\leq t\leq t'}$ is a super-martingale. □

Let $\tau^* = \min\{\tau - 1, \frac{\epsilon\theta}{2+\epsilon}n\}$. Lemma 4.4 then implies that $\{A^+_t + \frac{\epsilon}{2}t\}_{0\leq t\leq \tau^*}$ is a super-martingale. To bound τ, we define another super-martingale $\{X_i\}_{i\geq 0}$ by $X_t = A_t$ for all $0 \leq t \leq \tau^*$ and $X_t = X_{t-1} - \frac{\epsilon}{2}$ for all $t > \tau^*$.

Proposition 4.5. *Let* $s = \left\lceil \frac{6\Delta^2}{\epsilon^2} \log(n) \right\rceil$. *Then* $\mathbf{Pr}(X_s \geq 0) = O\left(\frac{1}{\sqrt{n}}\right)$.

Proof: We have

$$\left| \left[X_{t+1} + \frac{\epsilon}{2}(t+1) \right] - \left(X_t + \frac{\epsilon}{2} t \right) \right| \leq \left| (\Delta - 1 + X_t) + \frac{\epsilon}{2} - X_t^+ \right| < \Delta$$

for $t \leq \tau^*$ and

$$|X_{t+1} - X_t| = \frac{\epsilon}{2} < \Delta$$

for $t > \tau^*$. Thus by Theorem 3.9 with $c_t = \Delta$ for all t,

$$\begin{aligned} \mathbf{Pr}(X_t \geq 0) &= \left[\left(X_t + \frac{\epsilon}{2} t \right) - \left(X_0 + \frac{\epsilon}{2}(0) \right) \geq \frac{\epsilon}{2} t - X_0 \right] \\ &\leq \exp\left[- \left[\frac{\epsilon}{2} t - X_0 \right]^2 / 2\Delta^2 t \right] \\ &\leq e^{-(\epsilon t - 2X_0)^2 / 8\Delta^2 t}. \end{aligned}$$

Recall that $s = \left\lceil \frac{6\Delta^2}{\epsilon^2} \log(n) \right\rceil$, and so

$$\begin{aligned} \mathbf{Pr}(X_s \geq 0) &\leq e^{-[\epsilon s - 2X_0]^2 / 8\Delta^2 s} \\ &= \exp\left[- \left(\frac{3\Delta^2}{\epsilon} \log(n) - X_0 \right)^2 / \left(\frac{12\Delta^4}{\epsilon^2} \log(n) \right) \right] \\ &\leq e^{-(3\Delta^2 \log(n) - \Delta\epsilon)^2 / (12\Delta^4 \log(n))} \\ &= e^{-(3/4)\log(n) + (\epsilon/2\Delta) - (\epsilon^2 / 12\Delta^2 \log(n))} \\ &< e^{-(3/4)\log(n) + (1/4)\log(n)} \\ &= n^{-1/2}. \end{aligned}$$

□

Proposition 4.5 immediately implies that a.a.s. $\tau \leq \left\lceil \frac{6\Delta^2}{\epsilon^2} \log(n) \right\rceil$, which completes the proof of Theorem 4.1(i) by equation (4.1) and Proposition 3.17. In the remaining sections of this chapter, we complete the proof of Theorem 4.1(ii).

4.2 The Supercritical Case: $\lambda(D_n) > 1$

In this section, we will prove Theorem 4.1.(ii). This will require a new exploration process Γ that consists of a sequence of subroutines $\widehat{\Gamma}_i$. Each $\widehat{\Gamma}_i$ begins by choosing a bin v in

some manner and then searches both the fan-in and fan-out of v at the same time. A subroutine stops when some bad event (described in Section 4.2.2) occurs or it successfully finds a "good" bin v (defined in Section 4.2.2). In Section 4.2.4 we will show that a.a.s. v is contained in a giant SCC.

To analyze this Γ, we will couple each $\widehat{\Gamma}_i$ with two independent random walks $\{Y_t^-\}_{t\geq 0}$ and $\{Y_t^+\}_{t\geq 0}$ defined in Section 4.2.1. In Section 4.2.3, we use these walks to define the stopping times of the subroutines as well as show that a.a.s. Γ finds a good bin v.

4.2.1 The Random Walks $\{Y_t^+\}_{t\geq 0}$ and $\{Y_t^-\}_{t\geq 0}$

Recall that D_n is a well-behaved degree sequence, $\lambda(D_n) = 1+\epsilon$ for some $\epsilon > 0$, and $C \in \mathcal{C}(D_n)$. Let x be a point in C and define $\text{bin}(x)$ to be the bin containing x. As before, for each $v \in V(C)$ let $\mathcal{B}(v) = \{x \in \mathcal{B} \mid v = \text{bin}(x)\}$, $\mathcal{R}(v) = \{x \in \mathcal{R} \mid v = \text{bin}(x)\}$, $B(v) = |\mathcal{B}(v)|$, and $R(v) = |\mathcal{R}(v)|$. Also, recall $\mathcal{N}_{i,j} = \{v \in V(C) \mid B(v) = i, R(v) = j\}$ and $n_{i,j} = |\mathcal{N}_{i,j}|$.

Unlike the exploration process in Section 4.1.1, the exploration process Γ we informally describe below will expose both the fan-in and fan-out of a bin at the same time. This process continues to explore the fan-ins and fan-outs of various bins until a bin is found with certain properties. These particular properties (discussed in Section 4.2.2) make it likely to be in a giant SCC and so we call such a bin "good." However, some events can occur in the exploration process that interfere with checking these properties. These events (discussed in Section 4.2.2) will force the exploration process to terminate and so are considered "bad."

We now informally describe the procedure of the subroutine $\widehat{\Gamma}_i$ in the exploration process Γ. Its formal definition can be found in Section 4.2.2. As in the previous exploration process, the points in C are assigned one of three states: *active*, *used*, or *asleep*. Let $\mathcal{A}_t^+$ and $\mathcal{A}_t^-$ be the sets of active red and active blue points respectively at the end of iteration t. Also, let $\mathcal{U}_t^+$ and $\mathcal{U}_t^-$ be the sets of used red and used blue points respectively at the end of iteration t. Define $A_t^+ = |\mathcal{A}_t^+|$, $A_t^- = |\mathcal{A}_t^-|$, $U_t^+ = |\mathcal{U}_t^+|$, and $U_t^- = |\mathcal{U}_t^-|$.

In each iteration of its procedure, $\widehat{\Gamma}_i$ will choose some $\mathbf{r}_t \in \mathcal{A}_{t-1}^+$ and $\mathbf{b}_t \in \mathcal{A}_t^-$. It then uniformly at random chooses $b_t \in \mathcal{B}\backslash\mathcal{U}_{t-1}^-$ and pairs $\mathbf{r}_t$ with b_t. The states of all sleeping red points in $\text{bin}(b_t)$ change to active and $\mathbf{r}_t$ and b_t change their state to used. If $\mathbf{b}_t \neq b_t$, $\widehat{\Gamma}_i$ uniformly at random chooses some $r_t \in \mathcal{R}\backslash(\mathcal{U}_{t-1}^+ \cup \{\mathbf{r}_t\})$ and pairs $\mathbf{b}_t$ with r_t. The states of all sleeping blue points in $\text{bin}(r_t)$ changes to active and $\mathbf{b}_t$ and r_t change their states to used. The subroutine will then terminate if some bad event occurred, such as $\mathbf{b}_t = b_t$.

By the nature of this process, it is clear that it is difficult to analyze $\{A_t^+\}_{t\geq 0}$ and $\{A_t^-\}_{t\geq 0}$ directly. Therefore, we will define two random walks $\{Y_t^+\}_{t\geq 0}$ and $\{Y_t^-\}_{t\geq 0}$ such that:

1. $\mathbf{Pr}(Y_t^+ - Y_{t-1}^+ = j)$ and $\mathbf{Pr}(Y_t^- - Y_{t-1}^- = i)$ are independent of $\{Y_i^+\}_{0\leq i<t-1}$ and $\{Y_i^-\}_{0\leq i<t-1}$,
2. $\{Y_t^+\}_{t\geq 0}$ and $\{Y_t^-\}_{t\geq 0}$ are independent, and
3. $\{A_t^+\}_{t\geq 0}$ almost always stochastically *dominates* $\{Y_t^+\}_{t\geq 0}$ and $\{A_t^-\}_{t\geq 0}$ almost always stochastically *dominates* $\{Y_t^-\}_{t\geq 0}$.

We then couple these walks with $\{A_t^+\}_{t\geq 0}$ and $\{A_t^-\}_{t\geq 0}$ in the definition of $\widehat{\Gamma}_i$ in order to analyze $\{A_t^+\}_{t\geq 0}$ and $\{A_t^-\}_{t\geq 0}$.

Definition 4.6. *Let I be an index set and X_i be a random discrete variable defined on a probability space (Ω_i, P_i) for each $i \in I$. A* coupling *of the X_i is a collection of random variables $\widehat{X}_i$ defined on a common probability space (Ω, P) such that the marginal distribution of $\widehat{X}_i$ is the same as the distribution of X_i for each $i \in I$.*

Definition 4.7. *A random walk $\{X_i\}_{i\geq 0}$* dominates *a random walk $\{Y_i\}_{i\geq 0}$ if there exists a coupling of the X_i and Y_i such that $\widehat{X}_i \geq \widehat{Y}_i$ for all $i \geq 0$.*

Let $p_j^+(t)$ be the conditional probability that $\text{bin}(b_t)$ contains exactly j sleeping red points at the start of iteration t given the states of all points at the end of iteration $t-1$. Similarly, let $p_i^-(t)$ be the conditional probability in iteration t that $\text{bin}(r_t)$ contains exactly i sleeping blue points after b_t changes its state to used given the states of all points after the pair $(\mathbf{r}_t, b_t)$ is exposed by $\widehat{\Gamma}_i$. Roughly speaking, $p_j^+(t)$ is the conditional probability $A_t^+ - A_{t-1}^+ = j-1$ and $p_i^-(t)$ is the conditional probability $A_t^- - A_{t-1}^- = i-1$.

To satisfy the desired properties for $\{Y_t^+\}_{t\geq 0}$ and $\{Y_t^-\}_{t\geq 0}$, we require the following lemma to hold.

Lemma 4.8. *There exist two distribution functions ϕ^+ and ϕ^- such that:*

1. *There exists a constant k such that the domain of ϕ^+ and ϕ^- is $\{0, 1, \ldots, k\}$.*
2. *There exists $\sigma > 0$ such that for all $0 \leq t \leq \sigma n + 1$ and $i, j \geq 1$, a.a.s. $p_j^+(t) \geq \phi_j^+$ and $p_i^-(t) \geq \phi_i^-$, and*

3. $\sum_{j=0}^{k} j\phi_j^+ \geq 1 + \frac{\epsilon}{4}$ *and* $\sum_{i=0}^{k} i\phi_i^- \geq 1 + \frac{\epsilon}{4}$.

We prove Lemma 4.8 later in this section. For now, assume the statement holds.

Let η^+ be a random variable that takes on the value $j-1$ with probability ϕ_j^+ for all $j \geq 0$. Also, let η^- be a random variable that takes on the value $i-1$ with probability ϕ_i^- for all $i \geq 0$. For all $t \geq 0$, let η_t^+ be an independent copy of η^+ and η_t^- be an independent copy of η^-. We define Y_t^+ and Y_t^- such that:

1. $Y_0^+ = Y_0^- = 1$ and
2. $Y_{t+1}^+ = Y_t^+ + \eta_t^+$ and $Y_{t+1}^- = Y_t^- + \eta_t^-$ for all $t \geq 1$.

Since η^+ and η^- are independent random variables, $\{Y_t^+\}_{t\geq 0}$ and $\{Y_t^-\}_{t\geq 0}$ are independent walks. Furthermore, $\mathbf{Pr}(Y_{t+1}^+ - Y_t^+ = j-1) = \phi_j^+$ and $\mathbf{Pr}(Y_{t+1}^- - Y_t^- = i-1) = \phi_i^-$ for all $t \geq 0$. Hence $\{Y_t^+\}_{t\geq 0}$ and $\{Y_t^-\}_{t\geq 0}$ satisfy the first two desired properties. We will prove that $\{A_t^+\}_{t\geq 0}$ and $\{A_t^-\}_{t\geq 0}$ almost always stochastically dominate $\{Y_t^+\}_{t\geq 0}$ and $\{Y_t^-\}_{t\geq 0}$ for some number of iterations of Γ after formally defining Γ in Section 4.2.2.

In the remainder of this section, we provide the proof of Lemma 4.8.

Proof of Lemma 4.8

Recall that D_n is well-behaved and $\lambda(D_n) = 1 + \epsilon$ for some $\epsilon > 0$. Thus there exists constants k and N such that for all $n > N$,

$$\left| \sum_{i=0}^{k} \sum_{j=0}^{k} ij \frac{n_{i,j}}{\theta n} - \lambda \right| < \frac{\epsilon}{2}. \tag{4.2}$$

Hence $\sum_{i=0}^{k} \sum_{j=0}^{k} ij\frac{n_{i,j}}{\theta n} > 1 + \frac{\epsilon}{2}$. We will define ϕ^+ and ϕ^- on $\{0, \ldots, k\}$ for k satisfying (4.2).

Before defining ϕ^+ and ϕ^-, we need some lower bounds for $p_j^+(t)$ and $p_i^-(t)$ for all $t \geq 1$. At the start of iteration $t = 1$, the probability of choosing a blue point b_0 such that $R(\text{bin}(b_0)) = j$ is

$$p_j^+ = p_j^+(1) = \sum_{i\geq 0} i\frac{n_{i,j}}{\theta n}.$$

Furthermore, the probability of then choosing a red point r_0 such that $B(\text{bin}(r_0)) = i$ and $b_0 \notin \mathcal{B}(\text{bin}(r_0))$ is

$$p_i^- = p_i^-(1) = \sum_{j \geq 0} \frac{j}{\theta n}(n_{i,j} - \mathbb{1}_{\text{bin}(b_0) \in \mathcal{N}_{i,j}}).$$

We then have the following property.

Lemma 4.9. *For all $t \geq 1$, $p_j^+(t+1) > \sum_{i=0}^{k} i\frac{n_{i,j}}{\theta n} - \frac{4kt}{\theta n - 2t}$ and $p_i^-(t+1) > \sum_{j=0}^{k} j\frac{n_{i,j}}{\theta n} - \frac{4kt}{\theta n - 2t}$.*

Proof: Let $\mathcal{N}_{i,j}(t)$ be the set of bins contain exactly i sleeping blue and j sleeping red points at the beginning of iteration t and let $n_{i,j}(t) = |\mathcal{N}_{i,j}(t)|$. By the procedure of $\widehat{\Gamma}_i$, it is clear that $\text{bin}(\mathbf{r}_t)$, $\text{bin}(b_t)$, $\text{bin}(\mathbf{b}_t)$ and $\text{bin}(r_t)$ are the only bins that could be present in $\mathcal{N}_{i,j}(t)$ but not in $\mathcal{N}_{i,j}(t+1)$. Thus for all i and j, $n_{i,j}(t-1) - 4 \leq n_{i,j}(t) \leq n_{i,j}(t-1) + 4$. Hence for all $t \geq 1$,

$$\begin{aligned}
p_j^+(t+1) &= \sum_{i \geq 0} i\frac{n_{i,j}(t+1)}{\theta n - 2t} \\
&\geq \sum_{i=0}^{k} i\frac{n_{i,j}(t+1)}{\theta n - 2t} \\
&\geq \left(\sum_{i=0}^{k} i\frac{n_{i,j}}{\theta n - 2t}\right) - \frac{4kt}{\theta n - 2t} \\
&= \frac{\theta n}{\theta n - 2t}\left(\sum_{i=0}^{k} i\frac{n_{i,j}}{\theta n}\right) - \frac{4kt}{\theta n - 2t} \\
&> \sum_{i=0}^{k} i\frac{n_{i,j}}{\theta n} - \frac{4kt}{\theta n - 2t}.
\end{aligned}$$

A similar calculation shows that $p_i^-(t+1) > \sum_{i=0}^{k} j\frac{n_{i,j}}{\theta n} - \frac{4kt}{\theta n - 2t}$ for all $t \geq 1$. □

We will use these bounds for certain choices of j and i to define ϕ^+ and ϕ^-. First, let $q_j^+ = \sum_{i=0}^{k} i\frac{n_{i,j}}{\theta n}$ and $q_i^- = \sum_{j=0}^{k} j\frac{n_{i,j}}{\theta n}$. We then have the following lemma.

