ON PERMUTATION PATTERNS, PINNACLE SETS, AND BACKBONES OF BIPARTITE PROJECTIONS

By

Rachel Domagalski

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Mathematics – Doctor of Philosophy

2021

ABSTRACT

ON PERMUTATION PATTERNS, PINNACLE SETS, AND BACKBONES OF BIPARTITE PROJECTIONS

By

Rachel Domagalski

This dissertation encompasses the study of two different fields, one regarding permutations including pattern containment and pinnacle sets, and the other on weighted networks, specifically bipartite projections and their backbones.

The study of pattern containment and avoidance for linear permutations is a well-established area of enumerative combinatorics. A cyclic permutation is the set of all rotations of a linear permutation. Callan initiated the study of permutation avoidance in cyclic permutations and characterized the avoidance classes for all single permutations of length 4. We continue this work. In particular, we establish a cyclic variant of the Erdős-Szekeres Theorem that any linear permutation of length $mn + 1$ must contain either the increasing pattern of length $m + 1$ or the decreasing pattern of length $n + 1$. We then derive results about avoidance of multiple patterns of length 4. We also determine generating functions for the cyclic descent statistic on these classes.

We then study the pinnacle set, which is the value analogue of a well-studied permutation statistic, the peak set. Let $\pi = \pi_1\pi_2 \ldots \pi_n$ be a permutation in the symmetric group $\mathfrak{S}_n$ written in one-line notation. The pinnacle set of π, denoted $\operatorname{Pin} \pi$, is the set of all π_i such that $\pi_{i-1} < \pi_i > \pi_{i+1}$. The classic peak set statistic consists of the positions of these values. The pinnacle set was introduced by Davis, Nelson, Petersen, and Tenner who showed that it has many interesting properties. In particular, they proved that the number of subsets of $[n] = \{1, 2, \ldots, n\}$ which can be the pinnacle set of some permutation is a binomial coefficient. Their proof involved a bijection with lattice paths and was somewhat involved. We give a simpler demonstration of this result which does not need lattice paths. Moreover, we show that our map and theirs are different descriptions of the same function. Davis et al. also studied the number of pinnacle sets with maximum m and cardinality d which they denoted by $\mathfrak{p}(m, d)$. We show that these integers are the well-known ballot

numbers and give two proofs of this fact: one using finite differences and one bijective. Diaz-Lopez, Harris, Huang, Insko, and Nilsen found a summation formula for calculating the number of permutations in $\mathfrak{S}_n$ having a given pinnacle set. We derive a new expression for this number which is faster to calculate in many cases. We also show how this method can be adapted to find the number of orderings of a pinnacle set which can be realized by some $\pi \in \mathfrak{S}_n$. This concludes our research on permutations.

Bipartite projections are used in a wide range of network contexts including politics (bill co-sponsorship), geography (firm co-location), genetics (gene co-expression), economics (executive board co-membership), and innovation (patent co-authorship). However, because bipartite projections are always weighted graphs, which are inherently challenging to analyze and visualize, it is often useful to examine the 'backbone,' an unweighted subgraph containing only the most significant edges. We introduce the R package `backbone` for extracting the backbone of weighted bipartite projections, and use two empirical datasets to demonstrate its functionality, bill sponsorship data from the 114$^{\text{th}}$ session of the United States Senate and a Globalization and World Cities data set regarding firm locations in 2000.

After introducing and demonstrating five different models for backbone extraction, the fixed fill model (FFM), fixed row model (FRM), fixed column model (FCM), fixed degree sequence model (FDSM), and stochastic degree sequence model (SDSM), we compare them in terms of accuracy, speed, statistical power, similarity, and community detection. Here, we aim to find which models perform similarly to FDSM, since the FDSM model controls for both degree sequences exactly. We find that the computationally-fast SDSM offers a statistically conservative but close approximation of the computationally-impractical FDSM under a wide range of conditions, and that it correctly recovers a known community structure even when the signal is weak. Therefore, although each backbone model may have particular applications, we recommend SDSM for extracting the backbone of most bipartite projections.

ACKNOWLEDGEMENTS

First I would like to thank my advisor, Dr. Bruce Sagan, for all his help and support since I arrived at MSU. You have always made me feel like I can do and achieve anything and have helped me feel capable in my skills as a mathematician. I am so grateful for your advice and kindness. To Dr. Zachary Neal, thank you for taking me under your wing and into your research projects. You have shown me a new way of applying mathematics and it's been so much fun and so exciting working with you. Thank you for your support and trust. Without both of your encouragement and guidance, this thesis would not be possible.

To my committee, Drs. Robert Bell, Peter Magyar, and Elizabeth Munch, thank you so much for dedicating your time to preparing me to become a mathematician worthy of this degree. To Dr. Sivaram Narayan, thank you for your guidance and inspiring me to pursue graduate school. I greatly appreciate your wisdom, support, and friendship through my undergrad, masters, PhD, and beyond. My collaborators and friends, Jinting Liang, Quinn Minnich, Jamie Schmidt, Alex Sietsema, and Xiaoqin Yan, thank you for your encouragement and knowledge. It's been a pleasure working with you all both in-person and online. You are all destined for great things, and I can't wait to see what you accomplish.

Davis, words cannot even begin to come close to cover how grateful I am for your support over the past decade. Through every life step we've taken, you've always encouraged me to follow every dream and passion. I love you. Here's to our next chapter with the beautiful family we've made. Which brings me to our dogs, Atlas and Tsuki. You two are magic. All extra hours post-graduation go to you my loves.

Mom and Dad, thank you for instilling a love of math and science and discovery in me from day one. I'm so lucky to have such wonderful people as parents. Thank you for being my biggest cheerleaders and for all of your love. Steven, my best friend since birth. Thank you for always being there for me, no matter how many miles between us. You inspire me every single day. I love you all and can't wait to be together again.

To my cohort and beyond, I couldn't have done this without you. The laughter, the late nights, the self-deprecating humor, the game nights, the life talks, the tree climbing, the walks to the river, the El Oasis taco truck trips, the downs, and the so so many ups, I love you all and miss seeing you daily. To my friends, especially Emilee, Nikki, Olivia, Brooke, Paige, and to my extended family, thank you for your support, love, and all the joy we have shared.

Finally, thank you to the faculty and staff members of the mathematics department at Michigan State University for creating such a wonderful home during my time here, and those at Central Michigan University and Holly High School who got me here in the first place.

Portions of Chapter 3 appear in "Domagalski, R., Liang, J., Minnich, Q., Sagan, B. E., Schmidt, J., & Sietsema, A. (2021). Cyclic Pattern Containment and Avoidance. ArXiv:2106.02534 [Math]. http://arxiv.org/abs/2106.02534" and are reprinted here under a CC BY 4.0 license.

Portions of Chapter 4 appear in "Domagalski, R., Liang, J., Minnich, Q., Sagan, B. E., Schmidt, J., & Sietsema, A. (2021). Pinnacle Set Properties. ArXiv:2105.10388 [Math]. http://arxiv.org/abs/2105.10388" and are reprinted here under a CC BY 4.0 license.

Portions of Chapter 7 were originally published in "Domagalski, R., Neal, Z. P., & Sagan, B. (2021). Backbone: An R package for extracting the backbone of bipartite projections. Plos one, 16(1), e0244363," reprinted here under a CC BY 4.0 license, and in "Neal, Z. P., Domagalski, R., & Sagan, B. (2021). Analysis of Spatial Networks From Bipartite Projections Using the R Backbone Package. Geographical Analysis. https://doi.org/10.1111/gean.12275 [NDS21a]," and "Neal, Z. P., Domagalski, R., & Yan, X. (2022). Homophily in collaborations among US House Representatives, 1981–2018. Social Networks, 68, 97–106. https://doi.org/10.1016/j.socnet.2021.04.007," both reprinted with journal permissions.

Portions of Chapters 6 and 8 originally appeared in "Neal, Z. P., Domagalski, R., & Sagan, B. (2021). Comparing Models for Extracting the Backbone of Bipartite Projections. arXiv preprint arXiv:2105.13396." They are reprinted here under a CC BY-SA 4.0 license.

The work in Chapters 6-8 was supported by funding from the National Science Foundation (#1851625 & #2016320) and Michigan State University Center for Business and Social Analytics.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

This doctoral thesis is the culmination of two combinatorial projects. The first explores permutation patterns and statistics, specifically looking at pattern containment and avoidance of cyclic permutations, and generating functions of cyclic descent statistics. Additionally, we study a particular permutation statistic, the pinnacle set. The second project involves bipartite projections, a type of weighted graph. When a weighted graph represents a social relationship, it is of interest to know whether an edge weight should be considered particularly strong or weak. We provide various probabilistic null models to which one can compare an edge weight to determine its statistical significance. Edges deemed significant are part of the backbone subgraph.

The initial three chapters will describe the project on permutations, beginning with background information in chapter 2, then discussing permutation patterns and avoidance in chapter 3, and finally pinnacle set properties in chapter 4. We begin by expanding on the well-studied field of pattern avoidance in linear permutations by considering its implications in cyclic permutations. Specifically, we begin chapter 3 by proving a cyclic variant of the Erdős-Szekeres theorem in section 3.1. This new theorem states that in any cyclic permutation of size $mn+2$, there is either an increasing subsequence of length $m+2$ or a decreasing subsequence of length $n+2$. This theorem becomes of great use in our study of length four pattern avoidance in sections 3.2 and 3.3. While linear pattern avoidance has origins reaching back to the early 1900's, the study of cyclic pattern avoidance was introduced relatively recently by Callan [Cal02] in 2002. He was able to count the number of cyclic permutations that avoid single patterns of length four (length three pattern avoidance being relatively trivial). We complete this study of length four pattern avoidance by counting the number of cyclic permutations that avoid any set of length four patterns, specifically providing proof for all pairs and triples. These proofs utilize the proof technique of generating trees. As the cardinality of the set of patterns increases, the number of permutations that avoid the set decreases. These results allow us to completely count all avoidance sets of any size of length four patterns. After

this classification, we discuss cyclic descent generating functions in section 3.4. These generating functions allow us to count the numbers of cyclic descents in permutations that avoid a given set of patterns, refining our enumerations of the avoidance classes. Chapter 3 is concluded by a section on open problems raised within this work, namely now that patterns of length three and four are characterized, future projects could include looking for enumerative formulas for patterns of length five and higher. It is also of interest to look at the generating functions for other permutation statistics over the avoidance classes. We provide one result which counts the joint distribution of cyclic descents and cyclic peaks. Additionally vincular pattern avoidance can be studied. In this scenario, occurrences of the pattern in a permutation may require different elements to be adjacent to one another. We conjecture an exponential generating function which will count the number of permutations that avoid $\overline{123}$ and $\overline{213}$, concluding chapter 3. Recently, Sergi Elizalde and Bruce Sagan have proven this conjecture [ES21].

Using the background on permutations and permutation statistics presented in chapter 2, in chapter 4 we will explore the pinnacle set of a permutation and prove a number of results related to counting either the number of pinnacle sets or the number of permutations with a given pinnacle set. In section 4.1, we reprove a result of [DNKPT18] that counts the number of pinnacle sets. Their proof involved lattice paths and was somewhat complicated, while ours is a simpler demonstration that does not need lattice paths. In fact, we show that our map and theirs are different descriptions of the same function. We then turn our attention to counting pinnacle sets with a defined maximum and size in section 4.2. While [DNKPT18] proved these counts satisfied a nice recurrence, they did not provide a formula to find the exact count. We show that these counts are actually just ballot numbers, and do this in two ways: using the theory of finite differences and via a bijection. Since we now have counts of the number of pinnacle sets of given sizes, it is natural to turn one's attention to counting the number of permutations with a given pinnacle set. We address this area in section 4.3. While a summation formula that counts such permutations was given in [DLHH$^+$21], we construct a new formula that is more computationally efficient in many cases. We also show how this formula can be modified to answer a similar question: how many admissible orderings of a pinnacle set are

there? Both of the enumerations found in this section have been of great interest to the research community in recent weeks, and we conclude this chapter by describing the recent progress made in constructing even faster formulas in section 4.4, which completes our study of permutations.

The remaining chapters will discuss the backbone of a weighted network. We begin by introducing the concept of bipartite projections and backbone extraction in chapter 5. While bipartite networks are used to describe and represent a wide range of scenarios, their projections are challenging to analyze as they are dense and weighted. In addition, the projection loses information about the original row and column degree sequences of the bipartite network. Ideally, we'd like to reduce the complexity of these networks to a backbone network that contains only the most important edges. The edges retained should be those that had a higher or lower weight than would be expected in a random scenario. To find these backbone networks, we introduce five different bipartite ensemble backbone models in chapter 6. Each of the different bipartite ensemble models constrain the degree sequences of the set of all bipartite networks to which we compare our data. We prove the probability mass functions for the stochastic degree sequence model (SDSM), fixed row model (FRM), fixed column model (FCM), and fixed fill model (FFM). The FDSM is considered the 'gold standard' model as it exactly fixes both degree sequences. However, its distribution remains unknown, and therefore we must approximate it through Monte Carlo methods.