Lemma 4.10. *Let k be as in (4.2). Then for all $1 \leq j \leq k$ such that $\lim_{n\to\infty} q_j^+ > 0$, there exists a real number a_j^+ such that $0 < a_j^+ < \frac{\epsilon}{4+2\epsilon} \lim_{n\to\infty} q_j^+$. Similarly, for all $1 \leq i \leq k$ such that $\lim_{n\to\infty} q_i^- > 0$, there exists a real number a_i^- such that $0 < a_i^- < \frac{\epsilon}{4+2\epsilon} \lim_{n\to\infty} q_i^-$.*

Proof: Let $1 \leq j \leq k$ be an integer such that $\lim_{n\to\infty} p_j^+ > 0$. By the choice of j, it is clear $\frac{\epsilon}{4+2\epsilon} \lim_{n\to\infty} q_j^+ > 0$. Thus the density of the reals implies there exists a real number a_j^+ such that $0 < a_j^+ < \frac{\epsilon}{4+2\epsilon} \lim_{n\to\infty} q_j^+$.

The proof for the existence of a_i^- for $1 \leq i \leq k$ such that $\lim_{n\to\infty} q_i^- > 0$ is the same as for the existence of a_j^+ since $0 < \frac{\epsilon}{4+2\epsilon} \lim_{n\to\infty} q_i^-$ for all such i. □

Let k be as in (4.2) and a_j^+ and a_i^- be as in Lemma 4.10 for all $1 \leq j \leq k$ and $1 \leq i \leq k$ such that $\lim_{n\to\infty} q_j^+ > 0$ and $\lim_{n\to\infty} q_i^- > 0$. We define ϕ^+ and ϕ^- on $0 \leq j \leq k$ and $0 \leq i \leq k$ respectively, as follows.

$$\phi_j^+ = \begin{cases} \lim_{n\to\infty} q_j^+ - a_j^+ & \text{if } \lim_{n\to\infty} q_j^+ > 0 \text{ and } 1 \leq j \leq k \\ 0 & \text{if } \lim_{n\to\infty} q_j^+ = 0 \text{ and } 1 \leq j \leq k \\ 1 - \sum_{\ell=1}^{k} \phi_\ell^+ & \text{if } j = 0 \end{cases} \tag{4.3}$$

$$\phi_i^- = \begin{cases} \lim_{n\to\infty} q_i^- - a_i^- & \text{if } \lim_{n\to\infty} q_i^- > 0 \text{ and } 1 \leq i \leq k \\ 0 & \text{if } \lim_{n\to\infty} q_i^- = 0 \text{ and } 1 \leq i \leq k \\ 1 - \sum_{\ell=1}^{k} \phi_\ell^- & \text{if } i = 0 \end{cases} \tag{4.4}$$

Clearly (4.3) and (4.4) satisfy the first condition of Lemma 4.8. We now prove they satisfy the remaining two conditions in Lemmas 4.11 and 4.12. First, note that $\phi_j^+ \leq \lim_{n\to\infty} q_j^+$ for all $1 \leq j \leq k$. Furthermore, equality holds only if $\lim_{n\to\infty} q_j^+ = 0$. Similarly, $\phi_i^- \leq \lim_{n\to\infty} q_i^-$ for all $1 \leq i \leq k$ with equality holding only if $\lim_{n\to\infty} q_i^- = 0$. This leads to the following result.

Lemma 4.11. *Let ϕ^+ and ϕ^- be as in (4.3) and (4.4). Then there exists $\sigma > 0$ such that for all integers $0 \leq t \leq \sigma n + 1$ and $i, j \geq 1$, $p_j^+(t) \geq \phi_j^+$ and $p_i^-(t) \geq \phi_i^-$.*

Proof. Suppose $\lim_{n\to\infty} q_j^+ = 0$. Then $\phi_j^+ = 0$ and $p_j^+(t) \geq 0$ for all t, so there is nothing to show. Furthermore, for all $j > k$, $\phi_j^+ = 0$ and $p_j^+(t) \geq 0$ for all t. Hence for all $j > k$, $p_j^+(t) \geq \phi_j^+$ for all t. Similarly, $\phi_i^- = 0 \leq p_i^-(t)$ for all t if $\lim_{n\to\infty} q_i^- = 0$ and for all $i > k$, $p_i^-(t) \geq \phi_i^- = 0$ for all t. Thus, assume $1 \leq i \leq k$, $1 \leq j \leq k$, $\lim_{n\to\infty} q_j^+ > 0$, and $\lim_{n\to\infty} q_i^- > 0$.

Let $\delta^+ = \min\{a_j^+ \mid 1 \leq j \leq k\}$, $\delta^- = \min\{a_i^- \mid 1 \leq i \leq k\}$, and $\delta = \min\{\delta^+, \delta^-\}$. By Lemma 4.9, we have $p_j^+(t+1) > \sum_{i=0}^{k} i\frac{n_{i,j}}{\theta n} - \frac{4kt}{\theta n - 2t}$ and $p_i^-(t+1) > \sum_{j=0}^{k} j\frac{n_{i,j}}{\theta n} - \frac{4kt}{\theta n - 2t}$ for all $t \geq 1$. Thus for sufficiently large n,

$$p_j^+(t+1) > \lim_{n\to\infty} q_j^+ - \frac{6kt}{\theta n - 2t} \tag{4.5}$$

and

$$p_i^-(t+1) > \lim_{n\to\infty} q_i^- - \frac{6kt}{\theta n - 2t}. \tag{4.6}$$

For $t = \frac{\delta\theta}{6k+2\delta}n$, (4.5) and (4.6) imply $p_j^+(t+1) > \lim_{n\to\infty} q_i^- - \delta \geq \phi_i^-$ and $p_i^-(t+1) > \lim_{n\to\infty} q_i^- - \delta \geq \phi_i^-$. Thus the statement holds for $\sigma = \frac{\delta\theta}{6k+2\delta}$. □

In addition, we have the following lemma.

Lemma 4.12. *Let ϕ^+ and ϕ^- be as in (4.2) and (4.3). Then $\sum_{j=0}^{k} j\phi_j^+ > 1 + \frac{\epsilon}{4}$ and $\sum_{i=0}^{k} i\phi_i^- > 1 + \frac{\epsilon}{4}$.*

Proof: Recall $a_j^+ < \frac{\epsilon}{4+2\epsilon}\lim_{n\to\infty} q_j^+$. Thus,

$$\begin{aligned}
\sum_{j=0}^{k} j\phi_j^+ &> \sum_{j=0}^{k} j\left(\lim_{n\to\infty} q_j^+ - \frac{\epsilon}{4+2\epsilon}\lim_{n\to\infty} q_j^+\right) \\
&= \left(1 - \frac{\epsilon}{4+2\epsilon}\right)\sum_{j=0}^{k} j\lim_{n\to\infty} q_j^+ \\
&\geq \left(1 - \frac{\epsilon}{4+2\epsilon}\right)\left(1 + \frac{\epsilon}{2}\right) \\
&= 1 + \frac{\epsilon}{4}.
\end{aligned}$$

A similar calculation shows $\sum_{i=0}^{k} i\phi_i^- > 1 + \frac{\epsilon}{4}$. □

Lemma 4.8 then follows from Lemmas 4.11 and 4.12. In the next section, we complete the definition of the exploration process used in the proof of Theorem 4.1(ii). This process will couple Y_t^+ and Y_t^- with A_t^+ and A_t^- so that $A_t^+ \geq Y_t^+$ and $A_t^- \geq Y_t^-$ for all iterations of Γ.

4.2.2 The Exploration Process

Recall that Γ is a sequence of subroutines $\widehat{\Gamma}_i$. Each subroutine $\widehat{\Gamma}_i$ performs iterations of a procedure defined in this section until some bad event forces $\widehat{\Gamma}_i$ to stop or a good bin is found. This bad event as well as some bad events that will force Γ to terminate are discussed later in this section.

We first define a *good* bin. As mentioned in the previous section, the properties of a good bin make it likely to have a large fan-in and fan-out as well as only need a small number of iterations to determine these fans are large. These properties in turn make it likely that the bin is in a giant SCC.

Definition 4.13. *Assume a subroutine $\widehat{\Gamma}$ starts in iteration t_0. We say $\widehat{\Gamma}$ finds a* good *bin v if it starts at v and there exists some $s \leq \log^3(n)$ and constant $c > 0$ such that:*

1. $t_0 \leq s \leq t_0 + c\log(n)$,
2. $A_t^+, A_t^- > 0$ *for all* $t_0 \leq t \leq s$, *and*
3. $A_s^+, A_s^- \geq \frac{c\epsilon}{4}\log(n)$.

From this definition, it is clear a subroutine $\widehat{\Gamma}$ will not find a good bin after iteration $\log^3(n)$. We will later define a stopping time for Γ which ensures that each subroutine never takes more than $\log^3(n)$ iterations (see event (B1) in the definition of the procedure of $\widehat{\Gamma}$). Before defining the procedure of $\widehat{\Gamma}$, we discuss some notation.

For all $i \geq 0$, let $\mathcal{B}_i$ be the set of all blue points in bins that contain exactly i blue points and at least one red point, i.e.

$$\mathcal{B}_i = \{b \in \mathcal{B} \mid B(\text{bin}(b)) = i, R(\text{bin}(b)) > 0\}.$$

Similarly, for all $j \geq 0$, let

$$\mathcal{R}_j = \{r \in \mathcal{R} \mid R(\text{bin}(r)) = j, B(\text{bin}(r)) > 0\}.$$

We define

$$\mathcal{B}_{>i} = \{b \in \mathcal{B} \mid B(\text{bin}(b)) > i, R(\text{bin}(b)) > 0\}$$

and

$$\mathcal{R}_{>j} = \{r \in \mathcal{R} \mid R(\text{bin}(r)) > j, B(\text{bin}(r)) > 0\}.$$

Let $B_i = |\mathcal{B}_i|$, $R_j = |\mathcal{R}_j|$, $B_{>i} = |\mathcal{B}_{>i}|$, and $R_{>j} = |\mathcal{R}_{>j}|$.

In addition, define $\mathcal{B}_i(t)$ to be the set of blue points in bins containing exactly i sleeping blue and at least one unused red point at the start of iteration t, i.e.

$$\mathcal{B}_i(t) = \{b \in \mathcal{B} \mid |\mathcal{B}(\text{bin}(b))\backslash(\mathcal{A}^-_{t-1} \cup \mathcal{U}^-_{t-1})| = i, |\mathcal{R}(\text{bin}(b))\backslash\mathcal{U}^+_{t-1}| > 0\}.$$

Hence $\mathcal{B}_i = \mathcal{B}_i(0)$. Similarly, let

$$\mathcal{R}_j(t) = \{r \in \mathcal{R} \mid |\mathcal{R}(\text{bin}(r))\backslash(\mathcal{A}^+_{t-1} \cup \mathcal{U}^+_{t-1})| = j, |\mathcal{B}(\text{bin}(r))\backslash\mathcal{U}^-_{t-1}| > 0\}.$$

We then define $\mathcal{B}(\mathcal{R}_j(t))$ to be the set of blue points in bins that contain red points in $\mathcal{R}_j(t)$ and $\mathcal{R}(\mathcal{B}_i(t))$ to be the set of red points in bins that contain blue points in $\mathcal{B}_i(t)$. Let $\mathcal{B}(\mathcal{R}_{>j}(t))$ and $\mathcal{R}(\mathcal{B}_{>i}(t))$ be the sets of blue and red points respectively in bins containing more than j red or i blue sleeping points respectively at the start of iteration t.

We formally define Γ as follows. The process starts by assigning all points in C the sleeping state. It then begins the subroutine $\widehat{\Gamma}_1$ in iteration $t_0 = 0$ and sets $\mathcal{U}^+_{-1} = \mathcal{U}^-_{-1} = \emptyset$.

For each $i \geq 1$, $\widehat{\Gamma}_i$ begins iteration t_0 by returning all unused points to the sleeping state. A bin v such that $R(v), B(v) > 0$ is then chosen arbitrarily and all red and blue points in v change their state to active. It then defines $\widehat{Y}^+_{t_0} = \widehat{Y}^-_{t_0} = 1$. This ends iteration t_0 and so $\mathcal{A}^+_{t_0} = \mathcal{R}(v)$, $\mathcal{A}^-_{t_0} = \mathcal{B}(v)$, $\mathcal{U}^+_{t_0} = \mathcal{U}^+_{t_0-1}$, and $\mathcal{U}^-_{t_0} = \mathcal{U}^-_{t_0-1}$.

In each iteration $t > t_0$ of $\widehat{\Gamma}_i$, the following procedure is performed.

1. A red point $\mathbf{r}_t \in \mathcal{A}^+_{t-1}$ and a blue point $\mathbf{b}_t \in \mathcal{A}^-_{t-1}$ are chosen.
2. One of the following is performed with the corresponding probability:
   (a) With probability ϕ^+_j for each $1 \leq j \leq k$, let $\widehat{Y}^+_t = \widehat{Y}^+_{t-1} + (j-1)$ and b_t be chosen uniformly at random from $\mathcal{B}(\mathcal{R}_j(t))\backslash\mathcal{U}^-_{t-1}$. Pair $\mathbf{r}_t$ with b_t.

(b) With probability $p_j^+(t) - \phi_j^+$ for each $1 \le j \le k$, let $\widehat{Y}_t^+ = \widehat{Y}_{t-1}^+ - 1$ and b_t be chosen uniformly at random from $\mathcal{B}(\mathcal{R}_j(t)) \backslash \mathcal{U}_{t-1}^-$. Pair $\mathbf{r}_t$ with b_t.

(c) With the remaining probability, let $\widehat{Y}_t^+ = \widehat{Y}_{t-1}^+ - 1$ and b_t be chosen uniformly at random from $[\mathcal{B}(\mathcal{R}_{>k}(t)) \cup \mathcal{B}(\mathcal{R}_0(t))] \backslash \mathcal{U}_{t-1}^-$.

Change the states of $\mathbf{r}_t$ and b_t to used and the states of all sleeping red points in $\text{bin}(b_t)$ to active.

3. If $\mathbf{b}_t \neq b_t$, repeat Step 2 using ϕ_i^-, $p_i^-(t)$, Y_t^-, $\mathcal{R}(\mathcal{B}_i(t))$, $\mathcal{R}(\mathcal{B}_{>k}(t))$, $\mathcal{R}(\mathcal{B}_0(t))$, $\mathcal{U}_{t-1}^+ \cup \{\mathbf{r}_t\}$, $\mathbf{b}_t$, and r_t. Change the states of r_t and $\mathbf{b}_t$ to used and the state of all sleeping blue points in $\text{bin}(r_t)$ to active.
If $\mathbf{b}_t = b_t$, let $\widehat{Y}_t^- = \widehat{Y}_{t-1}^- - 1$ and change the state of all sleeping blue points in $\mathbf{r}_t$ to active.