While methods for backbone extraction including a few of the ones mentioned above have existed in the literature for several years, there did not exist one central software package or program where they were all implemented. This meant that researchers who wanted to find a backbone of their network would have to first find which method they wanted to use, potentially guessing which was best for their purposes, and then see if the algorithm was already implemented or available for use. To increase the ease of access for backbone methods, we've implemented the SDSM, FDSM, FRM, FCM, and FFM in the new R package `backbone`. The package and its usage are described in chapter 7. To demonstrate how to use `backbone`, we apply the functions to two different data sets, a legislative network and a spatial network. Through implementing the R package and increasing its user base, we're often met with the same question from researchers: "which model should be

used for my data?" This is the question we investigate in chapter 8.

In chapter 8 we consider each of the five aforementioned models and compare their accuracy, speed, statistical power, similarity, and community detection. These analyses are conducted in four studies. In section 8.1, we evaluate the accuracy and speed of different approaches for estimating cell-filling probabilities used by the SDSM. In section 8.2, we evaluate the statistical power of the SDSM relative to the FDSM. In section 8.3, we examine how degree distributions impact the similarity of backbones extracted using different models. In section 8.4, we examine the extent to which backbones extracted using different models accurately recover a known community structure. Finally, we conclude in section 8.5 with recommendations for backbone model selection and opportunities for future model development.

CHAPTER 2

BACKGROUND ON PERMUTATION PATTERNS AND STATISTICS

We begin by reviewing some notions from the well-studied theory of patterns in (linear) permutations. We then discuss permutation statistics and generating functions for cyclic descents. We'll finish by exploring what is known about the pinnacle set. The pinnacle set is the value analogue of a particular permutation statistic, the peak set. More information on the topic of patterns in permutations can be found in the texts of Bóna [Bón04], Sagan [Sag20], or Stanley [Sta97, Sta99].

Let $\mathbb{N}$ and $\mathbb{P}$ be the nonnegative and positive integers, respectively. If $m, n \in \mathbb{N}$ then we define $[m, n] = \{m, m+1, \ldots, n\}$; if $m = 1$ we then abbreviate to $[n] = [1, n]$. Consider the symmetric group $\mathfrak{S}_n$ of all permutations $\pi = \pi_1 \pi_2 \ldots \pi_n$ of $[n]$ written in one-line notation. We call n the *length* of π and write $|\pi| = n$. We will also use this notation to represent the cardinality of a set, where the difference should be clear by context. We will sometimes put commas between the elements of π for readability. We say that two sequences of distinct integers $\pi = \pi_1 \ldots \pi_k$ and $\sigma = \sigma_1 \ldots \sigma_k$ are *order isomorphic*, written $\pi \cong \sigma$, whenever $\pi_i < \pi_j$ if and only if $\sigma_i < \sigma_j$. If $\sigma \in \mathfrak{S}_n$ and $\pi \in \mathfrak{S}_k$ then σ *contains* π *as a pattern* if there is a subsequence σ' of σ with $|\sigma'| = k$ and $\sigma' \cong \pi$. If no such subsequence exists then σ *avoids* π. We use the notation

$$\mathrm{Av}_n(\pi) = \{\sigma \in \mathfrak{S}_n \mid \sigma \text{ avoids } \pi\}$$

for the *avoidance class* of π. For example $\sigma = 42351$ contains the pattern $\pi = 3241$ because of the subsequence 4251, among others. But it avoids 1234 because it has no increasing subsequence of length 4. One can extend this notion to sets of permutations Π by letting

$$\mathrm{Av}_n(\Pi) = \{\sigma \in \mathfrak{S}_n \mid \sigma \text{ avoids all } \pi \in \Pi\} = \bigcap_{\pi \in \Pi} \mathrm{Av}_n(\pi).$$

A famous theorem of Erdős and Szekeres [ES35] can be stated in terms of pattern containment and avoidance. Let

$$\iota_n = 12 \ldots n \qquad \text{and} \qquad \delta_n = n \ldots 21$$

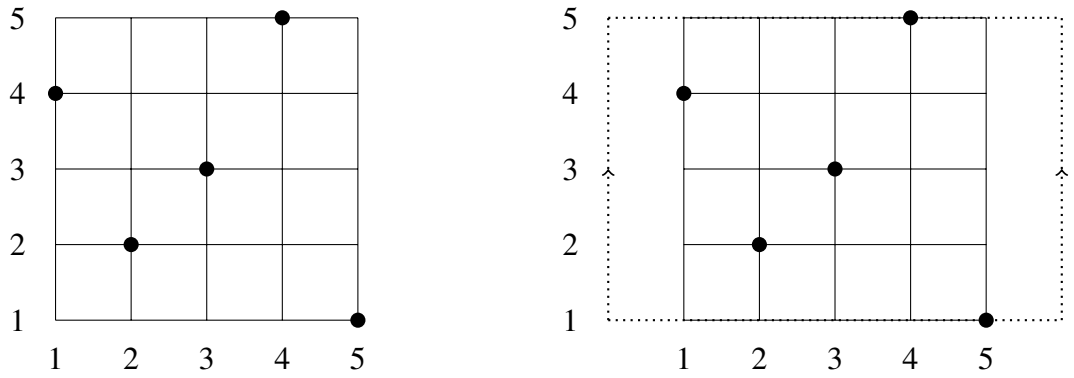

Figure 2.1: The graph of 42351 on the left and of [42351] on the right

be the increasing and decreasing permutations of length n, respectively.

Theorem 2.0.1 ([ES35])**.** *Suppose $m, n \in \mathbb{N}$. Then any $\sigma \in \mathfrak{S}_{mn+1}$ contains either ι_{m+1} or δ_{n+1}. This is the best possible in that there exist permutations in $\mathfrak{S}_{mn}$ which avoid both ι_{m+1} and δ_{n+1}.* □

The *diagram* of $\pi \in \mathfrak{S}_n$ is the collection of points (i, π_i) in the first quadrant of the Cartesian plane. The graphical representation of $\pi = 42351$ is given on the left in Figure 2.1. It follows that we can act on π with the dihedral group of the square

$$D_4 = \{\rho_0, \rho_{90}, \rho_{180}, \rho_{270}, r_0, r_1, r_{-1}, r_\infty\}$$

where ρ_θ is rotation counterclockwise through θ degrees and r_m is reflection in a line of slope m. We wish to write some of these rigid motions in terms of the one-line notation for $\pi = \pi_1\pi_2 \dots \pi_n$. Reflection in a vertical line gives the *reversal* of π which is

$$\pi^r = \pi_n \dots \pi_2\pi_1.$$

Similarly, reflection in a horizontal line results in the *complement* of π

$$\pi^c = n + 1 - \pi_1, \quad n + 1 - \pi_2, \quad \dots, \quad n + 1 - \pi_n.$$

Combining these two operations gives rotation by 180 degree or *reverse complement*

$$\pi^{rc} = n + 1 - \pi_n, \quad \dots, \quad n + 1 - \pi_2, \quad n + 1 - \pi_1.$$

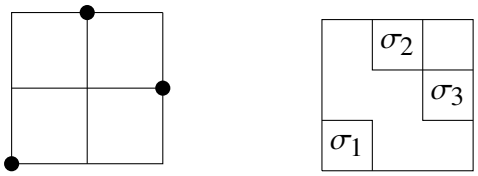

Figure 2.2: The diagram of 132 (left) and $132\langle\sigma_1, \sigma_2, \sigma_3\rangle$ (right)

We apply any of these operations to sets of permutations by applying them to each element of the set.

We can use diagrams to inflate permutations. If we are given $\pi = \pi_1\pi_2\ldots\pi_n \in \mathfrak{S}_n$ and permutations $\sigma_1, \sigma_2, \ldots, \sigma_n$ then the *inflation of π by the σ_i* is the permutation $\pi\langle\sigma_1, \sigma_2, \ldots, \sigma_n\rangle$ whose diagram is obtained from that of π by replacing each vertex (i, π_i) by a copy of σ_i. For example, given $\pi = 132$ and $\sigma_1, \sigma_2, \sigma_3$ then a schematic of the diagram of $132\langle\sigma_1, \sigma_2, \sigma_3\rangle$ is given on the right in Figure 2.2. More concretely, if $\sigma_1 = 21$, $\sigma_2 = 1$, and $\sigma_3 = 213$ then

$$132\langle\sigma_1, \sigma_2, \sigma_3\rangle = 216435.$$

We say that patterns π and π' are *Wilf equivalent*, written $\pi \equiv \pi'$, if $\#\operatorname{Av}_n(\pi) = \#\operatorname{Av}_n(\pi')$ for all $n \in \mathbb{N}$, where the hash symbol denotes cardinality. This definition extends in the obvious way to sets of patterns. Note that if π and π' are Wilf equivalent then both must be in the same $\mathfrak{S}_n$. It is easy to see that if $\phi \in D_4$ then $\pi \equiv \phi(\pi)$ and so these are called *trivial Wilf equivalences*. It is well known that all elements of $\mathfrak{S}_3$ are Wilf equivalent.

Theorem 2.0.2. *If $\pi \in \mathfrak{S}_3$ then*

$$\#\operatorname{Av}_n(\pi) = C_n$$

where $C_n = \frac{1}{n+1}\binom{2n}{n}$ is the nth Catalan number. □

Trivial Wilf equivalence carries over to sets Π of permutations. Simion and Schmidt [SS85] determined all Wilf equivalences among the $\operatorname{Av}_n(\Pi)$ for all $\Pi \subseteq \mathfrak{S}_3$.

A *permutation statistic* is a map $\mathrm{st} : \uplus_{n\geq 0}\mathfrak{S}_n \to S$ where S is some set. Famous permutation statistics include the *descent set* statistic

$$\operatorname{Des}\pi = \{i \mid \pi_i > \pi_{i+1}\},$$

where the elements $i \in \operatorname{Des}\pi$ are called *descents* and if $\pi_i < \pi_{i+1}$ then i is called an *ascent*, the *descent number* statistic

$$\operatorname{des}\pi = \#\operatorname{Des}\pi,$$

the *major index* statistic

$$\operatorname{maj}\pi = \sum_{i\in\operatorname{Des}\pi} i,$$

the *inversion* statistic

$$\operatorname{inv}\pi = \#\{(i,j) \mid i < j \text{ and } \pi_i > \pi_j\},$$

the *excedance statistic*

$$\operatorname{exc}\pi = \#\{i \mid \pi(i) > i\},$$

and the *peak set statistic*

$$\operatorname{Pk}\pi = \{i \mid \pi_{i-1} < \pi_i > \pi_{i+1}\}.$$

Returning to the example given in fig. 2.1, the permutation $\pi = 42351$ has $\operatorname{Des}\pi = \{1,4\}$, $\operatorname{des}\pi = 2$, $\operatorname{maj}\pi = 5$, $\operatorname{inv}\pi = 6$, and $\operatorname{exc}\pi = 2$, and $\operatorname{Pk}\pi = \{4\}$.

Let st be a statistic whose range is $\mathbb{N}$ and let q be a variable. If Π is a set of patterns then its avoidance class has a corresponding generating function

$$F_n^{\mathrm{st}}(\Pi) = F_n^{\mathrm{st}}(\Pi; q) = \sum_{\sigma\in\operatorname{Av}_n(\Pi)} q^{\mathrm{st}\,\sigma}.$$

Say that Π and Π' are st-*Wilf equivalent* and write $\Pi \overset{\mathrm{st}}{\equiv} \Pi'$ if $F_n^{\mathrm{st}}(\Pi) = F_n^{\mathrm{st}}(\Pi')$ for all $n \geq 0$. Clearly st-Wilf equivalence implies Wilf equivalence. The maj- and inv-Wilf equivalence classes for $\Pi \subseteq \mathfrak{S}_3$ were determined by Dokos, Dwyer, Johnson, Sagan, and Selsor [DDJ$^+$12].