4. Terminate the procedure and repeat it for subroutine $\widehat{\Gamma}_{i+1}$ in iteration $t_0 = t + 1$ if

(E) $\widehat{Y}_t^+ = 0$ or $\widehat{Y}_t^- = 0$

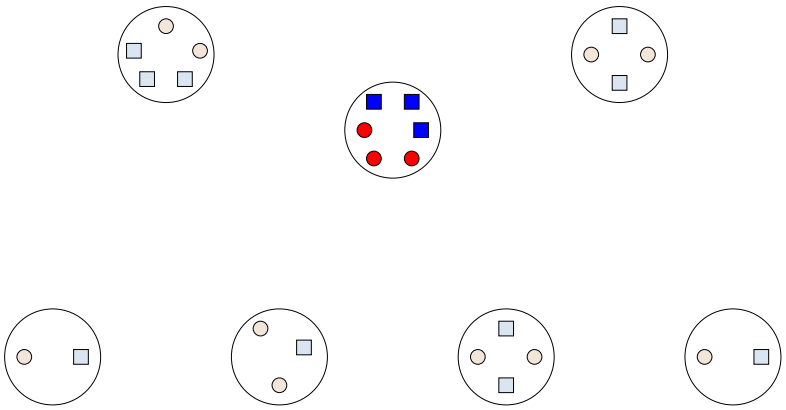
and abort Γ if one of the following occurs:

(B1) $t \ge t_0 + \lceil \log^2(n) \rceil$ and for each $t_0 \le s \le t$, $\widehat{Y}_s^+ < \frac{c\epsilon}{4} \log(n)$ or $\widehat{Y}_s^- < \frac{c\epsilon}{4} \log(n)$.

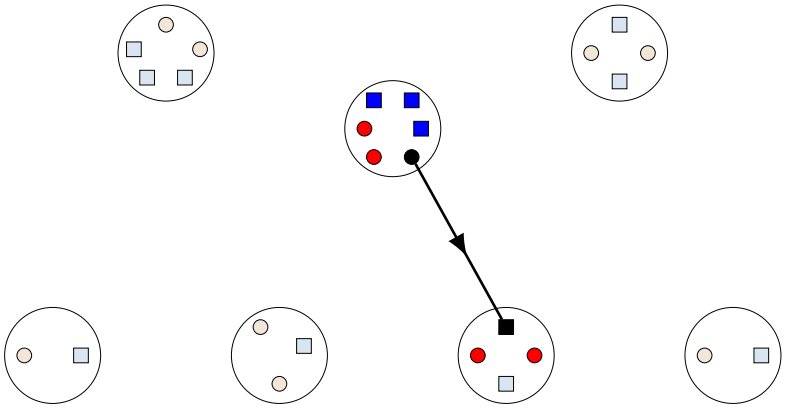
(B2) $r_t \in \mathcal{A}_{t-1}^+ \cup \mathcal{R}(\text{bin}(b_t))$ and $A_t^+ - \widehat{Y}_t^+ < \log^2(n)$.

(B3) $b_t \in \mathcal{A}_{t-1}^- \cup \mathcal{B}(\text{bin}(r_t))$ and $A_t^- - \widehat{Y}_t^- < \log^2(n)$.

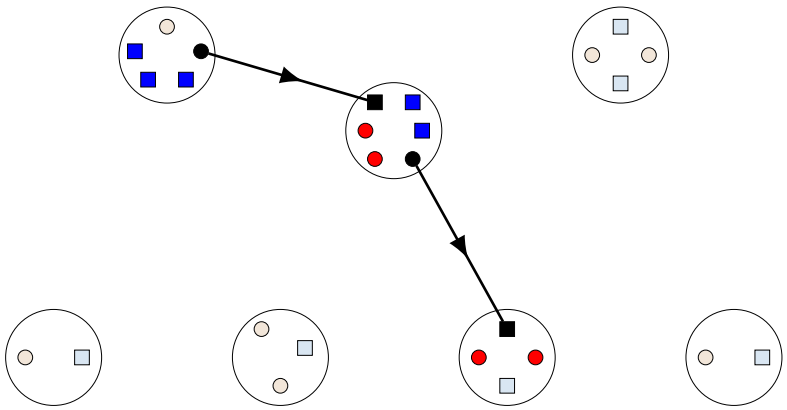
Figure 4.2 outlines one iteration of this procedure. The events (E), (B1), (B2), and (B3) are considered bad events for different reasons. The event (E) allows for the possibility that the fan-in or fan-out of the initial bin is too small to satisfy the conditions of Definition 4.13. Similarly, (B1) allows for the possibility that $\widehat{\Gamma}_i$ takes too many iterations to determine that the fan-in and fan-out of the initial bin is large. Events (B2) and (B3) allow for the possibility that $A_t^+ < \widehat{Y}_t^+$ and $A_t^- < \widehat{Y}_t^-$ respectively. This will prevent us from using Y_t^+ and Y_t^- to describe lower bounds for the size of A_t^+ and A_t^-, which is the approach we use in Section 4.2.3 as part of the proof of Theorem 4.1.ii.



(a) The start of subroutine $\widehat{\Gamma}$.



(b) Choosing the pair of one active red point (Step 2).



(c) Choosing the pair of one active blue point (Step 3).

Figure 4.2: One iteration of the supercritical subroutine procedure.

Let τ^* be the minimum $t > 0$ such that (B1), (B2), or (B3) occurs in iteration t. We define

$$\tau = \min\{\tau^*, \log^3(n)\}. \tag{4.7}$$

We now prove that $\{A_t^+\}_{0\le t<\tau}$ and $\{A_t^-\}_{0\le t<\tau}$ dominate $\{Y_t^+\}_{0\le t<\tau}$ and $\{Y_t^-\}_{0\le t<\tau}$.

Lemma 4.14. *Let τ be as defined in (4.5). Then $\{A_t^+\}_{0\le t<\tau}$ dominates $\{Y_t^+\}_{0\le t<\tau}$ and $\{A_t^-\}_{0\le t<\tau}$ dominates $\{Y_t^-\}_{0\le t<\tau}$.*

Proof: Since $\log^3(n) < \sigma n$ where σ is as in Lemma 4.8, $p_j^+(t) \ge \phi_t^+$ and $p_i^-(t) \ge \phi_i^-$ for all $i, j \ge 1$ and $0 \le t \le \tau$ by Lemma 4.8. Hence Γ is well-defined in this interval. Furthermore, the marginal distributions of $\widehat{Y}^+$ and $\widehat{Y}^-$ are the same as the distributions of Y^+ and Y^- respectively. Hence the procedure of Γ_i defines a coupling of A_t^+, A_t^-, Y_t^+, and Y_t^-.

From the procedure of Γ, each initial bin v of a subroutine $\widehat{\Gamma}_i$ starting in iteration t_0 of Γ is chosen so that $A_{t_0}^+ \ge \widehat{Y}_{t_0}^+ = 1$ and $A_{t_0}^- \ge \widehat{Y}_{t_0}^- = 1$. It therefore suffices to show $A_t^+ \ge \widehat{Y}_t^+$ and $A_t^- \ge \widehat{Y}_t^-$ for all iterations t in $\widehat{\Gamma}_i$.

Note that for $t > t_0$ in $\widehat{\Gamma}_i$, if $b_t \in \mathcal{B}(\mathcal{R}_j(t))$, then $A_t^+ - A_{t-1}^+ = j - 1$ unless $r_t \in \mathcal{A}_{t-1}^+$ or r_t is in $\text{bin}(b_t)$. Similarly, for $r_t \in \mathcal{R}(\mathcal{B}_i(t))$, $A_t^- - A_{t-1}^- = i - 1$ unless $b_t \in \mathcal{A}_{t-1}^-$ or b_t is in $\text{bin}(r_t)$. However, the choice of τ implies that for all $0 \le t < \tau$, $A_t^+ - \widehat{Y}_t^+ \ge \log^2(n)$ and $A_t^- - \widehat{Y}_t^- \ge \log^2(n)$ if $r_t \in \mathcal{A}_{t-1}^+ \cup \mathcal{R}(\text{bin}(b_t))$ and $b_t \in \mathcal{A}_{t-1}^- \cup \mathcal{B}(\text{bin}(r_t))$ respectively. Thus $A_t^+ > \widehat{Y}_t^+$ for all iterations $0 \le t < \tau$ such that $b_t \in \mathcal{B}(\mathcal{R}_j(t))$ and $A_t^+ - A_{t-1}^+ \ne j - 1$.

Similarly, $A_t^- > \widehat{Y}_t^-$ for all iterations $0 \le t < \tau$ such that $r_t \in \mathcal{R}(\mathcal{B}_i(t))$ and $A_t^- - A_{t-1}^- \neq i-1$.

For all iterations $0 \le t < \tau$ such that $b_t \in \mathcal{B}(\mathcal{R}_j(t))$ and $A_t^+ - A_{t-1}^+ = j-1$, the procedure of $\widehat{\Gamma}_i$ shows $\widehat{Y}_t^+ - \widehat{Y}_{t-1}^+ = j-1$ or $\widehat{Y}_t^+ - \widehat{Y}_{t-1}^+ = -1$. As $A_{t_0}^+ \ge \widehat{Y}_{t_0}^+ = 1$, this implies $A_t^+ \ge \widehat{Y}_t^+$ for all such iterations. Similarly, $\widehat{Y}_t^- - \widehat{Y}_{t-1}^- = i-1$ or $\widehat{Y}_t^- - \widehat{Y}_{t-1}^- = -1$ when $r_t \in \mathcal{R}(\mathcal{B}_i(t))$ and $A_t^- - A_{t-1}^- = i-1$. Hence $A_{t_0}^- \ge \widehat{Y}_{t_0}^- = 1$ implies $A_t^- \ge \widehat{Y}_t^-$ for all such iterations as well.

Therefore $A_t^+ \ge \widehat{Y}_t^+$ and $A_t^- \ge \widehat{Y}_t^-$ for all $0 \le t < \tau$. □

Thus the choice of $\{Y_t^+\}_{t\ge 0}$ and $\{Y_t^-\}_{t\ge 0}$ from Section 4.2.1 satisfies all three desired properties until iteration τ. In the next section, we will bound the probability that $\tau < \log^3(n)$, i.e. that a bad event occurs before $\log^3(n)$ iterations are performed. Furthermore, we show that the probability some $\widehat{\Gamma}_i$ of Γ finds a good bin v before τ is $1 - o(1)$.

4.2.3 Finding a Good Bin

We first prove the following proposition.

Proposition 4.15. *There exists a constant $c > 0$ such that* $\mathbf{Pr}\left(\tau \ge \left\lceil c\log^2(n)\right\rceil\right) = 1 - o(1)$.

We use this proposition to prove that a.a.s. some $\widehat{\Gamma}_i$ of Γ finds a good bin v.

To prove Proposition 4.15, it suffices to calculate the probability that (B1), (B2), or (B3) occurs before iteration $\left\lceil c\log^2(n)\right\rceil$ for some constant $c > 0$. We do so using the following lemmas.

Lemma 4.16. *There exists a constant $c > 0$ such that for $s = \lceil c\log(n)\rceil$,* $\mathbf{Pr}\left(Y_s^+ < \frac{c\epsilon}{8}\log(n)\right) = o(1)$ *and* $\mathbf{Pr}\left(Y_s^- < \frac{c\epsilon}{8}\log(n)\right) = o(1)$.

Proof: Recall $Y_t^+ = Y_{t-1}^+ + \eta_t^+$ for all $t > 0$ and $Y_0^+ = 1$. Thus $\mathbf{E}(Y_t^+) = 1 + [\mathbf{E}(\eta^+)]t$.

$$\begin{aligned}
\mathbf{E}(\eta^+) &= \sum_{j\ge 0}(j-1)\phi_j^+ \\
&= \left(\sum_{j=0}^{k} j\phi_j^+\right) - \sum_{j=0}^{k}\phi_j^+ \\
&\ge 1 + \frac{\epsilon}{4} - 1 \\
&= \frac{\epsilon}{4}.
\end{aligned}$$

Thus $\mathbf{E}(Y_t^+) \geq 1 + \frac{\epsilon}{4}t$ and so $\mathbf{E}(Y_s^+) \geq 1 + \frac{c\epsilon}{4}\log(n)$ and $\{Y_t^+ - \frac{\epsilon}{4}t\}_{0 \leq t \leq s}$ is a sub-martingale. Furthermore,

$$\left| \left[Y_{t+1}^+ - \frac{\epsilon}{4}(t+1) \right] - \left(Y_t^+ - \frac{\epsilon}{4}t \right) \right| \leq k$$

and so Corollary 3.10 implies

$$\begin{aligned} \mathbf{Pr}\left(Y_s^+ < \left\lceil \frac{c\epsilon}{8}\log(n) \right\rceil \right) &\leq \mathbf{Pr}\left[\left(Y_s^+ - \frac{c\epsilon}{4}\log(n) \right) - 1 \leq -\frac{c\epsilon}{8}\log(n) \right] \\ &\leq e^{-c^2\epsilon^2 \log^2(n)/128k^2 c\log(n)}. \end{aligned}$$

For $c = \frac{128k^2}{\epsilon^2}$, this probability is $O\left(\frac{1}{n}\right)$. A similar calculation shows the same result for Y_s^-. □

Corollary 4.17. *The probability (B1) occurs in iteration $\tau < \log^3(n)$ is $o(1)$.*

Proof: Let $\widehat{\Gamma}$ be an iteration of Γ that starts in iteration t_0 and stops because (B1) occurs. Thus (E), (B2), and (B3) do not occur during iterations $t_0 \leq t \leq t_0 + \lfloor \log^2(n) \rfloor$. Hence for a constant c and corresponding s as in Lemma 4.16, let (A) denote the event (E), (B2), and (B3) do not occur during iterations $t_0 \leq t \leq t_0 + s$. Then

$$\begin{aligned} \mathbf{Pr}(\text{(B1) occurs}) &\leq \mathbf{Pr}\left(\text{(A) and either } \widehat{Y}_s^+ < \frac{c\epsilon}{8}\log(n) \text{ or } \widehat{Y}_s^- < \frac{c\epsilon}{4}\log(n) \right) \\ &\leq \mathbf{Pr}\left(\text{(A) and } \widehat{Y}_s^+ < \frac{c\epsilon}{8}\log(n) \right) + \mathbf{Pr}\left(\text{(A) and } \widehat{Y}_s^- < \frac{c\epsilon}{8}\log(n) \right) \\ &\leq \mathbf{Pr}\left(Y_s^+ < \frac{c\epsilon}{8}\log(n) \right) + \mathbf{Pr}\left(Y_s^- < \frac{c\epsilon}{8}\log(n) \right) \\ &= o(1) + o(1). \end{aligned}$$

Hence the probability (B1) occurs is $o(1)$. □

Lemma 4.18. *For the same constant $c > 0$ as Lemma 4.16,*

(i) $\mathbf{Pr}($*(B2) does not occur for all* $0 \leq t \leq \lceil c\log^2(n) \rceil) = 1 - o(1)$

(ii) $\mathbf{Pr}($*(B3) does not occur for all* $0 \leq t \leq \lceil c\log^2(n) \rceil) = 1 - o(1)$.

Proof: (i) Recall that the probability (B2) occurs in iteration t is the probability of choosing

r_t such that $r_t \in (\mathcal{A}_{t-1}^+ \cup \mathcal{R}(u(b_t)))\backslash\{r_t\}$ and $A_t^+ - \widehat{Y}_t^+ < \log^2(n)$. Thus,

$$\begin{aligned}\mathbf{Pr}(\text{(B2) occurs in iteration } t) &\leq \frac{(A_{t-1}^+ - 1) + R(u(b_t))}{\theta n - 2(t-1) - 1} \\ &\leq \frac{k(t-1) - 1 + k + \log^2(n)}{\theta n - 2t + 1} \\ &< \frac{kt + \log^2(n)}{\theta n - 2t + 1}.\end{aligned}$$

For any $t < \log^3(n)$ this probability is $O\left(\frac{\log^3(n)}{n}\right)$. Hence
$\mathbf{Pr}(\text{(B2) does not occur for all } 0 \leq t \leq \lceil c\log^2(n)\rceil) = 1 - O\left(\frac{\log^6(n)}{n}\right)$.

(ii) Recall that the probability (B3) occurs in iteration t is the probability of choosing b_t such that $b_t \in \mathcal{A}_{t-1}^- \cup \mathcal{B}(u(r_t))$ and $A_t^- - \widehat{Y}_t^- < \log^2(n)$. Thus,

$$\begin{aligned}\mathbf{Pr}(\text{(B3) occurs in iteration } t) &= \frac{A_{t-1}^- + B(u(r_t))}{\theta n - 2(t-1)} \\ &\leq \frac{k(t-1) + k + \log^2(n)}{\theta n - 2t - 2}.\end{aligned}$$

For any $t < \log^3(n)$ this probability is $O\left(\frac{\log^3(n)}{n}\right)$. Hence
$\mathbf{Pr}(\text{(B3) does not occur for all } 0 \leq t \leq \lceil c\log^2(n)\rceil) = 1 - O\left(\frac{\log^6(n)}{n}\right)$. □

We now prove Proposition 4.15.

Proof of 4.15: Let $c = \frac{128k^2}{\epsilon^2}$. Then by Corollary 4.17 and Lemma 4.18,

$$\mathbf{Pr}\left(\tau \geq \left\lceil c\log^2(n)\right\rceil\right) \geq 1 - [o(1) + o(1) + o(1)] = 1 - o(1).$$

Hence the statement holds. □

We now use Proposition 4.15 to prove the following key proposition.

Proposition 4.19. $\mathbf{Pr}(\Gamma$ *contains a subroutine that finds a good bin*$) = 1 - o(1)$.

We begin the proof of this proposition with the following lemma.

Lemma 4.20. *Let $c > 0$ be the same as in Proposition 4.15 and $s = \lceil c \log(n) \rceil$. Then* $\mathbf{Pr}\left(Y_t^+, Y_t^- > 0 \text{ for all } 0 \le t \le s \text{ and } Y_s^+, Y_s^- \ge \frac{c\epsilon}{8}\log(n)\right) > \alpha$ *for some* $\alpha > 0$.

Proof: Let ρ^+ and ρ^- be the probability a Galton-Watson process with distribution $\eta^+ + 1$, respectively $\eta^- + 1$, is finite. The proof requires the following two claims.

Claim 1. $\mathbf{Pr}(Y_t^+ > 0 \text{ for all } 0 \le t \le \lceil c \log(n) \rceil) \ge 1 - \rho^+$.

Proof: Let $\{X_t\}_{t \ge 0}$ be a Galton-Watson process with distribution $\eta^+ + 1$. Recall that the distribution of $\{Y_t^+\}_{t \ge 0}$ is also $\eta^+ + 1$. Hence X_t can be coupled with Y_t^+ so that $X_t = Y_t^+$ for all t in $\widehat{\Gamma}$ until $Y_t^+ = 0$. Thus

$$\mathbf{Pr}(Y_t^+ > 0 \text{ for all } 0 \le t \le \lceil c \log(n) \rceil) \ge \mathbf{Pr}(\{X^+\}_{t \ge 0} \text{ is infinite}),$$

which is defined to be $1 - \rho^+$. □

Claim 2. $\mathbf{Pr}(Y_t^- > 0 \text{ for all } 0 \le t \le \lceil c \log(n) \rceil) \ge 1 - \rho^-$.

Proof: The proof is identical to that of Claim 1 with η^-, Y_t^-, and ρ^- replacing η^+, Y_t^+, and ρ^+ respectively. □

Recall from Lemma 4.17, $\mathbf{Pr}\left(Y_s^+ < \frac{c\epsilon}{8}\log(n)\right) = o(1)$ and $\mathbf{Pr}\left(Y_s^- < \frac{c\epsilon}{8}\log(n)\right) = o(1)$ for $s = \lceil c \log(n) \rceil$. Hence,

$$\begin{aligned}
&\mathbf{Pr}\left(Y_t^+, Y_t^- > 0 \text{ for all } 0 \le t \le s \text{ and } Y_s^+, Y_s^- \ge \frac{c\epsilon}{8}\log(n)\right) \\
&\ge \mathbf{Pr}(Y_t^+, Y_t^- > 0 \text{ for all } 0 \le t \le s) - \mathbf{Pr}\left(Y_s^+ < \frac{c\epsilon}{8}\log(n)\right) - \mathbf{Pr}\left(Y_s^- < \frac{c\epsilon}{8}\log(n)\right) \\
&\ge \left(\mathbf{Pr}(Y_t^+ > 0 \text{ for all } 0 \le t \le s)\right)\left(\mathbf{Pr}(Y_t^- > 0 \text{ for all } 0 \le t \le s)\right) - o(1) \\
&\ge (1 - \rho^+)(1 - \rho^-) - o(1).
\end{aligned}$$

As $\mathbf{E}(\eta^+ + 1) = \mathbf{E}(\eta^- + 1) \ge \frac{\epsilon}{4} + 1 > 1$, by Proposition 3.3.ii, $0 < \rho^+ < 1$ and $0 < \rho^- < 1$. Hence $\alpha = \frac{1}{2}(1 - \rho^+)(1 - \rho^-) > 0$ suffices. □

We can now prove Proposition 4.19.

Proof of 4.19: Recall that a subroutine $\widehat{\Gamma}$ stops only when (E) or one of (B1), (B2), and (B3) occurs. Furthermore, $\tau \le \log^3(n)$ and (B1) occurs if $\widehat{\Gamma}$ performs $\log^2(n)$ iterations without finding a good bin. Thus, if Γ does not contain a subroutine that finds a good bin, at least $\log(n)$ subroutines are completed before Γ aborts in iteration τ.

Define a subroutine $\widehat{\Gamma}$ to be a *success* if $\widehat{\Gamma}$ finds a good bin. Let S be the number of successful subroutines in the first $\log(n)$ subroutines in Γ.

$$\begin{aligned}\mathbf{Pr}(S=0) &\leq \mathbf{Pr}(\text{(B1), (B2), or (B3) occurs before iteration } \log^3(n)) \\ &\quad + \mathbf{Pr}(\text{(E) occurs } \log(n) \text{ times}) \\ &\leq o(1) + \left[1 - \mathbf{Pr}\left(Y_t^+, Y_t^- > 0 \text{ for all } 0 \leq t \leq s \text{ and } Y_s^+, Y_s^- \geq \frac{c\epsilon}{8}\log(n)\right)\right]^{\log(n)} \\ &< o(1) + [1-\alpha]^{\log(n)}.\end{aligned}$$

Thus a.a.s. Γ contains a subroutine that finds a good bin before τ. □

Proposition 4.20 implies that Γ finds a good bin v before $O(\log^3(n))$ pairs of C are exposed. It remains to show that v is in a giant SCC of C.

4.2.4 Proof of Theorem 4.1(ii)

In this section, we prove Theorem 4.1(ii). We do so by proving that with high probability, a good bin v is in a giant SCC. This is accomplished by showing that a.a.s. there are a linear number of pairs (r, b) with $r \in \mathcal{R}(\mathcal{F}^+(v))$ and $b \in \mathcal{B}(\mathcal{F}^-(v))$ and these pairs are in the same SCC. We then apply the following key lemma.

Lemma 4.21. *If $K \in \mathcal{K}(C)$ contains a linear number of pairs, then $|V(K)| = \Omega(n)$.*

Proof. Let $K \in \mathcal{K}(C)$ be a SCC with δn pairs for some $\delta > 0$. Since D_n is well-behaved and $C \in \mathcal{C}(D_n)$, there exist integers k and N such that for all $n > N$, $\left|\sum_{i=0}^{k}\sum_{j=0}^{k} ij\frac{n_{ij}}{\theta n} - \sum_{i\geq 0}\sum_{j\geq 0} ij\frac{n_{ij}}{\theta n}\right| < \frac{\delta\theta}{4}$ by uniform convergence. Thus,

$$\begin{aligned}\frac{\delta\theta}{4} &> \left|\sum_{i=0}^{k}\sum_{j=0}^{k} ij\frac{n_{ij}}{\theta n} - \sum_{i\geq 0}\sum_{j\geq 0} ij\frac{n_{ij}}{\theta n}\right| \\ &= \left|\sum_{i=0}^{k}\sum_{j=0}^{k} ij\frac{n_{ij}}{\theta n} - \left(\sum_{i=0}^{k}\sum_{j=0}^{k} ij\frac{n_{ij}}{\theta n} + \sum_{i\geq 0}\sum_{j>k} ij\frac{n_{ij}}{\theta n} + \sum_{i>k}\sum_{j=0}^{k} ij\frac{n_{ij}}{\theta n}\right)\right| \\ &= \sum_{i\geq 0}\sum_{j>k} ij\frac{n_{ij}}{\theta n} + \sum_{i>k}\sum_{j=0}^{k} ij\frac{n_{ij}}{\theta n},\end{aligned}$$

which implies $\frac{\delta}{4}n > \sum_{i\geq 0}\sum_{j>k} ijn_{ij} + \sum_{i>k}\sum_{j=0}^{k} ijn_{ij}$. Hence

$$R_{>k} = \sum_{i\geq 1}\sum_{j>k} jn_{ij} \leq \sum_{i\geq 0}\sum_{j>k} ijn_{ij} + \sum_{i>k}\sum_{j=0}^{k} ijn_{ij} < \frac{\delta}{4}n.$$

By symmetry, $B_{>k} = \sum_{i>k}\sum_{j\geq 1} in_{ij} < \frac{\delta}{4}n$.

Since K is strongly connected, each bin in K contains at least one red and at least one blue point. Thus for all $u, v \in V(K)$, if $B(u) > k$ then $\mathcal{B}(u) \subseteq \mathcal{B}_{>k}$ and if $R(v) > k$ then $\mathcal{R}(v) \subseteq \mathcal{R}_{>k}$. As $B_{>k} + R_{>k} < \frac{\delta}{2}n$, there exists at least $\frac{\delta}{2}n$ pairs in K whose red points are not in $\mathcal{R}_{>k}$ and blue points are not in $\mathcal{B}_{>k}$. This implies $|V(K)| \geq \frac{\delta}{2k}n$, as desired. □

Let v be a good bin and $\widehat{\Gamma}$ be the subroutine that starts at v in iteration t_0. Define $t_1 = t_0 + \lceil c\log(n)\rceil$ where c is the same constant as in Proposition 4.16. Instead of continuing $\widehat{\Gamma}$ in iteration t_1+1, we will explore $\mathcal{F}^+(v)$ and $\mathcal{F}^-(v)$ separately using two new procedures $\widehat{\Gamma}^+$ and $\widehat{\Gamma}^-$. This will make it easier for us to prove there is a linear number of pairs that contain a red point in $\mathcal{R}(\mathcal{F}^+(v))$ and blue point in $\mathcal{B}(\mathcal{F}^-(v))$.

We begin by performing the exploration procedure $\widehat{\Gamma}^+$ described below. This procedure will show that with high probability, within the first $\frac{\sigma}{4}n$ iterations of $\widehat{\Gamma}^+$, there will be an iteration that contains a linear number of active red points. First, $\widehat{\Gamma}^+$ couples A_t^+ with a new random variable X_t^+. X_t^+ is defined such that $X_{t_1}^+ = A_{t_1}^+$ and for all $t > t_1$, $X_t^+ = X_{t-1}^+ + j - 1$ with probability $p_j^+(t)$ for all $1 \leq j \leq k$ and $X_t^+ = X_{t-1}^+ - 1$ with the remaining probability.

In each iteration $t > t_1$ of $\widehat{\Gamma}^+$, the following procedure is performed.

1. Choose a red point $\mathbf{r}_t \in \mathcal{A}_{t-1}^+$.
2. One of the following is performed with the corresponding probability:
   (a) For each $1 \leq j \leq k$, with probability $p_j^+(t)$ let $\widehat{X}_t^+ = \widehat{X}_{t-1}^+ + j - 1$ and choose b_t uniformly at random from $\mathcal{B}(\mathcal{R}_j(t))\backslash\mathcal{U}_{t-1}^-$. Pair $\mathbf{r}_t$ with b_t.
   (b) Otherwise, let $\widehat{X}_t^+ = \widehat{X}_{t-1}^+ - 1$ and choose b_t uniformly at random from $[\mathcal{B}(\mathcal{R}_{>k}(t)) \cup \mathcal{B}(\mathcal{R}_0(t))]\backslash\mathcal{U}_{t-1}^-$. Pair $\mathbf{r}_t$ with b_t.
3. Change the states of $\mathbf{r}_t$ and b_t to used and the states of all sleeping red points in $\text{bin}(b_t)$ to active.

Note that this procedure exposes exactly one pair in C during each iteration and this pair is in $\mathcal{F}^+(v)$. Furthermore, $A_t^+ \geq \widehat{X}_t^+$ for all $t \geq t_1$.

We stop $\widehat{\Gamma}^+$ in iteration τ^+ where $\tau^+ = \min\left\{t > t_1 \mid A_t^+ = 0 \text{ or } A_t^+ = \left\lceil \frac{\epsilon\sigma}{32} n \right\rceil\right\}$. We then have the following lemmas about τ^+.

Lemma 4.22. *Let σ be as in Lemma 4.8.* $\mathbf{Pr}(X_t^+ > 0$ *for all* $t_1 < t \leq \sigma n) = 1 - o(1)$.

Proof: Note that the coupling of A_t^+ with Y_t^+ in Γ implies $A_{t_1}^+ \geq \widehat{Y}_{t_1}^+ \geq \frac{c\epsilon}{8}\log(n)$ where c is the constant from Proposition 4.15. Furthermore, $X_t^+ - X_{t-1}^+ \geq -1$ for all $t_1 < t$. Hence for all $t_1 < t < t_1 + \left\lceil \frac{c\epsilon}{8}\log(n) \right\rceil$, $\mathbf{Pr}(X_t^+ = 0) = 0$.

Let $\tau^* = \min\{t > t_1 \mid X_t^+ = 0\}$. Recall that for all $t_1 \leq t \leq \sigma n$,

$$\mathbf{E}(X_{t+1}^+ - X_t^+) = \left(\sum_{j=0}^{k} j p_j^+(t)\right) - 1 \geq \left(\sum_{j=0}^{k} j\phi_j^+\right) - 1 > \frac{\epsilon}{4}.$$

Thus, $\{X_t^+ - \frac{\epsilon}{4}(t - t_1)\}_{t_1 \leq t \leq \tau^*}$ is a sub-martingale. Hence by Corollary 3.10, for $t \geq t_1 + \left\lceil \frac{c\epsilon}{8}\log(n) \right\rceil$,

$$\begin{aligned}
\mathbf{Pr}(X_t^+ = 0) &\leq \mathbf{Pr}\left(\left[X_t^+ - \frac{\epsilon}{4}(t - t_1)\right] - X_{t_1}^+ \leq -\left[X_{t_1}^+ + \frac{\epsilon}{4}(t - t_1)\right]\right) \\
&\leq e^{-[4X_{t_1}^+ + \epsilon(t-t_1)]^2/32k^2(t-t_1)} \\
&= e^{-(X_{t_1}^+)^2/2k^2(t-t_1)} \cdot e^{-X_{t_1}^+\epsilon/2k^2} \cdot e^{-\epsilon^2(t-t_1)/32k^2} \\
&\leq e^{-c\epsilon^3\log(n)/256k^2} \\
&= n^{-c\epsilon^3/256k^2}.
\end{aligned}$$

Thus for $t^* = t_1 + \left\lceil \frac{c\epsilon}{8}\log(n) \right\rceil$,

$$\begin{aligned}
\mathbf{Pr}(X_t^+ > 0 \text{ for all } t_1 \leq t \leq \sigma n) &= 1 - \sum_{t=t_1}^{\sigma n} \mathbf{Pr}\left(X_t^+ = 0\right) \\
&= 1 - \sum_{t=t^*}^{\sigma n} \mathbf{Pr}\left(X_t^+ = 0\right) \\
&\geq 1 - (\theta n - t^*) n^{-c\epsilon^3/256k^2}.
\end{aligned}$$

Note that for $c = \max\{\frac{128k^2}{\epsilon^2}, \frac{256k^2}{\epsilon^3}\}$, all the results in Section 4.2.3 hold. Furthermore, $1 - (\theta n - t^*)n^{-c\epsilon^3/256k^2} = 1 - o(1)$ for this choice of c, and so the statement holds. □

Lemma 4.23. *Let* $t_2 = t_1 + \lfloor \frac{\sigma}{4} n \rfloor$. $\mathbf{Pr}(A_t^+ \geq \lceil \frac{\epsilon\sigma}{32} n \rceil$ *for some* $t_1 \leq t \leq t_2) = 1 - o(1)$.