If $\pi = \pi_1\pi_2\ldots\pi_n \in \mathfrak{S}_n$ then the corresponding *cyclic permutation* is the set of all rotations of π, denoted

$$[\pi] = \{\pi_1\pi_2\ldots\pi_n,\ \ \pi_2\ldots\pi_n\pi_1,\ \ \ldots,\ \ \pi_n\pi_1\ldots,\pi_{n-1}\}.$$

Continuing our example from the beginning of the section,

$$[42351] = \{42351,\ \ 23514,\ \ 35142,\ \ 51423,\ \ 14235\}.$$

If necessary, we will call permutations from $\mathfrak{S}_n$ *linear* to distinguish them from their cyclic cousins. We also use square brackets to denote cyclic analogues of objects defined in the linear case. For example, $[\mathfrak{S}_n]$ is the set of all cyclic permutations of length n. We say a cyclic permutation $[\sigma]$ *contains* $[\pi]$ *as a pattern* if there is some rotation σ' of σ which contains π linearly. Otherwise $[\sigma]$ *avoids* $[\pi]$. In our perennial example, even though 42351 avoids 1234 we have that $[42351]$ contains $[1234]$ since the rotation 14235 has the copy 1235 of this pattern. Given a set $[\Pi]$ of cyclic patterns the cyclic avoidance class $\mathrm{Av}_n[\Pi]$ is defined as expected. Note that when using a specific set of cyclic permutations the square brackets will be put around the permutations themselves, for example, $\mathrm{Av}_n([\pi], [\pi'])$. Callan [Cal02] determined $\#\,\mathrm{Av}_n[\pi]$ for all $[\pi] \in [\mathfrak{S}_4]$. Gray, Lanning, and Wang continued work in this direction considering cyclic packing of patterns [GLW18] and patterns in colored cyclic permutations [GLW19].

The graph of a cyclic permutation $[\pi]$ is obtained by embedding the graph of π on a cylinder. This is indicated on the right in Figure 2.1 by identifying the two dotted arrows. Cyclic Wilf equivalence has the obvious definition. But note that now there are fewer trivial cyclic Wilf equivalences since we need the chosen group element to preserved the cylinder, not just the square. So the only trivial equivalences are

$$[\pi] \equiv [\pi^r] \equiv [\pi^c] \equiv [\pi^{rc}]. \tag{2.1}$$

Certain linear permutation statistics have obvious cyclic analogues. For example, if $\pi \in \mathfrak{S}_n$ then its *cyclic descent number* is

$$\mathrm{cdes}[\pi] = \#\{i \mid \pi_i > \pi_{i+1} \text{ where subscripts are taken modulo } n\}.$$

Note that this is well defined because the cardinality does not depend on which representative of $[\pi]$ is chosen. To illustrate, $\pi = 23514$ has cyclic descents at indices 3 and 5 so $\mathrm{cdes}[\pi] = 2$. The

corresponding generating function $F_n^{\text{cdes}}[\Pi]$ where $[\Pi]$ is a set of cyclic permutations, and cdes-Wilf equivalence should now need no definition. Note that cdes is another form of the excedance statistic on linear permutations. In particular, if $\pi = \pi_1\pi_2\ldots\pi_n$ then

$$\text{cdes}[\pi] = \text{exc}(\pi_n, \ldots, \pi_2, \pi_1)$$

where $(\pi_n, \pi_{n-1}\ldots, \pi_1)$ is cycle notation for the linear permutation which, as a function, sends π_i to π_{i-1} for all i modulo n.

We return our attention to the *peak set* statistic on linear permutations,

$$\text{Pk}\,\pi = \{i \mid \pi_{i-1} < \pi_i > \pi_{i+1}\} \subseteq [2, n-1].$$

For example, if $\pi = 18524376$ then $\text{Pk}\,\pi = \{2, 5, 7\}$ since $\pi_2 = 8$, $\pi_5 = 4$, and $\pi_7 = 7$ are all bigger than the elements directly to their left and right. It is easy to see that $S \subseteq [2, n-1]$ is the peak set of some $\pi \in \mathfrak{S}_n$ if and only if no two elements of S are consecutive. So the number of possible peak sets is a Fibonacci number. One could also ask how many permutations have a given peak set. This question was answered by Billey, Burdzy and Sagan.

Theorem 2.0.3 ([BBS13])**.** *If $n \in \mathbb{P}$ and $S \subseteq [2, n]$ then*

$$\#\{\pi \mid \text{Pk}\,\pi = S\} = p(S; n)2^{n-\#S-1}$$

where # denotes cardinality and $p(S; n)$ is a polynomial in n depending on S. □

It is natural to study the values at the peak indices. This line of research was initiated by Davis, Nelson, Petersen, and Tenner [DNKPT18] and continued by Rusu [Rus20]; Diaz-Lopez, Harris, Huang, Insko, and Nilsen [DLHH$^+$21]; and Rusu and Tenner [RT]. Define the *pinnacle set* of a permutation $\pi \in \mathfrak{S}_n$ to be

$$\text{Pin}\,\pi = \{\pi_i \mid \pi_{i-1} < \pi_i > \pi_{i+1}\} \subseteq [3, n]$$

Continuing with the example $\pi = 18524376$ we see that $\text{Pin}\,\pi = \{4, 7, 8\}$. Following Davis et al., call a set S an *admissible pinnacle set* if there is some permutation π with $\text{Pin}\,\pi = S$. They found a

criterion for S to be admissible which will be useful in this work. This result was stated in recursive fashion, but it is clearly equivalent to the following non-recursive version.

Theorem 2.0.4 ([DNKPT18]). *Let $S = \{s_1 < s_2 < \ldots < s_d\} \subset \mathbb{P}$. The set S is an admissible pinnacle set if and only if we have*

$$s_i > 2i$$

for all $i \in [d]$. □

Davis et al. were able to count the number of admissible pinnacle sets for $\pi \in \mathfrak{S}_n$.

Theorem 2.0.5 ([DNKPT18]). *If*

$$\mathcal{A}_n = \{S \mid S = \operatorname{Pin} \pi \text{ for some } \pi \in \mathfrak{S}_n\}$$

then

$$\#\mathcal{A}_n = \binom{n-1}{\left\lfloor \frac{n-1}{2} \right\rfloor}.$$ □

They also studied the more refined constants

$$\mathfrak{p}(m, d) = \#\{S \in \mathcal{A}_n \mid \max S = m \text{ and } \#S = d\}$$

where $n \geq m$. Note that if $S = \operatorname{Pin} \pi$ for some $\pi \in \mathfrak{S}_n$ then S is also a pinnacle set of some $\pi' \in \mathfrak{S}_{n'}$ for all $n' \geq n$ since one can just add values larger than n to the beginning of π in decreasing order. It follows that the exact value of n does not play a role in the definition of $\mathfrak{p}(m, d)$.

A number of questions have been raised about pinnacle sets. For example, if

$$p_S(n) = \#\{\pi \in \mathfrak{S}_n \mid \operatorname{Pin} \pi = S\}$$

then how can one compute these numbers as there does not seem to be an analogue of Theorem 2.0.3 in the context of pinnacles. Davis et al. gave a recursive procedure for doing so, and then a non-recursive summation formula for determining the $p_S(n)$ was proposed in the paper of Diaz-Lopez et al.

Another problem suggested earlier is as follows. Given an admissible S, a permutation σ of S is called an *admissible ordering* if there is a $\pi \in \mathfrak{S}_n$ with $\operatorname{Pin} \pi = S$ and the pinnacles of π occur in the same order as they do in σ. Let

$$O(S) = \{\sigma \mid \sigma \text{ is an admissible ordering of } S\}.$$

For example, if $S = \{3, 5, 7\}$ then $\sigma = 537 \in O(S)$ as witnessed by $\pi = 4513276$. But $375 \notin O(S)$ since in order for 6 not to be a pinnacle, it must be directly to the left or right of 7 and both choices lead to a contradiction. The set O_S was studied in the articles of Rusu, and of Rusu and Tenner [Rus20, RT]. In the latter paper, the authors asked for a function to compute $\#O(S)$.

With these definitions and results in hand, we first examine cyclic pattern containment and avoidance in the following chapter 3. We begin by proving a cyclic version of the Erdős-Szekeres Theorem 3.1.1 in section 3.1. This result is used to help us count $\#\operatorname{Av}_n([\Pi])$ in sections 3.2 and 3.3, where $[\Pi] \subseteq [\mathfrak{S}_4]$ and $\#[\Pi] \geq 2$. We then consider cyclic descent generating functions over $\operatorname{Av}_n([\Pi])$ in section 3.4, and find $F_n^{\text{cdes}}[\Pi]$ for $\#[\Pi] = 1, 2$ and $[\Pi] \subseteq [\mathfrak{S}_4]$.

We then continue to our study of pinnacle sets in chapter 4. We begin by counting the number of admissible pinnacle sets in section 4.1. This quantity, given in theorem 2.0.5, was already found in [DNKPT18]. Here we provide a simpler proof using a bijection using interleaved and right canonical permutations. As mentioned, [DNKPT18] also studied the values $\mathfrak{p}(m; d)$. In section 4.2, we show these constants are actually ballot numbers, specifically $\mathfrak{p}(m; d) = \frac{m-2d+1}{m-1}\binom{m-1}{d-1}$. We do this in two ways, using finite differences and a bijection. Once we've counted the number of admissible pinnacle sets, we consider the number of permutations with a given pinnacle set in section 4.3. We provide a sum to count $p_S(n)$ in theorem 4.3.1 which is asymptotically more efficient than previously existing methods. We then extend this result to count $\#O(S)$ in theorem 4.3.12.

CHAPTER 3

CYCLIC PATTERN CONTAINMENT AND AVOIDANCE

This chapter contains material from Domagalski, Liang, Minnich, Sagan, Schmidt, and Sietsema [DLM+21a]. All results in this chapter come from this manuscript except as otherwise noted.

3.1 A cyclic Erdős-Szekeres Theorem

In this section we will use the linear Erdős-Szekeres Theorem to prove a cyclic analogue. We will need a variant of the decreasing permutation δ_n defined as follows. Given nonnegative integers n (the length), d (the difference), and s (the smallest value) define the decreasing sequence

$$\delta_{n,d,s} = s + (n-1)d, \quad s + (n-2)d, \quad \ldots, \quad s + d, \quad s.$$

For example

$$\delta_{5,2,3} = 11, 9, 7, 5, 3.$$

Theorem 3.1.1. *Suppose $m, n \in \mathbb{N}$. Then any $[\sigma] \in [\mathfrak{S}_{mn+2}]$ contains either $[\iota_{m+2}]$ or $[\delta_{n+2}]$. This is the best possible in that there exist permutations in $[\mathfrak{S}_{mn+1}]$ which avoid both $[\iota_{m+2}]$ and $[\delta_{n+2}]$.*

Proof. To prove the first statement we can assume, by rotating π if necessary, that

$$\sigma = \sigma_1, \sigma_2, \ldots, \sigma_{mn+1}, mn + 2.$$

So $\sigma' = \sigma_1\sigma_2 \ldots \sigma_{mn+1} \in \mathfrak{S}_{mn+1}$ and, by Theorem 2.0.1, contains a copy κ of either ι_{m+1} or δ_{n+1}. In the first case, the concatenation $\kappa, mn + 1$ is a copy of $[\iota_{m+2}]$ in $[\pi]$. In the second case, we have that $mn + 1, \kappa$ is a copy of $[\delta_{n+2}]$ in $[\sigma]$.

To prove the second statement, consider the concatenation

$$\sigma = 1,\ \delta_{n,m,2},\ \delta_{n,m,3},\ \ldots,\ \delta_{n,m,m+1}.$$

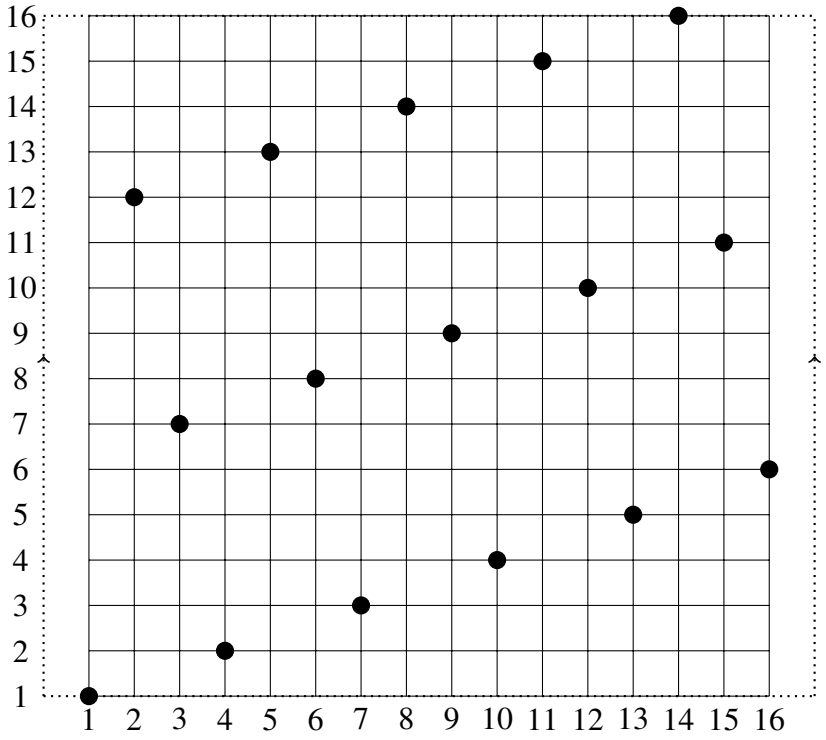

Figure 3.1: The graph of $[\sigma]$ when $m = 5$ and $n = 3$

For example, when $m = 5$ and $n = 3$, then

$$[\sigma] = [1, 12, 7, 2, 13, 8, 3, 14, 9, 4, 15, 10, 5, 16, 11, 6]$$

whose graph is shown in Figure 3.1. Define σ' by $\sigma = 1\sigma'$ and note that σ' can be written either as a disjoint union of m decreasing subsequences of length n, or of n increasing subsequences of length m. In a linear permutation, any increasing subsequence can intersect any decreasing subsequence at most once. So any increasing subsequence of σ' has length at most m, and any decreasing subsequence has length at most n. Now let $[\pi]$ be a subsequence of $[\sigma]$. We consider two cases.