Proof: By Lemma 4.22, with high probability $A_t^+ \geq \widehat{X}_t^+ > 0$ for all $t_1 < t \leq t_2$ and so $\widehat{\Gamma}^+$ stops in iteration $t < t_2$ only if $A_t^+ \geq \lceil \frac{\epsilon\sigma}{32} n \rceil$. Also, $\mathbf{E}(A_t^+ - A_{t-1}^+) \geq \mathbf{E}(X_t^+ - X_{t-1}^+) > \frac{\epsilon}{4}$ for all $t_1 < t \leq t_2$. Thus $\mathbf{E}(X_{t_2}^+) > \frac{\epsilon}{4}(t_2 - t_1)$ and so by Corollary 3.10,

$$\begin{aligned}
\mathbf{Pr}\left(X_{t_2}^+ < \left\lceil \frac{\epsilon\sigma}{32} n \right\rceil\right) &= \mathbf{Pr}\left(\left[X_{t_2}^+ - \frac{\epsilon}{4}(t_2 - t_1)\right] - X_{t_1}^+ \leq -\left[X_{t_1}^+ + \frac{\epsilon}{8}(t_2 - t_1)\right]\right) \\
&\leq e^{-[8X_{t_1}^+ + \epsilon(t_2 - t_1)]^2 / 128k^2(t_2 - t_1)} \\
&= e^{-(X_{t_1}^+)^2/2k^2(t_2-t_1)} \cdot e^{-\epsilon X_{t_1}^+/4k^2} \cdot e^{-\epsilon^2(t_2-t_1)/128k^2} \\
&\leq e^{-\epsilon^2\sigma n/512k^2}.
\end{aligned}$$

Hence $\mathbf{Pr}\left(X_{t_2}^+ \geq \lceil \frac{\epsilon\sigma}{32} n \rceil\right) = 1 - o(1)$ and so
$\mathbf{Pr}(A_t^+ \geq \lceil \frac{\epsilon\sigma}{32} n \rceil$ for some $t_1 \leq t \leq t_2) = 1 - o(1)$. □

As $A_t^+ \geq \widehat{X}_t^+$ for all $t \geq t_1$, Lemmas 4.22 and 4.23 imply that a.a.s. $\tau^+ < t_1 + \frac{\sigma}{4} n$ and $\widehat{\Gamma}^+$ stops because $A_{\tau^+}^+ = \lceil \frac{\epsilon\sigma}{32} n \rceil$. We then begin the exploration procedure $\widehat{\Gamma}^-$ in iteration $\tau^+ + 1$. This new procedure is essentially the same as $\widehat{\Gamma}^+$, but uses the appropriate sets for exposing the fan-in rather than the fan-out of a bin (i.e. U_{t-1}^+ instead of U_{t-1}^- and $\mathcal{R}(\mathcal{B}_i(t))$ instead of $\mathcal{B}(\mathcal{R}_j(t))$). It will also be used to show that with high probability, within the first $\frac{\sigma}{4} n$ iterations of $\widehat{\Gamma}^-$, there will be an iteration that contains a linear number of active blue points.

First, let X_t^- be a new random variable such that $X_{\tau^+}^- = A_{\tau^+}^-$ and for all $t > \tau^+$, $X_t^- = X_{t-1}^- + i - 1$ with probability $p_i^-(t)$ for all $1 \leq i \leq k$ and $X_t^- = X_{t-1}^- - 1$ with the remaining probability. $\widehat{\Gamma}^-$ is defined to be the same procedure as $\widehat{\Gamma}^+$, but using X_t^-, $\mathbf{b}_t$, $\mathcal{A}_{t-1}^-$, $p_i^-(t)$, r_t, $\mathcal{R}(\mathcal{B}_i(t))$, $\mathcal{U}_{t-1}^+$, and $[\mathcal{B}(\mathcal{R}_{>k}(t)) \cup \mathcal{B}(\mathcal{R}_0(t))]$ instead of X_t^+, $\mathbf{r}_t$, $\mathcal{A}_{t-1}^+$, $p_j^+(t)$, b_t, $\mathcal{B}(\mathcal{R}_j(t))$, $\mathcal{U}_{t-1}^-$, and $[\mathcal{R}(\mathcal{B}_{>k}(t)) \cup \mathcal{R}(\mathcal{B}_0(t))]$ respectively.

As before, $\widehat{\Gamma}^-$ stops in iteration τ^- where $\tau^- = \min\left\{t > \tau^+ \mid A_t^- = 0 \text{ or } A_t^- = \lceil \frac{\epsilon\sigma}{32} n \rceil\right\}$. The following lemmas about τ^- are similar to the results of Lemmas 4.22 and 4.23.

Lemma 4.24. *Let* σ *be as in Lemma 4.8.* $\mathbf{Pr}(X_t^- > 0$ *for all* $\tau^+ < t \leq \sigma n) = 1 - o(1)$.

Proof: The proof is essentially the same as that of Lemma 4.22 with the assumption that $A_{\tau^+}^- \geq \frac{c\epsilon}{8} \log(n)$. We therefore only prove this assumption.

Note that the coupling of A_t^- with Y_t^- in Γ implies $A_{t_1}^- \geq Y_{t_1}^- \geq \frac{c\epsilon}{8}\log(n)$ where c is the constant as Proposition 4.16. Also, the procedure of $\widehat{\Gamma}^+$ implies $\mathcal{A}_{\tau^+}^- = \mathcal{A}_{t_1}^- \backslash \{b_t \mid t_1 < t \leq \tau^+\}$. Furthermore, for $t_1 < t \leq \tau^+$,

$$\begin{aligned}\mathbf{Pr}\left(b_t \in \mathcal{A}_{t_1}^- \text{ and } A_{t-1}^- = \left\lceil \frac{c\epsilon}{8}\log(n)\right\rceil\right) &\leq \frac{\left\lceil \frac{c\epsilon}{8}\log(n)\right\rceil}{\theta n - \tau^+} \\ &< \frac{c\epsilon\log(n)}{8\theta n - 4\sigma n}\end{aligned}$$

Hence with high probability, $\widehat{\Gamma}^+$ does not pair an active red point with an active blue point in iteration t when there are only $\left\lceil \frac{c\epsilon}{8}\log(n)\right\rceil$ active blue points. Thus with high probability $A_{\tau^+}^- \geq \left\lceil \frac{c\epsilon}{8}\log(n)\right\rceil$. The remainder of the proof follows the same form as Lemma 4.22. □

Lemma 4.25. *Let $t_3 = \tau^+ + \left\lfloor \frac{\sigma}{4}n\right\rfloor$. $\mathbf{Pr}(A_t^- \geq \left\lceil \frac{\epsilon\sigma}{32}n\right\rceil$ for some $\tau^+ \leq t \leq t_3) = 1 - o(1)$.*

Proof: The proof is essentially the same as that of Lemma 4.23, using Lemma 4.24 instead of Lemma 4.22. □

Recall that we aim to find a linear number of pairs (r, b) such that $r \in \mathcal{R}(\mathcal{F}^+(v))$ and $b \in \mathcal{B}(\mathcal{F}^-(v))$. We do so by showing there exists a linear number of pairs (r, b) such that $r \in \mathcal{A}_{\tau^+}^+$ and $b \in \mathcal{A}_t^-$ for some $\tau^+ < t \leq \tau^-$. Let $\mathcal{P}$ be the set of these pairs, i.e.

$$\mathcal{P} = \{(r, b) \mid r \in \mathcal{A}_{\tau^+}^+, b \in \mathcal{A}_t^- \text{ for some } \tau^+ < t \leq \tau^-\},$$

and let $P = |\mathcal{P}|$.

Note that if in a linear number of iterations $\widehat{\Gamma}^-$ pairs an active blue point with an active red point, then $P = \Omega(n)$ and we are done. Thus, we may assume $\widehat{\Gamma}^-$ does not do so, hence $A_{\tau^-}^+ > \frac{1}{4}A_{\tau^+}^+ \geq \left\lceil \frac{\epsilon\sigma}{128}n\right\rceil$.

Let $\mathcal{A}'$ be an arbitrary subset of $\mathcal{A}_{\tau^-}^+$ of size $\left\lceil \frac{\epsilon\sigma}{128}n\right\rceil$. We have the following lemma.

Lemma 4.26. *Suppose the pairs containing the points of $\mathcal{A}'$ are exposed in some order. Then the probability a given point $r \in \mathcal{A}'$ is paired with a point $b \in \mathcal{A}_{\tau^-}^-$ is at least $\frac{\epsilon\sigma}{64\theta}$.*

Proof: Let $t^* = \tau^- + \left\lceil \frac{\epsilon\sigma}{128}n\right\rceil$. Note that the probability of pairing a point $r \in \mathcal{A}'$ with a

point $b \in \mathcal{A}^-_{\tau^-}$ is at least the probability $b_{t^*} \in \mathcal{A}^-_{\tau^-}$ given $b_t \in \mathcal{A}^-_{\tau^-}$ for all $\tau^- < t < t^*$.

$$\begin{aligned}\mathbf{Pr}(b_{t^*} \in \mathcal{A}^-_{\tau^-}) &\geq \frac{A^-_{\tau^-} - 2(t^* - \tau^- - 1)}{\theta n - 2(t^* - 1)} \\ &\geq \frac{\frac{\epsilon\sigma}{32}n - 2\left(\frac{\epsilon\sigma}{128}n\right)}{\theta n - \left(\frac{\epsilon\sigma}{64}n\right)} \\ &> \frac{2\epsilon\sigma n - \epsilon\sigma n}{64\theta n}.\end{aligned}$$

Thus for each $r \in \mathcal{A}'$, r is in a pair with some $b \in \mathcal{A}^-_{\tau^-}$ with probability at least $\frac{\epsilon\sigma}{64\theta}$. □

This leads to the following result for P.

Lemma 4.27. $\mathbf{E}(P) > \left\lfloor \frac{\epsilon^2\sigma^2}{8192\theta}n \right\rfloor$ *and* $\mathbf{Pr}\left(P \leq \frac{\epsilon^2\sigma^2}{4096\theta}n\right) = o(1)$.

Proof: Using Lemma 4.26,

$$\mathbf{E}(P) > \left(\left\lfloor \frac{\epsilon\sigma}{128}n \right\rfloor\right)\left(\frac{\epsilon\sigma}{64\theta}\right) = \left\lfloor \frac{\epsilon^2\sigma^2}{8192\theta}n \right\rfloor.$$

We define a sequence of trials $X_1, X_2, \ldots, X_{\epsilon\sigma n/128}$ such that $X_i = 1$ if the pair of the i^{th} point in $\mathcal{A}'$ is in $\mathcal{P}$ and $X_i = 0$ otherwise. Let X be the random variable $X = \sum_{i=1}^{\epsilon\sigma n/128} X_i$. Clearly for each i, $\max|\mathbf{E}(X \mid X_1, \ldots, X_{i+1}) - \mathbf{E}(X \mid X_1, \ldots, X_i)| \leq 1$. Thus by the Azuma-Hoeffding inequality,

$$\begin{aligned}\mathbf{Pr}\left[|X - \mathbf{E}(X)| > \left(\frac{\epsilon^2\sigma^2}{16384\theta}n\right)\right] &\leq 2\exp\left(-\frac{\epsilon^4\sigma^4n^2}{(16384)^2\theta^2}/2\frac{\epsilon\sigma n}{128}\right) \\ &\leq e^{-\epsilon^3\sigma^3 n/(2048\theta)^2}.\end{aligned}$$

As $P \geq X$, this implies $P \geq \frac{\epsilon^2\sigma^2}{16384\theta}n$ a.a.s. □

By Lemma 4.27, with high probability $P = \Omega(n)$. To apply Lemma 4.22, it remains to show that all pairs in $\mathcal{P}$ are in the same SCC.

Lemma 4.28. *All pairs in $\mathcal{P}$ are in $K(v) \in \mathcal{K}(C)$.*

Proof: Let $\mathcal{B}(\mathcal{P})$ be the set of blue points that are in pairs in $\mathcal{P}$ and let $\mathcal{R}(\mathcal{P})$ be the set of red points in pairs in $\mathcal{P}$. Define $V(\mathcal{B}(\mathcal{P}))$ to be the set of bins that contain a point in $\mathcal{B}(\mathcal{P})$, i.e. $V(\mathcal{B}(\mathcal{P})) = \{v \in V(C) \mid v = \text{bin}(b), b \in \mathcal{B}(\mathcal{P})\}$. Similarly, let $V(\mathcal{R}(\mathcal{P})) = \{v \in V(C) \mid v = \text{bin}(r), r \in \mathcal{R}(\mathcal{P})\}$. By Lemma 2.9, it suffices to show that for all $u \in V(\mathcal{B}(\mathcal{P})) \cup V(\mathcal{R}(\mathcal{P}))$, there exists a directed uv-path and directed vu-path in C.

By the exploration procedures, it is clear that for all $u \in V(\mathcal{B}(\mathcal{P}))$ and $w \in V(\mathcal{R}(\mathcal{P}))$, there exists a uv-path and vw-path in C. It remains to show for all $u \in V(\mathcal{B}(\mathcal{P}))\backslash V(\mathcal{R}(\mathcal{P}))$ and $w \in V(\mathcal{R}(\mathcal{P}))\backslash V(\mathcal{B}(\mathcal{P}))$, there exists a vu-path and wv-path in C.

By the definition of $V(\mathcal{B}(\mathcal{P}))$, there exists a point $b_u \in \mathcal{B}(u)$ such that (r_u, b_u) is a pair in $\mathcal{P}$ for some $r_u \in \mathcal{R}$. Thus $r_u \in \mathcal{R}(\mathcal{P})$ and so there is exists a $v\text{bin}(r_u)$-path Q in C. Hence Qu is a vu-path in C. Similarly, there exists a point $r_w \in \mathcal{R}(w)$ such that (r_w, b_w) is a pair in $\mathcal{P}$ for some $b_w \in \mathcal{B}$. Hence $b_w \in \mathcal{B}(\mathcal{P})$ and so there is exists a $\text{bin}(b_w)v$-path Q' in C. Hence wQ' is a wv-path in C.

Hence by Lemma 2.9, $u \in V(K(v))$ for all $u \in V(\mathcal{B}(\mathcal{P})) \cup V(\mathcal{R}(\mathcal{P}))$. Thus all pairs in $\mathcal{P}$ are in $K(v)$.

□

We can now prove that $K(v)$ is a giant SCC.

Proposition 4.29. *Let $v \in V(C)$ be a good bin. Then a.a.s. $K(v)$ is a giant SCC.*

Proof. By Lemma 4.27, with high probability there exists $\Omega(n)$ pairs (r, b) such that $r \in \mathcal{R}(\mathcal{F}^+(v))$ and $b \in \mathcal{B}(\mathcal{F}^-(v))$ and all of these pairs are in $K(v)$ by Lemma 4.28. Hence by Lemma 4.21, $|V(K(v))| = \Omega(n)$ and so $K(v)$ is giant. □

Propositions 4.19 and 4.29 immediately imply that $C \in \mathcal{C}(D_n)$ a.a.s. contains a giant SCC. Thus applying Proposition 3.17 completes the proof of Theorem 4.1(ii).