Suppose first that $[\pi]$ contains 1. If $[\pi]$ is increasing then rotate, if necessary, until $\pi = 1\pi'$ for some π' which is a subsequence of σ'. But from the previous paragraph, $|\pi'| \leq m$ which implies $|\pi| \leq m + 1$ as desired. If $[\pi]$ is decreasing then we pick a representative $\pi = \pi'1$ and proceed as in the increasing case to get $|\pi| \leq n + 1$.

Now consider the possibility that $[\pi]$ does not contain 1. Again, we start with the subcase when

$[\pi]$ is increasing. Suppose, for simplicity, that π contains an element of $x \in \delta_{n,m,2}$ as the proof will be similar for the other deltas. As before, π can contain at most one element of each of $\delta_{n,m,3}$ through $\delta_{n,m,m+1}$. Now $[\pi]$ can wrap around and pick up other elements. But those elements must come before x. And since $\delta_{n,m,2}$ is decreasing, at most one other element can be added in this way. It follows that $|\pi| \leq m + 1$. On the other hand, if $[\pi]$ is decreasing then the proof is similar. The only difference is that if one attempts to pick up elements of $\delta_{n,m,2}$ before x then this is impossible since such elements are larger than x and $[\pi]$ is decreasing. So $|\pi| \leq n$ which is an even tighter bound. This completes the demonstration of the theorem. □

We'll see the advantages that Theorem 3.1.1 brings in our study of length four pattern avoidance, specifically, when considering $\#\operatorname{Av}_n(S)$ where S contains $[1234]$ and $[1432]$.

3.2 Pattern avoidance of doubletons

In this section we will enumerate $\operatorname{Av}_n[\Pi]$ for all $[\Pi] \subset [\mathfrak{S}_4]$ with $\#[\Pi] = 2$. Any cyclic Wilf equivalences stated without proof are trivial.

Let us first dispose of the simplest singleton avoidance classes where $[\pi] \in [\mathfrak{S}_k]$ for $k < 4$. In $[\mathfrak{S}_2]$ there is only one cyclic permutation $[12]$ and it is easy to see that every $[\sigma]$ of length at least 2 contains it. In $[\mathfrak{S}_3]$ there are only the patterns $[123]$ and $[321]$, and these are only avoided by $[\delta_n]$ and $[\iota_n]$, respectively.

Callan [Cal02] enumerated $\operatorname{Av}_n[\pi]$ for any given $[\pi] \in [\mathfrak{S}_4]$. Recall the version of the Fibonacci numbers defined by $F_1 = F_2 = 1$ and $F_n = F_{n-1} + F_{n-2}$ for $n \geq 3$. Unlike the case of linear permutations in $\mathfrak{S}_3$, there are no nontrivial Wilf equivalences.

Theorem 3.2.1 ([Cal02])**.** *For $n \geq 2$ we have*

$$\#Av_n[1234] = \#Av_n[1432] = 2^n + 1 - 2n - \binom{n}{3},$$

$$\#Av_n[1243] = \#Av_n[1342] = 2^{n-1} - n + 1,$$

$$\#Av_n[1324] = \#Av_n[1423] = F_{2n-3}.$$ □

In presenting the enumerations for doubletons, we make the following conventions to facilitate locating a given result. All cyclic patterns will be listed starting with 1. And all sets of cyclic patterns will be given in lexicographic order. We will also use terms like "just before" or "just after" in $[\sigma]$ to refer the left-to-right order on the cylinder of a cyclic permutation in the form of Figure 2.1. For example, in $[\sigma] = [42351]$ the 5 comes just before 1 and the 4 just after. We also say that an element x is between y and z if it is in the subsequence of $[\sigma]$ traversed going left-to-right around the cylinder from y to z. Continuing our example, between 2 and 5 we have 3, while between 5 and 2 we have 1 and 4.

One of our tools will be generating trees. To the best of our knowledge, these trees were introduced by Chung, Grahamm, Hoggatt, and Kleiman [CGHK78] for studying Baxter permutations. Since then, they have become an integral technique in the theory of pattern avoidance [BBMD$^+$02, BM03, Kre00, Wes95, Wes96]. The *generating tree* for an avoidance class $\mathrm{Av}[\Pi]$, denoted $T[\Pi]$, has as its root the permutation $[12]$. The children of any $[\sigma] \in \mathrm{Av}_n[\Pi]$ are all the $[\sigma'] \in \mathrm{Av}_{n+1}[\Pi]$ which can be formed by inserting $n+1$ into one of the spaces of $[\sigma]$. A space, also called a *site*, where insertion of $n+1$ produces a permutation of the avoidance class is called *active* while the other spaces are *inactive*. A useful observation is that if a space is inactive it must be because inserting $n+1$ there results in copy of a forbidden pattern $[\pi]$ where $n+1$ plays the role of the largest element of π. Once we have picked a representative $\sigma = \sigma_1\sigma_2\ldots\sigma_n$ for $[\sigma]$ we will label the spaces as $1, 2, \ldots, n$ left to right where space i comes between σ_i and σ_{i+1}. The nodes for $\mathrm{Av}_n[\Pi]$ will be said to be at *level* n in $T[\Pi]$. We call the number of children of a vertex its *degree* which is denoted $\deg[\sigma]$. Given $d \in \mathbb{N}$, suppose that every cyclic permutation with $\deg[\sigma] = d$ has children of degrees $c_1, c_2, \ldots, c_d$. Then this is denoted by the *production rule*

$$(d) \to (c_1)(c_2)\ldots(c_d).$$

There may be other nodes having some special characteristic X which always produces nodes having characteristics $Y_1, Y_2, \ldots, Y_d$ which correspond to a production rule

$$(X) \to (Y_1)(Y_2)\ldots(Y_d).$$

In particular, the characteristic of being the root of the tree is denote in a production rule by $(*)$. We can also have production rules which mix numbers for degrees and letters for characteristics. If $T[\pi]$ can be characterized by production rules, these can often be used to calculate $\#\operatorname{Av}_n[\Pi]$.

Theorem 3.2.2. *We have*

$$\{[1234],[1243]\} \equiv \{[1234],[1342]\} \equiv \{[1243],[1432]\} \equiv \{[1342],[1432]\}.$$

And for $n \geq 3$

$$\#\operatorname{Av}_n([1234],[1342]) = 2(n-2).$$

Proof. We claim that $T = T([1234],[1342])$ has the following production rules

$$(*) \to (2)(2),$$

$$(1) \to (1),$$

$$(2) \to (1)(2).$$

Once these are proven then the enumeration follows easily since one can inductively show that, for $n \geq 3$, level n consists of two nodes of degree 2 and $2(n-3)$ nodes of degree 1.

It is easy to check the production rule at levels $n = 2$ and 3, so we assume that $n \geq 4$ and also that $[\sigma] \in \operatorname{Av}_n([1234],[1342])$. First of all, note that the site before n is always active. If it were not then the result $[\sigma']$ of inserting $n+1$ would have a copy κ of one of the patterns containing $n+1$. But n can not be in κ since neither of the patterns have 4 followed immediately in the cycle by 3. So replacing $n+1$ by n in κ would give a forbidden pattern in $[\sigma]$ which is a contradiction. Thus every $[\sigma]$ at has at least one child. Also σ has at most two children. For suppose

$$\sigma' = n+1, \rho, n, \tau$$

is the result of inserting $n+1$ in σ. It follows that $|\rho| \leq 1$ since if $\rho \geq 2$ then $[\sigma']$ has a copy of either [4123] or [4213]. Thus $n+1$ must be inserted either directly before n or two elements before n.

Now consider

$$\delta = n, n-1, \ldots, 3, 2, 1, \quad \text{and} \quad \epsilon = n, n-1, \ldots, 3, 1, 2. \tag{3.1}$$

It is easy to check that both sites n and $n-1$ are active in these permutations and so both have degree 2. It is also obvious that if one inserts $n+1$ in site n in either permutation then one gets another permutation of the same form.

From what we have done, we can finish the proof if we show that $\deg[\sigma] = 2$ implies $[\sigma] = [\delta]$ or $[\sigma] = [\epsilon]$. Write

$$\sigma = n\rho m$$

where m is the last element of σ and ρ is everything between n and m. Since $\deg[\sigma] = 2$, site $n-1$ is active and inserting $n+1$ there yields

$$\sigma' = n, \rho, n+1, m.$$

Then $m \leq 2$ since otherwise $[\sigma]$ contains a copy of $[4123]$ or $[4213]$ since $n \geq 4$. In the case $m = 1$ we must have ρ decreasing. For if there is an ascent $x < y$ in ρ then $[\sigma']$ contains $[x, y, n+1, 1]$ which is a copy of $[2341]$, a contradiction. So in this case ρ is decreasing and $\sigma = \delta$. The other possibility is that $m = 2$. This forces the last element of ρ to be 1. For if 1 is elsewhere and x is the last element of ρ then then $[\sigma']$ contains $[1, x, n+1, 2]$ which is contradictory copy of $[1342]$. Similarly to the first case, one can now show that ρ is decreasing and so $\sigma = \epsilon$ as desired. □

Comparing our next result with the previous one will provide our first nontrivial Wilf equivalence.

Theorem 3.2.3. *We have*

$$\{[1234], [1324]\} \equiv \{[1423], [1432]\}.$$

And for $n \geq 3$

$$\#\operatorname{Av}_n([1234], [1324]) = 2(n-2).$$

Proof. Let D stand for the decreasing permutation and E for the decreasing permutation with its largest two elements swapped. We consider the root $[12]$ to be of type D. We will show that $T = T([1234], [1324])$ has production rules

$$(1) \to (1),$$
$$(D) \to (D)(E),$$
$$(E) \to (1)(1).$$

It follows by induction that level $n \geq 3$ of T has a D, an E, and $2(n-3)$ nodes of degree one, proving the theorem.

The same demonstration as in the previous theorem shows that the site before n in any $[\sigma] \in \mathrm{Av}_n([1234], [1324])$ is active. So again, every such permutation has at least one child. Also, every $[\sigma]$ has at most two children. Indeed, write

$$\sigma = 1\sigma_2 \dots \sigma_n \tag{3.2}$$

and put $n+1$ in site $i \geq 3$. Then $1, \sigma_2, \sigma_3, n+1$ is a copy of either 1234 or 1324, another contradiction.

Now consider permutations corresponding to D and E at level n

$$\delta = 1, n, n-1, n-2, n-3, \dots, 2 \quad \text{and} \quad \epsilon = 1, n-1, n, n-2, n-3, \dots, 2. \tag{3.3}$$

It is easy to check that both sites 1 and 2 are active in δ, ϵ. So, by the previous paragraph, they both have degree 2. Furthermore, the two children of δ have the form D and E.

We will be done if we can show that $[\sigma]$ having two children implies $[\sigma] = [\delta]$ or $[\epsilon]$. Write σ as in (3.2). Since the active sites must be 1 and 2, and the site before n must be active, either $\sigma_2 = n$ or $\sigma_3 = n$. If $\sigma_2 = n$ and there is an ascent $x < y$ in the rest of the permutation, then after inserting $n+1$ in position 2 we have $[x, y, n, n+1]$ which is a copy of $[1234]$, a contradiction. So in this case $[\sigma] = [\delta]$. Alternatively, suppose $\sigma_3 = n$. This forces $\sigma_2 = n-1$, since if $\sigma_2 = x < n-1$ then $n-1$ comes after n. But inserting $n+1$ in position 1 gives $[x, n, n-1, n+1]$ which is a copy

of $[1324]$. And similarly to the first case we see that the rest of σ is decreasing. The result is that $[\sigma] = [\epsilon]$. This completes the proof. □

Theorem 3.2.4. *We have*

$$\{[1234], [1423]\} \equiv \{[1324], [1432]\}.$$

And for $n \geq 1$

$$\#\operatorname{Av}_n([1234], [1423]) = 1 + \binom{n-1}{2}.$$

Proof. Suppose $[\sigma] \in \operatorname{Av}_n([1234], [1423])$ and write

$$\sigma = 1\rho n\tau \tag{3.4}$$

where ρ and τ are the subsequences between 1 and n, and between n and 1, respectively. Now ρ and τ must be decreasing since $[\sigma]$ avoids $[1234]$ and $[1423]$, respectively. Furthermore, ρ must consist of consecutive integers since, if not, then we have $x < y < z$ such that $1zxny$ is a subsequence of σ. So $[xnyz]$ is a copy of $[1423]$ in $[\sigma]$, which is a contradiction. Conversely, it is easy to check that if σ has the form (3.4) with ρ and τ decreasing and ρ consecutive then $[\sigma] \in \operatorname{Av}_n([1234], [1423])$. So we have characterized the elements of this class.