In Chapter 6, we will explain how Theorem 4.1 can be applied to predict a new site percolation threshold result. First, we will discuss some previous results and techniques from percolation theory in Chapter 5.

Chapter 5

Percolation on Directed Graphs

The *percolation model* is a simple stochastic model that was first introduced by Broadbent and Hammersley in 1957 [10]. It can be described using a plane square lattice and number p such that $0 < p < 1$. Each edge of the lattice is examined in turn and declared *open* with probability p and *closed* with probability $1 - p$, independently of all other edges. A path between two boundaries of some finite subsection of the lattice such that the path only contains open edges is called an *open path*.

The theory surrounding the use of percolation models is called *percolation theory* and has been used to study random physical processes, such as fluid flow through disordered media. Percolation has been studied by physicists and mathematicians alike as it is easy to formulate and its qualitative predictions are fairly realistic. It can also be used to derive results for more complex systems, such as inhomogeneous models and models where the states of different edges are not independent (sometimes called *dependent percolation*).

Percolation models are often used to study behaviors of systems that drastically change with respect to some natural parameter. This typically can be described as a change in the component structure of random subgraphs of graphs due to some change in an aspect of the graph. For example, the existence of an open path through a lattice is more likely when edges are likely to be open ($p \approx 1$) than when they are likely to be closed ($p \approx 0$). In fact, there exists a certain threshold p_c for which an open path almost surely exists when $p > p_c$ and almost surely does not exist when $p < p_c$.

Definition 5.1. *The* percolation threshold *with respect to some parameter ρ is the threshold function $t(\rho)$ for a behavior affected by ρ in the percolation model for a system.*

One area of research in percolation theory has been to determine the percolation thresholds for various properties. Another area has focused on studying the behavior of systems

at or around some percolation threshold. As a result, percolation models are categorized in terms of the percolation threshold.

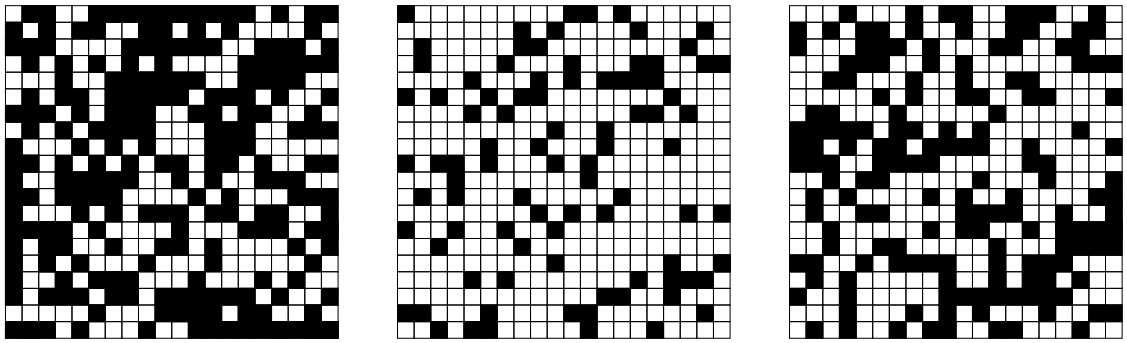
Definition 5.2. *Let p be the value of the parameter measured by the percolation threshold p_c in a percolation model.*

(i) If $p < p_c$, then the model is subcritical.

(ii) If $p > p_c$, then the model is supercritical.

(iii) If $p = p_c$, then the model is critical.

Figure 5.1 provides an example of a subcritical, supercritical, and critical percolation model for the tiling of a grid.



(a) A subcritical model. (b) A supercritical model. (c) A critical model.

Figure 5.1: A percolation model where tiles are open (white) with probability p and closed (black) with probabiltiy $1 - p$.

Another example of a percolation model can be used to describe Erdős and Rényi's result on the presence of a giant component in $G(n, N)$, which was stated in Proposition 2.4 [14]. For $G(n, N)$, define c such that $N = cn + o(1)$. Then the percolation threshold for the presence of a giant component is $c = \frac{1}{2}$. Thus, in the subcritical model $c < \frac{1}{2}$ and so asymptotically almost surely $G(n, N)$ has no giant component. In the supercritical model, $c > \frac{1}{2}$ and so asymptotically almost surely $G(n, N)$ has a giant component. The critical model was shown by Łuczak in [28] to have its largest component be of size $\theta(n^{2/3})$.

As suggested by the 30 years between [14] and [28], it is often difficult to analyze the behavior of a system when the system is modeled by a critical or nearly critical percolation

model. Such models are often referred to as being in the *critical window* and typically require different techniques.

In this chapter, we will focus on two types of discrete percolation, known as *site* and *bond* percolation. These types of percolation are closely related, which we discuss in Section 5.1.2. In Section 5.1, we will also discuss some known percolation thresholds for both site and bond percolation. Section 5.2 will present the technique of Janson from [22] that we use in Chapter 6.

5.1 Discrete Percolation

As mentioned at the start of this chapter, percolation theory studies fluid flow and other similar processes in random media. Depending on the medium, different models are used to analyze these behaviors. For example, if the medium being studied is only able to be modeled by uncountable domains, such as $\mathbb{R}^d$ or non-discrete subsets of $\mathbb{R}^d$, then the percolation model must also be described on a continuum. This is often referred to as continuum percolation theory and several models have been developed to study such percolation.

Graphs are an example of a medium that is a discrete rather than continuous set. Such media fall under the domain of discrete percolation theory.

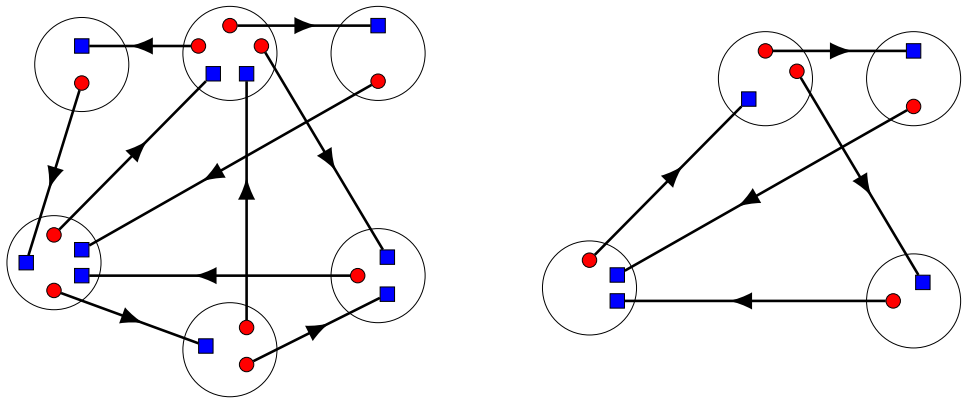
Definition 5.3. Discrete percolation *is the term used to describe models of percolation theory whose media are discrete sets.*

Many common models for studying discrete percolation use a regular point lattice as the underlying model of the medium. On such a structure, several types of percolation can be applied. The two most common types of percolation applied to the point lattice are *site percolation* and *bond percolation*. We discuss these types of percolation and some results known for each in the following subsections.

5.1.1 Site Percolation

Although site percolation is often applied to a regular point lattice, we will analyze site percolation on directed and undirected graphs. We therefore define site percolation on a graph rather than on a lattice.

Definition 5.4. *Let G be a graph and $0 < p < 1$. For each vertex v in G, remove v and its incident edges with probability $1 - p$ independently of all other vertices. This model is called* site percolation.



(a) G (b) Site percolation on G, $p = .5$.

Figure 5.2: Site percolation on a directed graph G.

Site percolation gets its name from the standard percolation theory terminology where vertices are called "sites" and edges are called "bonds."

Although not as commonly used as bond percolation, site percolation can be used to study the presence of an infinite component in a random subgraph of the infinite graph. For finite graphs with a large number of vertices, site percolation instead studies the presence of a giant component in the graph. This can be further generalized to sequences of random graphs on n vertices with fixed degree sequences.

One of the first to study site percolation on sequences of random graphs was Fountoulakis. In [15] from 2008, he found the percolation threshold p_{site} that determines the existence of a giant component in a graph G' resulting from site percolation of a sparse random graph with fixed degree sequence. In other words, when $p < p_{site}$, for all $\epsilon > 0$, with high probability G' contains no components with at least ϵn vertices. However, if $p > p_{site}$, then (with high probability) G' contains a component of size $\theta(n)$.

Fountoulakis' work was extended by Janson in 2009 [22] through an "explosion" technique. We will discuss this technique in Section 5.2. Janson also applied the technique to find the threshold for the presence of a k-core after site percolation. Some other known results for site percolation on various types of undirected graphs can be found in [18] and [7].

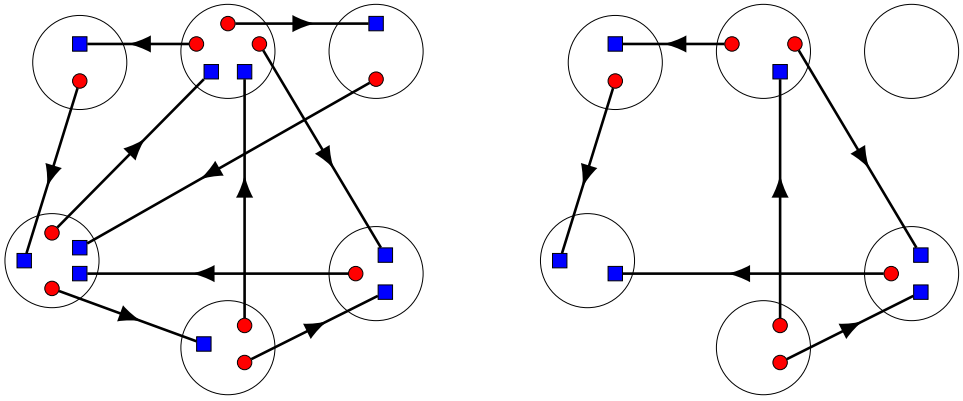
Site percolation on directed graphs is much less studied. Schwartz et al. [40] found that site percolation in directed graphs is greatly affected by the existence of correlations between a vertex's in-degree and its out-degree. However, their results are not proved rigorously. In Chapter 6, we present new results for site percolation in graphs with well-

behaved degree sequences.

5.1.2 Bond Percolation

As with site percolation, we define bond percolation on a graph rather than on a lattice.

Definition 5.5. *Let G be a graph and $0 < p < 1$. For each edge e in G, remove e with probability $1 - p$ independently of all other edges. All vertices remain in the graph. This model is called* bond percolation.



(a) G (b) Bond percolation on G, $p = .5$.

Figure 5.3: Bond percolation on a directed graph G.

Bond percolation also gets its name from standard percolation theory terminology as edges are called "bonds."

Note that $G(n, p)$ can be thought of as bond percolation on the complete graph on n vertices. Because of this relation, bond percolation has been well studied for a wide variety of properties in undirected graphs. In addition, the self-duality of the square lattice has made bond percolation on lattices one of the most studied percolation processes. Grimmett's text [18] consists almost entirely of results for bond percolation on d-dimensional cubic lattices.

In terms of sequences of finite graphs, Fountoulakis also found the percolation threshold p_{bond} for the existence of a giant component after bond percolation on sparse random graphs with fixed degree sequences [15]. In fact, he showed that for such degree sequences, $p_{site} = p_{bond}$. This relationship between the thresholds of the two models does not hold in general. However, the following relation always holds.

Proposition 5.6 ([18]). *Let G be a connected, infinite, locally finite multigraph. Then the percolation thresholds of G satisfy $p_{bond} \leq p_{site}$.*

Proof: Let v be a vertex of G. We define G_n to be the subgraph of G induced by the vertices of G within distance n of v. Consider site percolation on G_n and define $K_s(v)$ to be the open component containing v in G_n. Note that $K_s(v)$ is empty if v is closed by the percolation, so we may condition on v being open.

We explore $K_s(v)$ conditioned on v being open by testing the state after percolation of the vertices in G_n. This is done by constructing a random sequence $\Gamma = \{(\mathcal{O}_t, \mathcal{C}_t, \mathcal{U}_t)\}_{1 \leq t \leq \ell}$ of tripartitions of the vertex set $V(G_n)$ for some $\ell \leq |V(G_n)|$. In this sequence, $\mathcal{O}_t$ is the set of open vertices, $\mathcal{C}_t$ the set of closed vertices, and $\mathcal{U}_t$ the set of untested vertices after the state of the t^{th} vertex in $V(G_n)$ is determined. The sequence stops in iteration ℓ when every vertex adjacent to a vertex in $\mathcal{O}_\ell$ is in $\mathcal{O}_\ell \cup \mathcal{C}_\ell$. In this case, $K_s(v) = \mathcal{O}_\ell$.

To define Γ, let $\mathcal{O}_1 = \{v\}$, $\mathcal{C}_1 = \emptyset$, and $\mathcal{U}_1 = V(G_n) \backslash \{v\}$. Given $(\mathcal{O}_t, \mathcal{C}_t, \mathcal{U}_t)$, if there is no edge between a vertex of $\mathcal{O}_t$ and $\mathcal{U}_t$, then $t = \ell$ and the sequence terminates. Otherwise, choose an edge $o_t u_t$ with $o_t \in \mathcal{O}_t$ and $u_t \in \mathcal{U}_t$ and set $\mathcal{U}_{t+1} = \mathcal{U}_t \backslash \{u_t\}$. We test u_t to determine if it is open or closed. If u_t is open, then $\mathcal{O}_{t+1} = \mathcal{O}_t \cup \{u_t\}$ and $\mathcal{C}_{t+1} = \mathcal{C}_t$. If not, then $\mathcal{O}_{t+1} = \mathcal{O}_t$ and $\mathcal{C}_{t+1} = \mathcal{C}_t \cup \{u_t\}$. Note that at each step, the conditional probability u_t is open is p. Furthermore, this process terminates since G_n is finite.

By the construction, for every t, $\mathcal{O}_t$ is a set of open vertices such that the graph induced by $\mathcal{O}_t$ is connected and all of the vertices in $\mathcal{C}_t$ are closed. As no vertex in $\mathcal{O}_t$ has a neighbor in $\mathcal{U}_\ell = V(G_n) \backslash (\mathcal{O}_\ell \cup \mathcal{C}_\ell)$, O_ℓ is precisely the open component of v in G_n, or G_v^s.

To compare G_v^s to G_v^b, which is the open component containing v in the bond percolation of G_n, we explore G_v^b in a similar fashion. Let $\Gamma' = (\mathcal{O}'_t, \mathcal{C}'_t, \mathcal{U}'_t)_{t=1}^{\ell'}$ be a random sequence that is constructed in the same manner as Γ with one addition. Once edge $e_t = o_t u_t$ is chosen, e_t is tested to determine if it is open. As this is the first time e_t is tested, conditional on the sequence Γ' up to step t, the probability e_t is open is p. Hence Γ and Γ' have the same distribution and so $|G_v^s| = |\mathcal{O}_\ell|$ and $|\mathcal{O}'_{\ell'}|$ have the same distribution.

As $\mathcal{O}'_{\ell'}$ is contained in the open component G_v^b of v in G_n in the bond percolation,

$$\mathbf{Pr}_s(|G_v^s| \geq n \mid v \text{ is open}) \leq \mathbf{Pr}_b(|G_v^b| \geq n),$$

where $\mathbf{Pr}_s$ and $\mathbf{Pr}_b$ denote the probability measures on the set of subgraphs of G in site and bond percolation respectively. Furthermore, for every vertex x of G, every integer $n \geq 1$, and every probability $0 < p < 1$,

$$\mathbf{Pr}_s(|G_x^s| \geq n) \leq p \mathbf{Pr}_b(|G_x^b| \geq n).$$

Let θ_x denote the probability that the open component containing x is infinite. Then as $n \to \infty$, $\theta_x(G_x^s) \leq p\theta_x(G_x^b)$. Thus, if $\theta_x(G_x^b) = 0$, then $\theta_x(G_x^s) = 0$ and so $p_{bond} \leq p_{site}$. $\square$

In Chapter 6, we will study site percolation instead of bond percolation. This is because bond percolation on a graph G is equivalent to site percolation on the *line graph* of G, denoted $L(G)$. Hence percolation thresholds for bond percolation can be found using site percolation on the line graph of the original graph.