To finish the enumeration, if $\rho = \emptyset$ there is one corresponding σ. But if $\rho \neq \emptyset$ then choosing the smallest and largest element of ρ from the elements $2, 3, \ldots, n-1$ completely determines σ. Since these two elements could be equal, we are choosing 2 elements from $n-2$ elements with repetition which is counted by $\binom{n-1}{2}$. □

The following result follows immediately from Theorem 3.1.1

Theorem 3.2.5. *We have*

$$\#\operatorname{Av}_n([1234], [1432]) = 0$$

for $n \geq 6$. □

We now have, by comparison with Theorem 3.2.4, another nontrivial Wilf equivalence.

Theorem 3.2.6. *We have*

$$\{[1243], [1324]\} \equiv \{[1243], [1423]\} \equiv \{[1324], [1342]\} \equiv \{[1342], [1423]\}.$$

And for $n \geq 1$

$$\#\operatorname{Av}_n([1324], [1342]) = 1 + \binom{n-1}{2}.$$

Proof. Take $[\sigma] \in \operatorname{Av}_n([1324], [1342])$ and write σ as in (3.4). Then ρ is increasing since $[\sigma]$ avoids $[1324]$. And every element of ρ is smaller than every element of τ since $[\sigma]$ avoids $[1342]$. To avoid a copy of one of the forbidden patterns containing the 1 of σ we must have that τ avoids 213 and 231. And to avoid a copy of $[1324]$ where n plays the role of 4, it must be that τ avoids 132. The τ which avoid these three pattern are exactly those which are inflations of the form $\tau = 21\langle\delta_k, \iota_l\rangle$ for some $k, l \geq 0$ (see the chart on page 2773 of [DDJ+12]). Absorbing the 1 and n of σ into ρ and τ, respectively, we see that

$$\sigma = 132\langle\iota_j, \delta_k, \iota_l\rangle \tag{3.5}$$

where $j, k \geq 1$ and $l \geq 0$. Again, it is not hard to check that for every σ of this form we have $[\sigma] \in \operatorname{Av}_n([1324], [1342])$.

To enumerate these σ, we distinguish two cases. If $l \geq 2$ then picking the smallest and largest elements of the copy of ι_l from $2, 3, \ldots, n-1$ completely determines σ . So in this case there are $\binom{n-2}{2}$ choices. If $l \leq 1$ then the copy of ι_l can be appended to the copy of δ_k so that $\sigma = 12[\iota_j, \delta_{n-j}]$. Since we must have 1 and n in the ascending and decreasing subsequences, there are now $n-1$ choices. Adding the two counts given the desired result. □

Theorem 3.2.7. *For* $n \geq 4$ *we have*

$$\#\operatorname{Av}_n([1243], [1342]) = 4.$$

Proof. Take $[\sigma] \in \operatorname{Av}_n([1243], [1342])$ and write σ as in (3.4). Then ρ and τ can not both be nonempty. For if $x \in \rho$ and $y \in \tau$ then $1xny$ is a copy of either 1243 or 1342.

Assume first that $\rho = \emptyset$ so that

$$\sigma = 1n\tau. \tag{3.6}$$

Then τ must be increasing or decreasing. For suppose it was neither. Then it would contain a copy of one of the patterns 132, 231, 213, or 312. In the first two cases this would give, together with the 1, a copy of 1243 or 1342 in σ. And in the last two cases, prepending n gives a copy of 4213 or 4312. Conversely, if σ is given by (3.6) with τ increasing or decreasing then it is easy to verify that $[\sigma] \in \mathrm{Av}_n([1243], [1342])$.

Using the same ideas, one can also show that if $\tau = \emptyset$ then one gets exactly two elements of $\mathrm{Av}_n([1243], [1342])$, of the form $\sigma = 1\rho n$ where ρ is either increasing or decreasing. Thus there are a total of four elements in the avoidance class. □

Theorem 3.2.8. *For $n \geq 3$ we have*

$$\#\,\mathrm{Av}_n([1324], [1423]) = 2^{n-2}.$$

Proof. Take $[\sigma] \in \mathrm{Av}_n([1324], [1423])$ and write

$$\sigma = n, \rho, n-1, \tau.$$

Similar to the previous proof, one of ρ or τ must be empty since otherwise 4132 or 4231 is a pattern in σ. If $\rho = \emptyset$ then one shows similarly that $n-2$ either begins or ends τ. Continuing in this manner, we see that there are 2 choices for the positions of $n-1, n-2, \ldots, 2$. Checking, as usual, that all such permutations are actually in the avoidance set, the enumeration follows. □

This fully characterizes all non-trivial Wilf equivalences for all length four doubletons.

3.3 Three or more patterns

We will now compute $\#\,\mathrm{Av}_n[\Pi]$ for $\Pi \subseteq \mathfrak{S}_n$ having $\#\Pi \geq 3$. We will not consider those $[\Pi]$ containing both $[1234]$ and $[1432]$ since for such classes $\#\,\mathrm{Av}_n[\Pi] = 0$ for $n \geq 6$ as in Theorem 3.2.5.

Theorem 3.3.1. *We have*

$$\{[1234], [1243], [1324]\} \equiv \{[1234], [1324], [1342]\} \equiv \{[1243], [1423], [1432]\}$$
$$\equiv \{[1342], [1423], [1432]\}.$$

And for $n \geq 4$

$$\#\operatorname{Av}_n([1234], [1324], [1342]) = 3.$$

Proof. If $[\sigma] \in \operatorname{Av}_n([1234], [1324], [1342])$ then $[\sigma]$ avoids $[1324]$ and $[1342]$. So, by the proof of Theorem 3.2.6, we can write σ in the form (3.5) for $j, k, l \geq 1$. But since $[\sigma]$ also avoids $[1234]$ we must have $j + l \leq 3$. For the same reason, $j \leq 2$ since if $j = 3$ then the copy of ι_3 and one element of the copy of δ_k would form a $[1234]$. Thus the only possibilities are $(j, l) = (1, 1)$, $(1, 2)$, or $(2, 1)$ which proves the result. □

Theorem 3.3.2. *We have*

$$\{[1234], [1243], [1342]\} \equiv \{[1243], [1342], [1432]\}.$$

And for $n \geq 5$

$$\#\operatorname{Av}_n([1234], [1243], [1342]) = 2.$$

Proof. If $[\sigma] \in \operatorname{Av}_n([1234], [1243], [1342])$ then $[\sigma]$ avoids $[1243]$ and $[1342]$. So, by the proof of Theorem 3.2.7, we can write

$$\sigma = xy\rho \tag{3.7}$$

where $\{x, y\} = \{1, n\}$ and ρ is either increasing or decreasing. Since $n \geq 5$ we have $|\rho| \geq 3$. But $[\sigma]$ also avoides $[1234]$ and this forces ρ to be decreasing. So there are two choices for $[\sigma]$ depending on the values of x and y. □

Theorem 3.3.3. *We have*

$$\{[1234], [1243], [1423]\} \equiv \{[1234], [1342], [1423]\} \equiv \{[1243], [1324], [1432]\}$$
$$\equiv \{[1324], [1342], [1432]\}.$$

And for $n \geq 2$

$$\#\,\mathrm{Av}_n([1234], [1342], [1423]) = n - 1.$$

Proof. We will show that $T = T([1234], [1342], [1423])$ has production rules

$$(*) \rightarrow (1)(2),$$

$$(1) \rightarrow (1),$$

$$(2) \rightarrow (1)(2).$$

Then, by induction, level $n \geq 2$ of T will contain one node of degree 2 and $n-2$ nodes of degree 1. Checking the root is easy, so assume $n \geq 3$.

By Theorem 3.2.2, T is a subtree of $T([1234], [1342])$. So we just need to check which nodes of that tree also avoid $[1423]$. As in the proof of that theorem, the site before n in $[\sigma]$ at level n in T is still active since 4 is not followed immediately by 3 in $[1423]$. Thus it suffices to show that both sites of δ remain active, but only one in ϵ where δ, ϵ are defined by (3.1). Indeed, the two sites of δ give rise to copies of δ and ϵ at level $n+1$ of T. But site $n-1$ of delta which was active in the larger tree is now inactive since inserting $n+1$ there gives the copy $[1, n+1, 2, n]$ of $[1423]$. This completes the proof. □

We now have, in comparison with the previous theorem, a nontrivial Wilf equivalence.

Theorem 3.3.4. *We have*

$$\{[1234], [1324], [1423]\} \equiv \{[1324], [1423], [1432]\}.$$

And for $n \geq 2$

$$\#\,\mathrm{Av}_n([1234], [1324], [1423]) = n - 1.$$

Proof. It suffices to show that $T = T([1234], [1324], [1423])$ satisfies the same production rules as in the previous theorem. Now T is a subtree of $T([1234], [1324])$ which was constructed in the proof of Theorem 3.2.3. And we see in the usual way that the site before n in any $[\sigma]$ remains active in T because 4 is not immediately followed by 3 in $[1423]$.

So it suffices to show, with δ and ϵ as in (3.3), that site 1 remains active in δ, but not in ϵ. Indeed, inserting $n+1$ in this site of δ just produces another descending sequence. But in ϵ such a placement gives the copy $[1, n+1, n-1, n]$ of $[1423]$. □

We now have another nontrivial Wilf equivalence with Theorem 3.3.1.

Theorem 3.3.5. *We have*

$$\{[1243], [1324], [1342]\} \equiv \{[1243], [1342], [1423]\}.$$

And for $n \geq 4$

$$\#\operatorname{Av}_n([1243], [1324], [1342]) = 3.$$

Proof. By Theorem 3.2.7, we just need to show that exactly 3 of the 4 permutations $[\sigma]$ avoiding $\{[1243], [1342]\}$ also avoid $[1324]$. These permutations are described in equation (3.7). If $x = n$ and $y = 1$ then $[\sigma]$ contains the copy $[n132]$ of this pattern. It is also easy to check that the other three avoid it. □

We now have our last nontrivial Wilf equivalence for triples.

Theorem 3.3.6. *We have*

$$\{[1243], [1324], [1423]\} \equiv \{[1324], [1342], [1423]\}.$$

And for $n \geq 2$

$$\#\operatorname{Av}_n([1324], [1342], [1423]) = n - 1.$$

Proof. Comparing the description of $\operatorname{Av}_n([1324], [1342])$ in the proof of Theorem 3.2.6 and that of $\operatorname{Av}_n([1324], [1423])$ in the proof of Theorem 3.2.8, we see that any
$[\sigma] \in \operatorname{Av}_n([1324], [1342], [1423])$ can be put in the form

$$\sigma = 21[\delta_k, \iota_{n-k}]$$

with $k \geq 1$. Also, $k = n-1$ and $k = n$ yield the same permutation. So there are $n-1$ choices for k and we are done. □

$[\Pi]$	$\#\,\mathrm{Av}_n[\Pi]$
$\{[1234], [1243], [1324], [1342]\}$ $\{[1243], [1342], [1423], [1432]\}$	1
$\{[1234], [1243], [1324], [1423]\}$ $\{[1234], [1243], [1342], [1423]\}$ $\{[1234], [1324], [1342], [1423]\}$ $\{[1243], [1324], [1342], [1423]\}$ $\{[1243], [1324], [1342], [1432]\}$ $\{[1243], [1324], [1423], [1432]\}$ $\{[1324], [1342], [1423], [1432]\}$	2
$\{[1234], [1243], [1324], [1342], [1423]\}$ $\{[1243], [1324], [1342], [1423], [1432]\}$	1

Table 3.1: Wilf equivalence classes and cardinalities of $\mathrm{Av}_n[\Pi]$ for certain $[\Pi]$ and $n \geq 5$

When $\#[\Pi] \geq 4$ where $[\Pi] \subset [\mathfrak{S}_4]$, the size of $\mathrm{Av}_n[\Pi]$ becomes constant for $n \geq 5$. And this size is trivial to calculate for $n \leq 4$. Furthermore, the description of the surviving permutations for large n is easy to obtain given our previous proofs. So we content ourselves with a listing of the equivalence classes and associated constants in Table 3.1. Classes are separated by double horizontal line. As usual, we do not consider classes containing both the increasing and decreasing permutations because of the cyclic Erdős-Szekeres Theorem.