The next section describes the technique introduced by Janson in [22]. The version of this method presented is for directed graphs and will be used in Chapter 6.

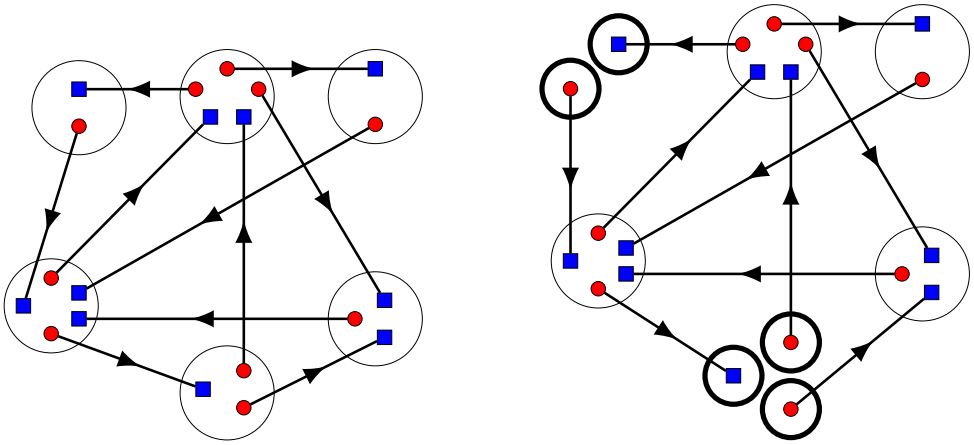
5.2 Percolation by Exploding Vertices and Edges

In [22] from 2009, Janson introduced a method of performing site and bond percolation that completes the required vertex and edge deletions in two steps. First, vertices (respectively edges) are "exploded." Then, uniformly at random some vertices of degree 1 and their incident edges are removed. In this section, we adapt this technique for directed graphs with well-behaved degree sequences.

We first describe the method for site percolation using the configuration model. Let C be a configuration with a well-behaved degree sequence and $0 < p < 1$ a probability. Recall that for a bin $v \in V(C)$, $R(v)$ and $B(v)$ are the number of red and blue points respectively contained in v. For each bin v in C, with probability $1 - p$ replace it with $R(v)$ bins containing 1 red and no blue points and $B(v)$ bins containing 1 blue and no red points. This process of replacing a bin is referred to as *exploding* that bin.

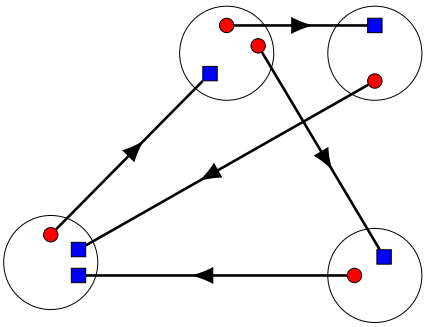
The result of this first step is a new configuration C^*. We let $\mathcal{N}_{i,j}^*$ denote the set of bins in C^* that contain exactly i blue and j red points and let $n_{i,j}^* = |\mathcal{N}_{i,j}^*|$. The next step is to delete $n_{1,0}^* - n_{1,0}$ randomly chosen bins in $\mathcal{N}_{1,0}^*$ as well as $n_{0,1}^* - n_{0,1}$ randomly chosen bins in $\mathcal{N}_{1,0}^*$ together with all points in pairs with points in these bins. This results in a configuration C' with the same degree sequence as the configuration obtained by deleting the selected bins and their pairs in C directly. Figure 5.4 demonstrates this procedure.

The technique for bond percolation is similar. However, instead of exploding bins into new bins that contain exactly one point, we "explode" pairs. This is accomplished by replacing the partner (i.e. other point in its pair) of each point in the pair with a new point such that this point is in a new bin, it is the same color as the partner, and it is the only point in its bin. Note that this replaces the exploded pair with two new pairs. Again,



(a) The original configuration C.

(b) The exploded configuration C^*.



(c) The resulting configuration C'.

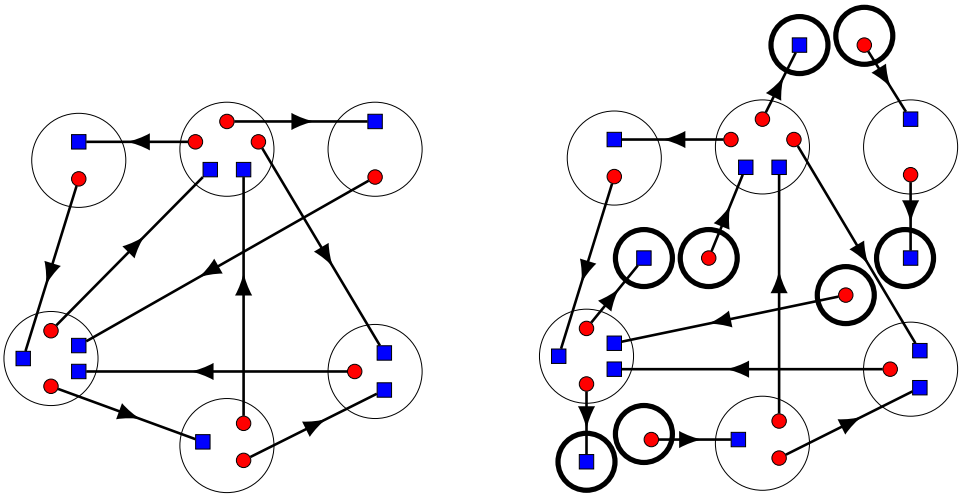
Figure 5.4: Site percolation using Janson's method.

the second step involves deleting a certain amount of bins containing only one point and the points in pairs with points in these bins.

To be more precise, let C be a configuration on n bins with a well-behaved degree sequence and $0 < p < 1$ a probability. For each entry d_i^+ and d_i^- in the degree sequence, replace them by independent random degrees f_i^+ and f_i^- which have binomial distributions $\mathrm{Bi}(d_i^+, \sqrt{p})$ and $\mathrm{Bi}(d_i^-, \sqrt{p})$ respectively. Then add $n^+ = \sum_{i=1}^{n}(d_i^+ - f_i^+)$ in-degree 0 and out-degree 1 as well as $n^- = \sum_{i=1}^{n}(d_i^- - f_i^-)$ in-degree 1 and out-degree 0 terms to the degree

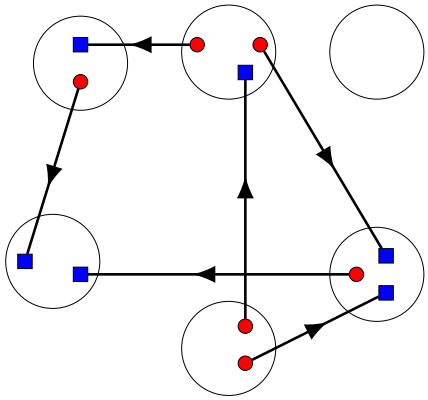
sequence. This defines the degree sequence of the configuration obtained by *exploding* a pair with probability $1 - p$.

The next step is to construct a random configuration C^* with this new degree sequence and then delete n^- randomly chosen bins of $\mathcal{N}^*_{1,0}$ as well as n^+ randomly chosen bins of $\mathcal{N}^*_{0,1}$ together with all points in pairs with points in these bins. This results in a configuration with the same degree sequence as the configuration obtained by deleting the selected pairs of C directly. Figure 5.5 demonstrates this procedure.



(a) The original configuration C.

(b) The exploded configuration C^*.



(c) The resulting configuration C'.

Figure 5.6: Bond percolation using Janson's method.

The benefit of using this method is that it reduces percolation problems to a simple random modification of the degree sequence followed by a random removal of a set of bins that contain only one point. Since a bin must have at least one red and one blue point to be included in a SCC, removing this set of bins does not affect the size of the largest SCC. Thus, the size of the largest SCC in configurations with the modified degree sequence is the same as the size of the largest SCC in the random configurations resulting from percolation on the original configuration.

Therefore, assuming the degree sequence of the exploded graph is well-behaved, we may apply Theorem 4.1 to these modified degree sequences to determine the percolation threshold for graphs with well-behaved degree sequences. We will see an example of such a result in Chapter 6.

Chapter 6

Percolation Thresholds for Giant Strongly Connected Components - A Heuristic Investigation

As mentioned in Chapter 5, both site and bond percolation have been studied on undirected graphs. Percolation thresholds for the presence of a giant connected component are known for many types of graphs, including lattices $\mathbb{Z}^d$ for $d \geq 2$ [18] and $G(n,p)$ [14]. Other results are known for sequences of graphs, including sequences of hypercubes [1, 9], graphs with uniformly bounded maximum degree [2], and sparse random graphs on n vertices [15, 22].

In this chapter, we use a heuristic approach to predict a certain site percolation threshold for the presence of a giant strongly connected component for graphs with well-behaved degree sequences. This approach will use the adaptation of Janson's technique shown in Section 5.2 to reduce the problem to studying the modified degree sequence of the exploded graph. We then present some arguments that suggest the exploded graph is likely to have a well-behaved degree sequence. If this is true, then Theorem 4.1 can be applied to the exploded graph's degree sequence to obtain the site percolation threshold.

6.1 Site Percolation on $\mathcal{G}(D_n)$

Let D_n be the well-behaved degree sequence. Recall that $\mathcal{G}(D_n)$ is the set of all graphs with degree sequence D_n and $\lambda(D_n) = \sum_{i\geq 0}\sum_{j\geq 0} ij\frac{n_{i,j}}{\theta n}$. We trivially have the following result.

Lemma 6.1. *Let D_n be a well behaved degree sequence such that $\lambda(D_n) = 1 - \epsilon$ for some $\epsilon > 0$. For any $0 \leq p \leq 1$, let G be a random graph in $\mathcal{G}(D_n)$ and G'_p be a random graph obtained by site percolation on G where vertices are retained with probability p. Then a.a.s. every SCC of G'_p has size $O([\Delta(D_n)]^2 \log(n))$.*

Proof: Let $0 \leq p \leq 1$. Note that the largest SCC of any graph resulting from site percolation on G is at most the size of the largest SCC of G. By Theorem 4.1, a.a.s. every SCC of G has size $O([\Delta(D_n)]^2 \log(n))$. Thus every SCC of G'_p has size $O([\Delta(D_n)]^2 \log(n))$. $\square$

For the remainder of this section, we will assume $\lambda(D_n) = 1 + \epsilon$ for some $\epsilon > 0$. Also, we assume G is a random graph in $\mathcal{G}(D_n)$ and p is a probability such that $0 < p < 1$. Let G^* be a random graph obtained by exploding the vertices of G with probability $1 - p$ and D_n^* be the degree sequence of G^*.

We believe it should be easy to prove that if D_n is a well-behaved degree sequence, then a.a.s. D_n^* is a well-behaved degree sequence such that $\lambda(D_n^*) = p\lambda(D_n)$. However, we will only present some partial and supporting arguments for this result in this section. We begin with some arguments that suggest a.a.s. D_n^* is well-behaved.

Let $\mathcal{N}_{i,j}^*$ to be the set of all vertices in G^* of in-degree i and out-degree j and let $n_{i,j}^* = |\mathcal{N}_{i,j}^*|$. Also, let n^* denote the number of vertices in G^*. We have the following result.

Lemma 6.2. *Let D_n be a well behaved degree sequence such that $\lambda(D_n) = 1 + \epsilon$ for some $\epsilon > 0$, G be a random graph in $\mathcal{G}(D_n)$, and $0 < p < 1$ be a probability. Let G^* be a random graph obtained by exploding the vertices of G with probability $1-p$ and let D_n^* be the degree sequence of G^*. Then,*

(i) D_n^* is feasible;

(ii) $\Delta(D_n^*) = o[(n^*)^{1/4}]$; and

(iii) $\max\left(\sum_{i\geq 0}\sum_{j\geq 0} i^2 \frac{n_{i,j}^*}{n^*}, \sum_{i\geq 0}\sum_{j\geq 0} j^2 \frac{n_{i,j}^*}{n^*}\right) \leq A_2^*$ *for some absolute constant A_2^*.*

Proof. (i) $G^* \in \mathcal{G}(D_n^*)$ and so this holds trivially.

(ii) For all such G^*, $n^* \geq n$. Furthermore, $\Delta^+(D_n^*) \leq \Delta^+(D_n)$ and $\Delta^-(D_n^*) \leq \Delta^-(D_n)$. Hence $\Delta(D_n) = o(n^{1/4})$ implies $\Delta(D_n^*) = o[(n^*)^{1/4}]$.

(iii) Note that when a vertex in $\mathcal{N}_{i,j}$ is exploded, the amount it contributes to $\sum_{i\geq 0}\sum_{j\geq 0} i^2\frac{n^*_{i,j}}{n^*}$ is at most the amount it contributes to $\sum_{i\geq 0}\sum_{j\geq 0} i^2\frac{n_{i,j}}{n}$ (since $1^2 \leq i^2$ for all $i > 0$ and $n^* \geq n$). Similarly for $\sum_{i\geq 0}\sum_{j\geq 0} j^2\frac{n^*_{i,j}}{n^*}$. Thus,

$$\max\left(\sum_{i\geq 0}\sum_{j\geq 0} i^2\frac{n^*_{i,j}}{n^*}, \sum_{i\geq 0}\sum_{j\geq 0} j^2\frac{n^*_{i,j}}{n^*}\right) \leq \max\left(\sum_{i\geq 0}\sum_{j\geq 0} i^2\frac{n_{i,j}}{n}, \sum_{i\geq 0}\sum_{j\geq 0} j^2\frac{n_{i,j}}{n}\right) \leq A_2$$

for some absolute constant A_2. Hence $A_2^* = A_2$ suffices. □

By Lemma 6.2, D_n^* is feasible and satisfies the first and third conditions of Definition 2.17. It remains to show that D_n^* is smooth and satisfies the second and fourth conditions. We provide some support for these conditions by performing the necessary calculations using $\mathbf{E}(n^*_{i,j})$ and $\mathbf{E}(n^*)$ instead of $n^*_{i,j}$ and n^* for all $i, j \geq 0$. We expect some concentration arguments can be used to show the results of the calculations with $n^*_{i,j}$ and n^* are similar to those with $\mathbf{E}(n^*_{i,j})$ and $\mathbf{E}(n^*)$.