3.4 Cyclic descent generating functions

We will now consider the generating function for the number of cyclic descents over various avoidance classes $[\Pi] \subset [\mathfrak{S}_4]$, starting with those defined by a single element. We will sometimes use the characterizations given by Callan [Cal02] for these classes to facilitate our work, and use the abbreviation

$$D_n([\Pi]) = D_n([\Pi]; q) = \sum_{\sigma \in \mathrm{Av}_n[\Pi]} q^{\mathrm{cdes}\,\sigma}$$

for the generating function.

To begin, we have a lemma showing that trivial Wilf equivalences also give simple relationships between the corresponding generating functions.

Lemma 3.4.1. *For any* $[\Pi]$*, we have*

$$D_n([\Pi]^r; q) = D_n([\Pi]^c; q) = q^n D_n([\pi]; 1/q)$$

and

$$D_n([\Pi]^{rc}; q) = D_n([\Pi]; q).$$

Proof. Reversing or complementing a permutation turns all cyclic descents into cyclic ascents and vice-versa. Translating this into generating functions gives the first displayed equalities. And the second displayed equation follows from the the previous display. □

Now consider the possible $D_n([\pi])$ for $[\pi] \in [\mathfrak{S}_4]$. We begin with the simplest case.

Theorem 3.4.2. *We have* $D_n([1423]; q) = q^n D_n([1324]; 1/q)$ *where, for* $n \geq 2$*,*

$$D_n([1324]; q) = \sum_{k=1}^{n-1} \binom{n+k-3}{n-k-1} q^k.$$

Proof. We use Callan's characterization of this avoidance class to obtain a recursion for $D_n([1324])$. If $[\sigma] \in \mathrm{Av}_n([1324])$ and $n \geq 3$ then write $\sigma = \sigma_1\sigma_2 \ldots \sigma_{n-1}n$. Let k be the index such that $\sigma_k = n-1$. There are two cases.

If $k = n-1$ then $\sigma = \tau, n-1, n$ where $[\tau, n-1] \in \mathrm{Av}_{n-1}([1324])$ and this is a bijection. Since $\mathrm{cdes}[\sigma] = \mathrm{cdes}[\tau, n-1]$, this case contributes $D_{n-1}([1324])$ to the recursion.

If $1 \leq k \leq n-2$ then this forces

$$\sigma = 2314[\iota_{k-1}, 1, \tau, 1]$$

for some τ such that $[\tau n]$ avoids $[1324]$. Because of the extra descent caused by $n-1$ we have $\mathrm{cdes}[\sigma] = 1 + \mathrm{cdes}[\tau n]$. So this case gives a contribution of $\sum_{k=1}^{n-2} qD_{n-k}([1324])$.

Putting everything together, we have

$$D_n([1324]) = D_{n-1}([1324]) + \sum_{k=1}^{n-2} qD_{n-k}([1324]).$$

for $n \geq 3$ and $D_2([1324]) = q$. It is now a simple manner of manipulating binomial coefficients to show that the formula given in the theorem satisfies this initial value problem. □

For the next case, we will use a characterization of the class different from the one found by Callan. This will permit us to avoid the use of a recurrence.

Lemma 3.4.3. *Suppose* $[\sigma] \in [\mathfrak{S}_n]$ *and write* $\sigma = 1\rho n\tau$. *We have* $[\sigma] \in \mathrm{Av}_n([1342])$ *if and only if the following three conditions are satisfied:*

(a) ρ *and* τ *both avoid* $\{213, 231\}$,

(b) $\max \rho < \min \tau$,

(c) there is not both a descent in ρ *and an ascent in* τ.

Proof. For the forward direction, suppose $[\sigma] \in \mathrm{Av}_n([1342])$. Condition (a) is true since if either ρ or τ contains 213 then, together with n, we have that $[\sigma]$ contains $[2134]$. Similarly, if either contains 231 then $[\sigma]$ contains the forbidden pattern by prepending the 1. As far as (b), if there is $y > x$ with $y \in \rho$ and $x \in \tau$ then $[1ynx]$ is a copy of $[1342]$. Finally for (c), if there were a descent in ρ and an ascent in τ then, because of (b), putting them together would again give a copy of the pattern to avoid.

The converse is similar where one assumes that a copy of $[1342]$ exists and then considers all the different intersections it could have with 1, ρ, n, and τ. We leave the details to the reader. □

In order to use this lemma, we will need a result about the ordinary descent statistic on linear permutations avoiding $\{123, 231\}$. The next result is a specialization of Proposition 5.2 of the paper of Dokos, Dwyer, Johnson, Sagan, and Selsor [DDJ+12] and so the proof is ommited.

Lemma 3.4.4 ([DDJ+12])**.** *We have*

$$\sum_{\sigma \in \mathrm{Av}_n(213,231)} q^{\operatorname{des}\sigma} = (1+q)^{n-1}. \qquad \square$$

We need one last well-known definition. Call a polynomial $f(q) = \sum_{k=0}^{n} a_k q^k$ of degree n *symmetric* if $a_k = a_{n-k}$ for all $0 \le k \le n$. Note that $f(q)$ of degree n is symmetric if and only if

$$q^n f(1/q) = f(q). \tag{3.8}$$

Theorem 3.4.5. *We have* $D_n([1243];q) = D_n([1342];q)$ *where, for* $n \geq 2$*,*

$$D_n([1342];q) = 2q(1+q)^{n-2} - q \cdot \frac{1-q^{n-1}}{1-q}$$

is symmetric.

Proof. It is easy to prove from the explicit form of $D_n([1342])$ that it satisfies equation (3.8) and so is symmetric. So once this is proved, the equality of the two generating functions follows from Lemma 3.4.1.

We adopt the notation of Lemma 3.4.3 and let $\sigma_k = n$ where $2 \leq k \leq n$. We will consider cases depending on whether ρ or τ is empty. If $\rho = \emptyset$ then by Lemma 3.4.3 (a) and Lemma 3.4.4 we have that the generating function for the possible linear τ is $(1+q)^{n-3}$. Also, $\operatorname{cdes}[\sigma] = 2 + \operatorname{des}\tau$ by the form of σ, so the contribution of such $[\sigma]$ to $D_n([1342])$ is $q^2(1+q)^{n-3}$. In an analogous way, we see that those $[\sigma]$ with $\tau = \emptyset$ yield $q(1+q)^{n-3}$. Adding these, we have a total of $q(1+q)^{n-2}$ so far.

We now assume that ρ, τ are both nonempty so that $3 \leq k \leq n-1$. By parts (b) and (c) of Lemma 3.4.3, either ρ must be an increasing subsequence of consecutive integers or τ must be a decreasing one. Using Lemma 3.4.4 again, we see that in the first subcase a contribution of $q^2(1+q)^{n-k-1}$ is obtained. And in the second, taking into account the descents in ρ, the contribution is $q^{n-k+1}(1+q)^{k-3}$. However, these two subcases overlap when ρ is increasing and τ is decreasing. So we must subtract q^{n-k+1}.

Thus we get a grand total of

$$D_n([1342]) = q(1+q)^{n-2} + \sum_{k=3}^{n-1} [q^2(1+q)^{n-k-1} + q^{n-k+1}(1+q)^{k-3} - q^{n-k+1}].$$

Summing the geometric series and simplifying completes the proof. □

For the avoidance class of the increasing (or decreasing) pattern in $[\mathfrak{S}_4]$, we will need another concept. Given sequences ρ and τ of distinct integers, their *shuffle set* is

$$\rho \sqcup\!\sqcup \tau = \{\sigma \ : \ |\sigma| = |\rho| + |\tau| \text{ and both } \rho, \tau \text{ are subsequences of } \sigma\}.$$

For example,

$$12 \sqcup\!\sqcup 34 = \{1234, 1324, 1342, 3124, 3142, 3412\}.$$

In the statement of the next result we make the usual convention that $\binom{n}{k} = 0$ if $k > n$.

Theorem 3.4.6. *We have* $D_n([1234]; q) = q^n D_n([1432]; 1/q)$ *where, for* $n \geq 2$,

$$D_n([1432]; q) = q + (2^{n-1} - n)q^2 + \sum_{j \geq 3} \binom{n}{2j-1} q^j.$$

Proof. We use Callan's description of the avoidance for [1234] translated by complementation to apply to [1432]. We are going to derive a recursion for $D_n([1432]; q)$. If $[\sigma] \in \mathfrak{S}_n[1432]$ then suppose $\sigma_n = 1$ and $\sigma_k = 2$ for some $1 \leq k \leq n-1$. There are three cases.

If $k = 1$ then there is a bijection between such $[\sigma]$ and $\mathrm{Av}_{n-1}[1432]$ obtained by removing 1 and taking the order isomorphic cyclic permutation on $[n-1]$. Since 2 immediately follows 1 cyclically in $[\sigma]$, the descent into 1 remains a descent after applying the map. So the contribution of this case is $D_{n-1}([1432]; q)$.

Now suppose that $2 \leq k \leq n-1$ and write

$$\sigma = \rho 2 \tau 1.$$

where $|\rho| = k-1$, $|\tau| = n-k-1$. As Callan proves, ρ must be increasing. So there are two more cases depending upon whether the elements of ρ are consecutive or not. Suppose first that they are not consecutive. In this case, τ must also be increasing so $\mathrm{cdes}[\sigma] = 2$. To compute the number of such σ, note that once the elements of ρ have been picked from $[3, n]$, all of σ is determined. The total number of nonempty subsets of this interval is $2^{n-2} - 1$. And those which consist of consecutive integers are determined by their minimum and maximum element, which could be equal. So there are $\binom{n-1}{2}$ subsets to exclude. The contribuion of this case is then

$$\left(2^{n-2} - \binom{n-1}{2} - 1\right) q^2.$$

Finally we consider the case when $\rho \neq \emptyset$ is consecutive (and still increasing), say with minimum $m+1$ and maximum $M-1$. Note that if $l = |\tau|$ then $0 \leq l \leq n-3$. Callan shows that the possible τ

are the elements of $(34\ldots m) \sqcup\!\sqcup (M, M+1, \ldots, n)$. Since a permutation can be written as a shuffle in many ways, the same shuffle could occur for different ρ. So it will be convenient to color the elements of the second sequence by marking them with a hat. Thus the σ in this case are in bijection with colored shuffles $(34\ldots m) \sqcup\!\sqcup (\widehat{M}, \widehat{M+1}, \ldots, \widehat{n})$. It will also be convenient to consider these as corresponding to the sequences 2τ by prepending a 2 to each shuffle and considering 2 as an uncolored element. Set S be the set of such sequences $s = 2s_2s_3\ldots s_{l+1}$ where l, m, M are allowed to vary over all possible values. Note that if s corresponds to σ then $\operatorname{des}\sigma = 2 + \operatorname{des} s$. To compute $\operatorname{des} s$, we consider the transition indices

$$\operatorname{Tr} s = \{i \mid s_i \text{ is colored and } s_{i+1} \text{ is not, or vice-versa}\}.$$

For example, if $s = 23\widehat{6}45\widehat{7}\widehat{8}$ then $\operatorname{Tr} s = \{2, 3, 5\}$. It is easy to see that the map $\operatorname{Tr} : S \to 2^{[l]}$, the range being all subsets of $[l]$, is a bijection. Also, every other transition index of s starting with the second corresponds to a descent. So, using the round down function, $\operatorname{des} s = \lfloor \#\operatorname{Tr} s/2 \rfloor$. We can now complete this case using $i = \#\operatorname{Tr} s$ to see that the contribution is

$$\begin{aligned}
\sum_{l=0}^{n-3}\sum_{i=0}^{l}\binom{l}{i}q^{\lfloor i/2\rfloor+2} &= \sum_{i=0}^{n-3} q^{\lfloor i/2\rfloor+2}\sum_{l=i}^{n-3}\binom{l}{i} \\
&= \sum_{i=0}^{n-3}\binom{n-2}{i+1}q^{\lfloor i/2\rfloor+2} \\
&= q^2\sum_{j\geq 0}\left[\binom{n-2}{2j+1}+\binom{n-2}{2j+2}\right]q^j \\
&= q^2\sum_{j\geq 0}\binom{n-1}{2j+2}q^j.
\end{aligned}$$

Putting all the cases together we have

$$D_n([1432];q) = D_{n-1}([1432];q) + q^2\left[2^{n-2} - \binom{n-1}{2} - 1 + \sum_{j\geq 0}\binom{n-1}{2j+2}q^j\right].$$

As usual, the routine verification that our desired formula satisfies this recursion and the initial condition is left to the reader. □

We now turn to the cyclic descent polynomials for pairs in $[\mathfrak{S}_4]$. To simplify notation, for any polynomial $f(q)$ and $n \in \mathbb{N}$ we let

$$f^{(n)}(q) = q^n f(1/q).$$

Theorem 3.4.7. *We have the following descent polynomials.*

(a) We have

$$D_n([1234],[1243]) = D_n([1342],[1432]) = D_n^{(n)}([1243],[1432]) = D_n^{(n)}([1234],[1342]).$$

And for $n \geq 3$

$$D_n([1234],[1342];q) = (2n-5)q^{n-2} + q^{n-1}.$$

(b) We have

$$D_n([1423],[1432]) = D_n^{(n)}([1234],[1324]).$$

And for $n \geq 3$

$$D_n([1234],[1324];q) = (2n-5)q^{n-2} + q^{n-1}.$$

(c) We have

$$D_n([1324],[1432]) = D_n^{(n)}([1234],[1423]).$$

And for $n \geq 1$

$$D_n([1234],[1423];q) = q^{n-1} + \binom{n-1}{2} q^{n-2}.$$

(d) We have

$$D_n([1243],[1423]) = D_n([1342],[1423]) = D_n^{(n)}([1243],[1324]) = D_n^{(n)}([1324],[1342]).$$

And for $n \geq 1$

$$D_n([1324],[1342];q) = q + \sum_{k=2}^{n-1} (n-k)q^k.$$

(e) For $n \geq 4$ we have

$$D_n([1243], [1342]; q) = q + q^2 + q^{n-1} + q^{n-2}.$$

(f) For $n \geq 3$ we have

$$D_n([1324], [1423]; q) = q(1+q)^{n-2}.$$

Proof. We will only prove (a) as the others follow easily in a similar fashion from the descriptions of the avoidance classes in Section 3.2. We adopt the notation of the proof of Theorem 3.2.2.

We will use the description of the generating tree to obtain a recursion for $D_{n+1}[1243], [1432])$. Note that if $n+1$ is inserted in site i of σ to form σ' then

$$\mathrm{cdes}[\sigma'] = \begin{cases} \mathrm{cdes}[\sigma] & \text{if } i \text{ is a cyclic descent,} \\ \mathrm{cdes}[\sigma] + 1 & \text{if } i \text{ is a cyclic ascent.} \end{cases}$$

Since the site before n is always active, and such a site is a cyclic ascent, these children will give a contribution of $qD_n([1243], [1432])$. In δ and ϵ, insertion in the other active site gives permutations with $n-1$ descents. So

$$D_{n+1}[1243], [1432]) = 2q^{n-1} + qD_n([1243], [1432]).$$

It is now easy to check that the formula in (a) satisfies this recursion and is also valid at $n = 3$, completing the proof. □

For classes avoiding 3 or more patterns, we will only write down the results for those which are not eventually constant. The interested reader can easily compute the polynomials for the remaining classes. We also content ourselves with stating the polynomial for one member of every trivial Wilf equivalence class since the rest can be computed from Lemma 3.4.1.

Theorem 3.4.8. *We have the descent polynomials*

$$D_n([1234], [1342], [1423]; q) = D_n([1234], [1324], [1423]; q) = (n-2)q^{n-2} + q^{n-1}$$

and

$$D_n([1324], [1342], [1423]; q) = q \cdot \frac{1-q^{n-1}}{1-q}$$

for $n \geq 2$. □

3.5 Open problems and concluding remarks

We collect here various areas for future research in the hopes that the reader will be interested in pursuing this work.

3.5.1 Longer patterns

There has been very little work about containment and avoidance for cyclic patterns of length longer than 4. Of course, the cyclic Erdős-Szekeres Theorem, Theorem 3.1.1 above, is one such result. There is also a paper of Gray, Lanning and Wang [GLW18] where the authors consider cyclic packing (maximizing the number of copies of a given pattern among all the permutations $[\sigma] \in [\mathfrak{S}_n]$ for some n) and superpatterns (permutations containing all the patterns $[\pi] \in [\mathfrak{S}_k]$ for some k). It would be interesting to see if there are nice enumerative formulas for classes consisting of cyclic patterns of length 5 and up.

3.5.2 Other statistics

We have previously mentioned the *peak set* of a linear permutation,

$$\operatorname{Pk} \pi = \{i \mid \pi_{i-1} < \pi_i > \pi_{i+1}\}$$

which has corresponding *peak number*

$$\operatorname{pk} \pi = \# \operatorname{Pk} \pi.$$

Peaks are an important part of Stembridge's theory of enriched P-partitions [Ste97] where P is a partially ordered set. On the enumerative side, the study of permutations which have a given peak set has been a subject of current interest [BBPS15, BBS13, BFT16, CVDLO$^+$17, DLHIO17a, DLHIO17b, DLHIPL17]. Now define the *cyclic peak number* to be

$$\operatorname{cpk}[\pi] = \#\{i \mid \pi_{i-1} < \pi_i > \pi_{i+1} \text{ where subscripts are taken modulo } n\}.$$

As with cdes, this is well defined since it is independent of the choice of representative of $[\pi]$. There should be interesting generating functions for the distribution of cpk over avoidance classes, or even for the joint distribution of cdes and cpk. As evidence, we prove one such result.

Theorem 3.5.1. *For $n \geq 3$*

$$\sum_{[\sigma] \in \mathrm{Av}_n([1234],[1342])} q^{\mathrm{cdes}[\sigma]} t^{\mathrm{cpk}[\sigma]} = q^{n-2}t + (2n-6)q^{n-2}t^2 + q^{n-1}t.$$

Proof. Let $F_n(q,t)$ denote the desired generating function. We proceed as in the proof of Theorem 3.4.7 (a) to find a recursion for $F_{n+1}(q,t)$. Since the largest element of $[\sigma]$ is always a cyclic peak, inserting $n+1$ before n does not change cpk. So this contributes $qF_n(q,t)$ to the recursion. For δ and ϵ, inserting $n+1$ in the other active site increases the number of peaks to 2. So the contribution from these cases is $2q^{n-1}t^2$. In summary

$$F_{n+1}(q,t) = 2q^{n-1}t^2 + qF_n(q,t)$$

and the desired polynomial is easily seen to be the solution. □

In a recent paper Adin, Gessel, Reiner, and Roichman [AGRR20] defined a cyclic analogue of the Hopf algebra of quasisymmetric functions. In this context the cyclic descent set of a linear permutation arises naturally in the description of the product in this algebra. They also raise the following intriguing question.

Question 3.5.2. *Find an analogue of the major index for cyclic permutations that has nice properties, such as a generating function over $[\mathfrak{S}_n]$ which factors nicely as does the generating function for the ordinary major index over $\mathfrak{S}_n$.*

3.5.3 Vincular patterns

The study of vincular patterns was originated by Babson and Steingrímsson [BS00] and has since become a mainstay of the pattern field. We consider π as a *vincular pattern* if one only counts occurrences in σ where certain adjacent elements of π must also be adjacent in the copy in σ. Such

adjacent elements are overlined in π. For example, $\sigma = 24513$ contains two copies of $\pi = 132$, namely 243 and 253. But only 243 is a copy of $\overline{13}2$. Avoidance and Wilf equivalence are defined in the obvious way. These notions and the corresponding notation carry over to cyclic patterns without change. There are undoubtedly nice results which can be proven about vincular cyclic patterns. As an example, we show how one vincular class is enumerated by the Catalan numbers.

Theorem 3.5.3. *We have*

$$[13\overline{24}] \equiv [1\overline{42}3] \equiv [\overline{13}24] \equiv [2\overline{31}4].$$

And for $n \geq 1$

$$\#\operatorname{Av}_n[13\overline{24}] = C_{n-1}.$$

Proof. The Wilf equivalences are trivial. To prove the Catalan formula, suppose that $[\sigma] \in \operatorname{Av}_n[13\overline{24}]$ for $n \geq 2$ and write σ so that $\sigma_n = n$ and $\sigma_{n-1} = m$ for some $m \in [n-1]$. First notice that $\sigma = \rho\tau mn$ where ρ and τ are permutations of $[m+1, n-1]$ and $[m-1]$, respectively. For if there are $x < m < y < n$ with x before y in σ then $[xymn]$ is a copy of $[13\overline{24}]$. Furthermore, it is clear that $[m\rho]$ and $[\tau m]$ must avoid the forbidden pattern.

We claim the if $\sigma = \rho\tau mn$ where ρ and τ obey the restrictions of the previous paragraph then $[\sigma]$ avoids $[13\overline{24}]$. Suppose, towards a contradiciton, that a copy $[\kappa] = [wyxz]$ exists with $wyxz$ order isomorphic to $13\overline{24}$. Consider the elements x and z which play the roles of 2 and 4. The possibility that they are m and n, respectively, is ruled out by the fact that every element of ρ is larger than every element of τ. If $z \in \tau m$ then all of κ must be in this subsequence since z is the largest element of the copy. But this is impossible since $[\tau m]$ avoids the bad pattern. Finally, suppose $z \in \rho$. This forces $x \in \rho$ since it is comes cyclically just before z, and n is too large to be x. We must also have $y \in \rho$ since $x < y < z$. But now there is no possible choice for w. Indeed, if $w \in [m\rho]$ then $[\kappa]$ is in this subsequence, contradicting our assumption. And if $w \in \tau$ then it could be replaced by m since $x, y, z > m$, yielding the same contradiction as before.

From the first two paragraphs we immediately get the recursion

$$\#\operatorname{Av}_n[13\overline{24}] = \sum_{m=1}^{n-1} \#\operatorname{Av}_m[13\overline{24}] \cdot \#\operatorname{Av}_{n-m}[13\overline{24}].$$

From this the Catalan enumeration follows by induction. □

It appears that sometimes rather than trying to find the size of the avoidance class directly, it may be easier to use exponential generating functions. Given a set of (possibly vincular) patterns $[\Pi]$, let

$$E[\Pi] = \sum_{n\geq 0} \#\operatorname{Av}_n[\Pi]\frac{x^n}{n!}.$$

We have the following conjectures for two vincular avoidance classes. Once the corresponding differential equation is proved, an explicit solution can easily found using separation of variables.

Conjecture 3.5.4. *We have the following.*

1. *If* $E = E[\overline{123}]$ *then*

$$E' = E^2 - E + 1.$$

2. *If* $E = E[\overline{213}]$ *then*

$$E' = e^{E-\frac{x^2}{2}}.$$

Recently, Sergi Elizalde and Bruce Sagan have constructed proofs of both conjectured results through a more general result using generating functions which keep track of the number of cyclic occurrences, instead of just avoidance [ES21].

CHAPTER 4

PINNACLE SET PROPERTIES

This chapter contains material from Domagalski, Liang, Minnich, Sagan, Schmidt, and Sietsema [DLM+21c]. All results in this chapter are from this manuscript except as otherwise noted.

4.1 Counting admissible pinnacle sets

In this section we give our proof of Theorem 2.0.5, which gives the number of admissible pinnacle sets $S = \operatorname{Pin}\pi$ for some $\pi \in \mathfrak{S}_n$. Our strategy will be as follows. First, we will introduce the set of interleaved permutations which are obviously counted by the desired binomial coefficient. Next, we will associate with each admissible pinnacle set S a particular permutation π such that $\operatorname{Pin}\pi = S$. This permutation will be called right canonical because its pinnacles will be as far right as possible. Finally, we will show that the set of interleaved permutations and the set of right canonical permutations are, in fact, the same. This will complete the proof of the theorem.

An *interleaved permutation* $\pi \in \mathfrak{S}_n$ is one constructed in the following manner. Pick any $A \subseteq [2, n]$ with $\#A = \left\lfloor \frac{n-1}{2} \right\rfloor$.

I1 Fill the first $\left\lfloor \frac{n-1}{2} \right\rfloor$ even positions of π with the elements of A in increasing order.

I2 Fill the remaining positions of π with the elements of $\overline{A} = [n] - A$ in increasing order.

As an example, suppose $n = 9$ and $A = \{2, 3, 7, 9\}$. After step I1 we have

$$\pi = \_\ 2\ \_\ 3\ \_\ 7\ \_\ 9\ \_.$$

Since $\overline{A} = \{1, 4, 5, 6, 8\}$, after I2 we have the full interleaved permutation

$$\pi = 1\ 2\ 4\ 3\ 5\ 7\ 6\ 9\ 8. \tag{4.1}$$

Let

$$\mathcal{I}_n = \{\pi \in \mathfrak{S}_n \mid \pi \text{ is interleaved}\}.$$

Clearly $\pi \in \mathcal{I}_n$ is completely determined by the choice of A. It follows immediately that

$$\#\mathcal{I}_n = \binom{n-1}{\lfloor \frac{n-1}{2} \rfloor}. \tag{4.2}$$

Now given an admissible pinnacle set $S = \{s_1 < s_2 < \ldots < s_d\} \subset [n]$ we wish to construct a permutation $\pi \in \mathfrak{S}_n$ with $\operatorname{Pin} \pi = S$. We use the following algorithm to construct the *right canonical permutation* π from S. We first deal with the case where n is odd. Let $\overline{S} = [n] - S$.

C1 Place elements of $\overline{S}$ in π moving right to left, starting with the largest unused element of $\overline{S}$ and then decreasing until an element less than the largest unused element of S is placed.