Therefore, we begin by calculating $\mathbf{E}(n^*_{i,j})$ and $\mathbf{E}(n^*)$ for all $i, j \geq 0$. Note that for all $(i, j) \notin \{(1, 0), (0, 1)\}$, the number of vertices in $\mathcal{N}_{i,j}$ that are not exploded is equal to $n^*_{i,j}$ and has binomial distribution $\mathbf{Bi}(n_{i,j}, p)$. Thus

$$\mathbf{E}(n^*_{i,j}) = pn_{i,j}.$$

Furthermore, $\mathcal{N}^*_{1,0}$ consists of the vertices in $\mathcal{N}_{1,0}$ as well as all the new vertices of in-degree 1 and out-degree 0 that were created by vertex explosions. Hence

$$\mathbf{E}(n^*_{1,0}) = n_{1,0}+(1-p)\left[\left(\sum_{i\geq 1}\sum_{j\geq 0} in_{i,j}\right) - n_{1,0}\right] = n_{1,0}+(1-p)(\theta n-n_{1,0}) = pn_{1,0}+(1-p)\theta n.$$

Similarly,

$$\mathbf{E}(n^*_{0,1}) = n_{0,1}+(1-p)\left[\left(\sum_{i\geq 1}\sum_{j\geq 0} jn_{i,j}\right) - n_{0,1}\right] = n_{0,1}+(1-p)(\theta n-n_{0,1}) = pn_{0,1}+(1-p)\theta n.$$

Thus,

$$\mathbf{E}(n^*) = \mathbf{E}\left(\sum_{i\geq 0}\sum_{j\geq 0} n^*_{i,j}\right) = 2(1-p)\theta n + p\sum_{i\geq 0}\sum_{j\geq 0} n_{i,j} = [2\theta(1-p)+p]n.$$

Hence for $(i,j) \notin \{(1,0),(0,1)\}$,

$$\begin{aligned}\lim_{n\to\infty} \frac{\mathbf{E}(n^*_{i,j})}{\mathbf{E}(n^*)} &= \lim_{n\to\infty} \frac{pn_{i,j}}{[2\theta(1-p)+p]n} \\ &= \lim_{n\to\infty} \frac{p}{2\theta(1-p)+p}\left(\frac{n_{i,j}}{n}\right) \\ &= \frac{p}{2\theta(1-p)+p}\kappa_{i,j}.\end{aligned}$$

Also,

$$\begin{aligned}\lim_{n\to\infty} \frac{\mathbf{E}(n^*_{1,0})}{\mathbf{E}(n^*)} &= \lim_{n\to\infty} \frac{pn_{1,0}+(1-p)\theta n}{[2\theta(1-p)+p]n} \\ &= \lim_{n\to\infty} \frac{p}{2\theta(1-p)+p}\left(\frac{n_{i,j}}{n}\right) + \frac{(1-p)\theta}{2\theta(1-p)+p} \\ &= \frac{p}{2\theta(1-p)+p}\kappa_{1,0} + \frac{(1-p)\theta}{2\theta(1-p)+p}\end{aligned}$$

and

$$\begin{aligned}\lim_{n\to\infty} \frac{\mathbf{E}(n^*_{0,1})}{\mathbf{E}(n^*)} &= \lim_{n\to\infty} \frac{pn_{0,1}+(1-p)\theta n}{[2\theta(1-p)+p]n} \\ &= \frac{p}{2\theta(1-p)+p}\kappa_{0,1} + \frac{(1-p)\theta}{2\theta(1-p)+p}.\end{aligned}$$

Since $\kappa_{i,j}$ is constant for all $i \geq 0$ and $j \geq 0$ and $p < 1$, for all $i \geq 0$ and $j \geq 0$ there exists constants $\kappa^*_{i,j}$ such that $\lim_{n\to\infty} \frac{\mathbf{E}(n^*_{i,j})}{\mathbf{E}(n^*)} = \kappa^*_{i,j}$. Thus, if $n^*_{i,j}$ and n^* can be shown to be concentrated around their expectations for all $i \geq 0$ and $j \geq 0$, then D^*_n is smooth.

We continue the discussion with some more calculations that suggest D^*_n satisfies the two remaining conditions of Definition 2.17. First, note that the number of arcs in G and G^* is the same. Hence $\theta^* n^* = \theta n$ and so $\theta^* = \frac{n}{n^*}\theta$. Thus,

$$\frac{n}{\mathbf{E}(n^*)}\theta = \frac{n}{[2\theta(1-p)+p]n}\theta = \frac{1}{2\theta(1-p)+p}(1+o(1))A_1.$$

This suggests that $\theta^* = (1+o(1))A^*_1$ for an absolute constant A^*_1 and so the first condition holds for D^*_n.

The final condition requires $\lambda^* = \lim_{n\to\infty} \sum_{i,j\geq 0} ij\frac{n^*_{i,j}}{\theta^* n^*}$ to exist, be finite, and have this sum approach the limit uniformly. Since D_n is well-behaved, for all $\delta > 0$, there exists k and N such that for all $n > N$, $\left|\sum_{i=0}^{k}\sum_{j=0}^{k} ij\frac{n_{i,j}}{\theta n} - \lambda\right| < \delta$. Thus,

$$\begin{aligned}
\left|\left(\sum_{i=0}^{k}\sum_{j=0}^{k} ij\frac{\mathbf{E}(n^*_{i,j})}{\mathbf{E}(\theta^* n^*)}\right) - p\lambda\right| &= \left|\left(\sum_{i=1}^{k}\sum_{j=1}^{k} ij\frac{\mathbf{E}(n^*_{i,j})}{\theta n}\right) - p\lambda\right| \\
&= \left|\left(\sum_{i=1}^{k}\sum_{j=1}^{k} ij\frac{pn_{i,j}}{\theta n}\right) - p\lambda\right| \\
&= \left|p\left(\sum_{i=1}^{k}\sum_{j=1}^{k} ij\frac{n_{i,j}}{n} - \lambda\right)\right| \\
&= p\left|\sum_{i=0}^{k}\sum_{j=0}^{k} ij\frac{n_{i,j}}{n} - \lambda\right| \\
&< p\delta.
\end{aligned}$$

Hence

$$\lambda^* = \lim_{n\to\infty} \sum_{i\geq 0}\sum_{j\geq 0} ij\frac{\mathbf{E}(n^*_{i,j}(n))}{\theta^*[n^*(n)]} = p\lambda$$

and the sum approaches this limit uniformly. Since λ is finite, λ^* is also finite. This suggests D^*_n satisfies the fourth condition of Definition 2.17.

Lemma 6.2 and the above calculations support the claim that for a well-behaved degree sequence D_n, D^*_n is well-behaved a.a.s. Furthermore,

$$\begin{aligned}
\lambda(D^*_n) &= \sum_{i\geq 0}\sum_{j\geq 0} ij\frac{\mathbf{E}(n^*_{i,j})}{\theta^* n^*} \\
&= \sum_{i\geq 1}\sum_{j\geq 1} ij\frac{\mathbf{E}(n^*_{i,j})}{\theta n} \\
&= \sum_{i\geq 1}\sum_{j\geq 1} ij\frac{pn_{i,j}}{\theta n} \\
&= p\sum_{i\geq 0}\sum_{j\geq 0} ij\frac{n_{i,j}}{\theta n} \\
&= p\lambda(D_n).
\end{aligned}$$

This final calculation supports the claim that $\lambda(D_n^*) = p\lambda(D_n)$. This leads us to believe that with some concentration arguments, it can likely be proved that for a well-behaved degree sequence D_n, a.a.s. D_n^* is well-behaved and $\lambda(D_n^*) = p\lambda(D_n)$. We can use this and Theorem 4.1 to predict a site percolation threshold for the presence of a giant strongly connected component in graphs with well-behaved degree sequences where $\lambda(D_n) > 1$.

First, consider a configuration with degree sequence D_n. With some additional arguments, it can be shown that the site percolation technique outlined in Section 5.2 will uniformly produce configurations with a degree sequence D_n^* that is well-behaved and has $\lambda(D_n^*) = p\lambda(D_n)$. We can then apply Theorem 4.1 to D_n^* to see that for $\delta > 0$,

1. If $p < \frac{1-\delta}{\lambda(D_n)}$, then a.a.s. every SCC of G_p' has size $O([\Delta(D_n)]^2 \log(n))$.
2. If $p > \frac{1+\delta}{\lambda(D_n)}$, then a.a.s. there exists a SCC in G_p' of size $\Theta(n)$.

This suggests that the site percolation threshold for the presence of a giant strongly connected component in a graph with well-behaved degree sequence D_n is $p = \frac{1}{\lambda(D_n)}$. This threshold is not surprising as it is the directed graph equivalent to the site percolation threshold for undirected graphs found by Fountoulakis in [15].

Chapter 7

Concluding Remarks

Although much is known about the size of the largest component of random graphs, little is known about the size of the largest strongly connected component of a random directed graph. One of the results in the area was presented by Cooper and Frieze in 2004 [12]. They found a threshold function for the presence of a giant strongly connected component in random directed graphs with proper degree sequences. However, the constraints on such degree sequences, such as the maximum term in the sequence being $\frac{n^{1/12}}{\log n}$, make these results have limited applications for modeling real-world systems.

In this thesis, we defined a similar type of degree sequence, called a well-behaved degree sequence. These degree sequences permit terms of higher degree (of size $o(n^{1/4})$) in the sequence, but require a stronger regularity condition than proper degree sequences. Using the configuration model and some coupling arguments, we found a threshold function for the presence of a giant strongly connected component in a random directed graph with a well-behaved degree sequence. This threshold function is identical to the function found by Cooper and Frieze.

However, like the results of Cooper and Frieze, directed graphs with well-behaved degree sequences have limited applications to real-world models. This is because most complex networks, such as the World Wide Web and metabolic networks, have been shown to exhibit power-law degree distributions with $2 < \gamma < 3$ [40, 15]. Recall that for power-law degree distributions, the number of vertices of in-degree i is proportional to $ci^{-\gamma}$ for some constants $c, \gamma > 0$ or the number of vertices of out-degree j is proportional to $c'j^{-\gamma'}$ for some constants $c', \gamma' > 0$. These real-world networks could have degree sequences that are not well-behaved since they may contain many vertices with in-degree or out-degree that are functions of n (so the final condition of Definition 2.17 isn't satisfied). However, it

is likely that the proof techniques presented in Chapter 4 can be extended to study such power-law degree sequences, whereas the technique used by Cooper and Frieze does not extend. Thus, more work is needed to study random directed graphs with these power-law degree distributions.

In addition to studying the presence of strongly connected components in well-behaved degree sequences, we discussed how our results could be applied to percolation theory. Specifically, we used a heuristic approach to predict a site percolation threshold for random directed graphs with well-behaved degree sequences. We provided some arguments supporting the claim that the exploded graph from Janson's technique has a well-behaved degree sequence. Thus, if this claim is true, our threshold function predicts a site percolation threshold for well-behaved degree sequences. Further work is need to prove this claim, but we suspect some concentration arguments can be used to prove it.

Another area for further investigation includes determining the bond percolation threshold for the presence of a giant strongly connected component in directed graphs with well-behaved degree sequences. This could be accomplished by applying the modified version of Janson's bond percolation technique. As with site percolation, some arguments for the resulting exploded directed graph having a well-behaved degree sequence will be necessary. Once this is shown, our result can be applied to the degree sequence of the exploded graph to find the exact bond percolation threshold.

References

[1] M. Ajtai, J. Komlós, and E. Szemerédi. Largest random component of the k-cube. *Combinatorica*. 2: 1-7, 1982.

[2] N. Alon, I. Benjamini, and A. Stacey. Percolation on finite graphs and isoperimetric inequalities. *Annals of Probability*. 32: 17-27-1745, 2004.

[3] N. Alon and A. Stacey. *The Probabilistic Method* (3rd ed.). John Wiley & Sons. Hoboken: 2008.

[4] M. Bloznelis, F. Götze, and J. Laworski. Birth of a strongly connected giant in an inhomogeneous random digraph. *Journal of Applied Probability*. 49: 601-611, 2012.

[5] B. Bollobás. The evolution of random graphs. *Transactions of the American Mathematical Society*. 286: 257-274, 1984.

[6] B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures and Algorithms*. 31: 3-122, 2007.

[7] B. Bollobás and O. Riordan. *Percolation*. Cambridge University Press. Cambridge: 2006.

[8] B. Bollobás and O. Riordan. An old approach to the giant component problem. *Journal of Combinatorial Theory, Series B*. 113: 236-260, 2015.

[9] C. Borgs, J. Chayes, R. van der Hofstad, G. Slade, and J. Spencer. The phase transition for the n-cube. *Combinatorica*. 26: 395-410, 2006.

[10] S. Broadbent and J. Hammersley. Percolation process I. Crystals and mazes. *Proceedings of the Cambridge Philosophical Society*. 53: 629-641, 1957.

[11] F. Chung and L. Lu. *Complex Graphs and Networks*. AMS. 2006.

[12] C. Cooper and A. Frieze. The size of the largest strongly connected component of a random digraph with a given degree sequence. *Combinatorics, Probability and Computing*. 13: 319-337, 2004.

[13] P. Erdős and A. Rényi. On random graphs I. *Publicationes Mathematicae* 6: 290-297, 1959.

[14] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl*. 5: 17-61, 1960.

[15] N. Fountoulakis. Percolation on sparse random graphs with given degree sequence. *Internet Mathematics*. 4(4): 329-356, 2007.

[16] E. Gilbert. Random graphs. *Annals of Mathematical Statistics* 30(4): 1141-1144, 1959.

[17] C. Greenhill, B. McKay, and X. Wang. Asymptotic enumeration of sparse 0-1 matrices with irregular row and column sums. *Journal of Combinatorial Theory*. Series A, 113: 291-324, 2006.

[18] G. Grimmett. *Percolation* (2nd ed.). Springer. Berlin: 1999.

[19] A. Gut. *Probability: A Graduate Course*. Springer. New York: 2005.

[20] T. Harris. *The Theory of Branching Processes*. Dover Publications. Mineola: 1989.

[21] H. Hatami and M. Molloy. The scaling window for a random graph with a given degree sequence. *Random Structures and Algorithms*. 41(1): 99-123, 2012.

[22] S. Janson. On percolation in random graphs with given vertex degrees. *Electronic Journal of Probability*. 14(5): 87-118, 2009.

[23] S. Janson and M. Luczak. A new approach to the giant component problem. *Random Structures and Algorithms*. 34(2): 197-216, 2009.

[24] O. Kallenberg. *Foundations of Modern Probability* (2nd ed.). Springer. New York: 2002.

[25] R. Karp. The transitive closure of a random digraph. *Random Structures and Algorithms*. 1(1): 73-93, 1990.

[26] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. *Computer Networks: The International Journal of Computer and Telecommunications Networking*. 31(11-16): 1481-1493, 1999.

[27] M. Kang and T. Seierstad. The critical phase for random graphs with a given degree sequence. *Combinatorics, Probability and Computing*. 17(1): 67-86, 2008.

[28] T. Łuczak. Component behavior near the critical point of the random graph process. *Random Structures and Algorithms*. 1(3): 287-310, 1990.

[29] T. Łuczak. The phase transition in the evolution of random digraphs. *Journal of Graph Theory*. 14(2): 217-223, 1990.

[30] T. Łuczak and J. Cohen. Giant components in three-parameter random directed graphs. *Advances in Applied Probability*. 24(4): 845-857, 1992.

[31] T. Łuczak, B. Pittel, and J. Wierman. The structure of a random graph at the point of phase transition. *Transactions of the American Mathematical Society*. 341(2): 721-748, 1994.

[32] T. Łuczak and T. Seierstad. The critical behavior of random digraphs. *Random Structures and Algorithms*. 35(3): 271-293, 2009.

[33] B. McKay. Asymptotics for 0-1 matrices with prescribed line sums. *Enumeration and Design*. 225-238, 1984.

[34] B. McKay and N. Wormald. Uniform generation of random regular graphs of moderate degree. *Journal of Algorithms*. 11: 52-67, 1990.

[35] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*. 6(2-3): 161-180, 1995.

[36] M. Molloy and B. Reed. The size of the largest component of a random graph on a fixed degree sequence. *Combinatorics, Probability and Computing*. 7(3): 295-305, 1998.

[37] A. Nachmias and Y. Peres. The critical random graph, with martingales. *Israel Journal of Mathematics*. 176: 29-41, 2010.

[38] B. Pittel. Edge percolation on a random regular graph of low degree. *The Annals of Probability*. 36(4): 1359-1389, 2008.

[39] O. Riordan. The phase transition in the configuration model. *Combinatorics, Probability and Computing*. 21(1-2): 265-299, 2012.

[40] N. Schwartz, R. Cohen, D. ben-Avraham, A. Barabási, and S. Havlin. Percolation in directed scale-free networks. *Physical Review*. 66(015104): 1-4, 2002.

[41] H. Watson and F. Galton. On the probability of the extinction of families. *Journal of the Anthropological Institute of Great Britain and Ireland*. 4: 138-144, 1875.

[42] N. Wormald. Models of random regular graphs. In: *Surveys in Combinatorics, London Mathematical Society Lecture Note Series*. 267: 239-298, 1999.