C2 Place the largest unused element of S in the rightmost unused position.

C3 Iterate C1 and C2 until all elements of S and $\overline{S}$ are placed.

If n is even, the only change to this procedure is that we fill both π_n and π_{n-1} with elements of $\overline{S}$ before considering whether to place an element of S. To illustrate, consider $n = 9$ and $S = \{4, 7, 9\}$. So $\overline{S} = \{1, 2, 3, 5, 6, 8\}$. Here is the construction of π where, at each stage, we note whether C1 or C2 is being used.

step	C1	C2	C1	C2	C1	C1	C2	C1	C1
π	8	98	698	7698	57698	357698	4357698	24357698	124357698

So the right canonical permutation for $S = \{4, 7, 9\}$ is $\pi = 124357698$. Note that $\operatorname{Pin} \pi = S$. Furthermore, this is the same permutation as obtained in (4.1). However, neither the sets A nor $\overline{A}$ equals S. Let

$$C_n = \{\pi \in \mathfrak{S}_n \mid \pi \text{ is right canonical}\}.$$

We first need to show that C1–C3 is well defined in that every position of π gets filled and that we always have $\operatorname{Pin} \pi = S$. Recall that $\mathcal{A}_n = \{S \mid S = \operatorname{Pin} \pi \text{ for some } \pi \in \mathfrak{S}_n\}$.

Lemma 4.1.1. *If $S \subset [n]$ is an admissible set then C1–C3 produces a permutation π with* $\operatorname{Pin} \pi = S$. *Thus*

$$\#C_n = \#\mathcal{A}_n.$$

Proof. Clearly the second sentence follows from the first. For the first sentence, we will present details for the case when n is odd. If n is even, then one can just place the largest element of $\overline{S}$ in position n and proceed as in the odd case.

The following notation will be useful. Let

$$S = \{s_1 < s_2 < \ldots < s_d\},$$
$$\overline{S} = \{\bar{s}_1 < \bar{s}_2 < \ldots < \bar{s}_{n-d}\}.$$

We will also let S_p and $\overline{S}_p$ denote the elements of S and of $\overline{S}$, respectively, which have not been used during the placement of $\pi_n, \pi_{n-1}, \ldots, \pi_p$.

We will use reverse induction on the position p being filled in π. When $p = n$, we have $\overline{S} \neq \emptyset$ since 1, which is always a non-pinnacle, must be in $\overline{S}$. So there is an element $\bar{s}_{n-d}$ to place in position n. Furthermore this element can not be a pinnacle since it is the last element of the permutation, which agrees with the fact that it is in $\overline{S}$.

Suppose that $\pi_n, \pi_{n-1}, \ldots, \pi_{p+1}$ have been constructed. Suppose first that $\pi_{p+1} \in \overline{S}$. One subcase is if either $S_p = \emptyset$, or $S_p \neq \emptyset$ and $\pi_{p+1} > \max S_p$. We must show that $\overline{S}_p \neq \emptyset$ so that we can let $\pi_p = \max \overline{S}_p$. This is true when $S_p = \emptyset$ since $|S_p \uplus \overline{S}_p| = p$. If the second option holds then we have $\pi_{p+1} > \max S_p$. But there must be at least two elements of $\overline{S}$ smaller than $\max S_p$ since S is admissible and so there is some permutation making $\max S_p$ a pinnacle. Also, these elements must still be in $\overline{S}_p$ since elements of this set are placed in decreasing order right to left. Thus this set is nonempty as desired. Furthermore, π_p is not a pinnacle since it is smaller than π_{p+1}.

Now consider the subcase when $\pi_{p+1} < \max S_p$. Then we let $\pi_p = \max S_p$ which is well defined. But we must show that π_p is a pinnacle. We know $\pi_p > \pi_{p+1}$. So there remains to check whether one can construct π_{p-1} with $\pi_{p-1} < \pi_p$. For this, it suffices to show that $\overline{S}_{p-1} \neq \emptyset$ since then we will have $\pi_{p-1} = \max \overline{S}_{p-1} < \pi_{p+1} < \pi_p$. Note that this will also finish the induction step.

We claim that if $\pi_p = s_i$ and $\pi_{p+1} = \bar{s}_j$ then $j > i$. It will then follow that $\bar{s}_{j-1}$ exists and can be used for π_{p-1}. But by Theorem 2.0.4 we have $s_i > \bar{s}_{i+1}$. Indeed, if $s_i < \bar{s}_{i+1}$ then at most the elements $s_1, \ldots, s_{i-1}, \bar{s}_1, \ldots, \bar{s}_i$ are less than s_i so that $s_i \leq 2i$, a contradiction. Also, elements of

$\overline{S}$ are placed in decreasing order with s_i being placed as early as possible with a smaller element to its right. The desired bound on j follows. □

We are now ready to give our proof of Theorem 2.0.5.

Theorem 4.1.2. *We have $C_n = \mathcal{I}_n$. Thus*

$$\#\mathcal{A}_n = \binom{n-1}{\lfloor \frac{n-1}{2} \rfloor}.$$

Proof. The second statement follows directly from the first, Lemma 4.1.1, and equation (4.2). So we only need to prove that the two sets are the same. We will consider the case when n is odd, as the even case is similar.

We begin by showing that any right canonical permutation π is interleaved. That is to say, the subword consisting of all even indices is an increasing sequence, and the subword consisting of all odd indices is an increasing sequence starting with 1.

In terms of the placement of 1, note that π_1 is not a pinnacle. And since non-pinnacles are placed in decreasing order from right to left, we must have $\pi_1 = 1$.

To finish this direction, it is enough to show that for any elements π_i and π_{i+2}, we have that $\pi_{i+2} > \pi_i$. Note that we are done immediately if π_i and π_{i+2} are either both pinnacles or both non-pinnacles since the construction places them in decreasing order from right to left. If π_{i+2} is a pinnacle and π_i is not, then by the pinnacle assumption $\pi_{i+2} > \pi_{i+1}$. And since non-pinnacles are placed in decreasing order right to left, $\pi_{i+1} > \pi_i$. Combining the two inequalities gives the desired result. Finally, suppose π_i is a pinnacle and π_{i+2} is not. Then π_{i+1} is not a pinnacle, being adjacent to π_i. And, by construction, π_{i+1} must be the first available non-pinnacle right to left which is smaller than π_i. It follows that $\pi_{i+2} > \pi_i$.

For set containment the other way, let π be an interleaved permutation. It suffices to show that if the elements of π are placed right to left then they follow C1–C3. Consider π_i placed after π_{i+1} with $1 < i < n$. The boundary cases when $i = 1$ or n are similar. If $\pi_i < \pi_{i+1}$ then π_i is a non-pinnacle and π_{i+1} is either a non-pinnacle or a pinnacle. In the first case, the non-pinnacles are being placed in decreasing order as desired. In the second, the previously placed non-pinnacle is π_{i+2}. So the

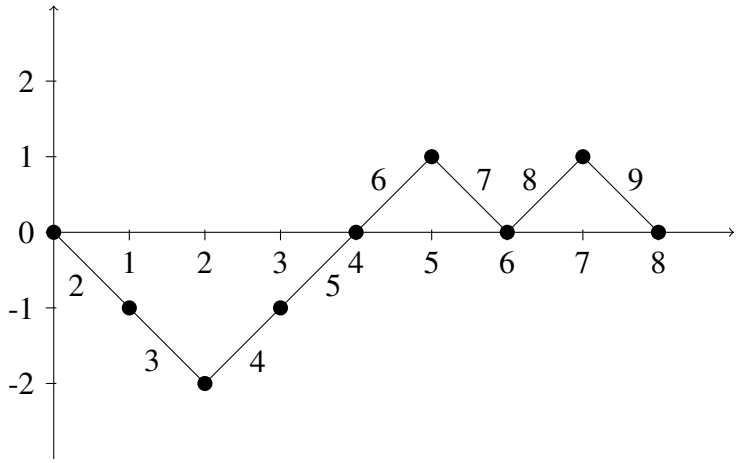

Figure 4.1: The lattice path L for $A = \{2, 3, 7, 9\}$

same conclusion holds by the interleaving condition. Now consider the possibility $\pi_i > \pi_{i+1}$. By the interleaving condition, $\pi_{i-1} < \pi_{i+1}$ so π_i is a pinnacle. Either π_{i+2} is a pinnacle or not, the latter possibility including the case that π_{i+2} does not exist. If it is, then the interleaving condition shows that pinnacles are being placed in decreasing order. If π_{i+2} is not a pinnacle, then this fact and the interleaving condition again imply $\pi_i < \pi_{i+2} < \pi_{i+3}$. It follows that π_i was placed after the first smaller non-pinnacle and, by the interleaving condition one last time, that any pinnacles to its right are larger. This completes the proof of the other containment. □

Given a set A and $k \in \mathbb{N}$ we let $\binom{A}{k}$ be the set of all k-element subsets of A. The above construct gives us a bijection

$$\psi : \binom{[2,n]}{\lfloor \frac{n-1}{2} \rfloor} \to \mathcal{A}_n$$

given by

$$\psi(A) = \operatorname{Pin} \pi$$

where π is the interleaving permutation corresponding to A.

In [DNKPT18], the authors proved Theorem 2.0.5 using a bijection

$$\phi : \binom{[2,n]}{\lfloor \frac{n-1}{2} \rfloor} \to \mathcal{A}_n$$

defined as follows. An *up-down lattice path* L starts at the origin and uses steps which are either up (U) or down (D) parallel to the vectors $[1, 1]$ and $[1, -1]$, respectively. For more information about lattice paths, see the text of Sagan [Sag20]. It will be convenient to index the steps of L with $[2, n]$ and write $L = s_2 s_3 \dots s_n$. Associate with $A \in \binom{[2,n]}{\lfloor \frac{n-1}{2} \rfloor}$ the lattice path L such that

$$s_i = \begin{cases} D & \text{if } i \in A, \\ U & \text{if } i \notin A. \end{cases}$$

To illustrate, if $n = 9$ and $A = \{2, 3, 7, 9\}$ as in the example beginning this section then

$$L = DDUUUDUD$$

as depicted in Figure 4.1 where each step is labeled by its index. We now define

$$\phi(A) = \{i \mid \text{in } L \text{ either } s_i = U \text{ strictly below the } x\text{-axis, or } s_i = D \text{ weakly above the } x\text{-axis}\}.$$

Continuing our example, $\phi(\{2, 3, 7, 9\}) = \{4, 7, 9\} = \psi(\{2, 3, 7, 9\})$. This is not an accident.

Proposition 4.1.3. *We have*

$$\phi = \psi.$$

Proof. We will give the proof for n odd as the even case is similar. Let $l = (n - 1)/2$. We need to show that $\phi(A) = \psi(A)$ for all $A \in \binom{[2,n]}{l}$. Suppose $A = \{a_1 < a_2 < \dots < a_l\}$ and $\overline{A} = [n] - A = \{\overline{a}_1 < \overline{a}_2 < \dots < \overline{a}_{n-l}\}$. Let L and π be the lattice path and interleaved permutation, respectively, associated with A. So $\psi(A) = \operatorname{Pin} \pi$ and there will be two cases depending on whether a pinnacle of π comes from A or $\overline{A}$

In the first case, suppose $a_i \in \operatorname{Pin} \pi$. Since π is interleaved, this is equivalent to $a_i = \pi_{2i} > \pi_{2i+1} = \overline{a}_{i+1}$. Recall that a_i indexes the ith D step of L, and similarly for $\overline{a}_{i+1}$ and U steps. So the previous inequality is equivalent to step $s_{a_i} = D$ being preceded by more up steps than down steps. And this is precisely the condition for a_i to be the index of a down step weakly above the x-axis, which means it is in $\phi(A)$. Thus this case is complete.

In a similar manner, one proves that $\overline{a}_i \in \operatorname{Pin} \pi$ if and only if $\overline{a}_i$ is the index of an up step strictly below the x-axis. This completes the second case and the proof. □

4.2 Ballot numbers

Davis et al. derived a number of properties of the constants $\mathfrak{p}(m, d)$ which count the number of admissible pinnacle sets S with d elements and maximum m. In this section we prove that these constants are, in fact, ballot numbers. We give two proofs of this result. In the first, we derive a formula for $\mathfrak{p}(m, d)$ using finite differences and then show that it agrees with the well-known expression for ballot numbers. In the second, we give an explicit bijection between these admissible sets and ballot sequences.

Suppose we are given nonnegative integers $p > q$. A (p, q) *ballot sequence* is a permutation $\beta = \beta_1\beta_2 \dots \beta_{p+q}$ of p copies of the letter X and q copies of the letter Y such that in any nonempty prefix $\beta_1\beta_2 \dots \beta_i$ the number of X's is greater than the number of Y's. Let

$$\mathcal{B}_{p,q} = \{\beta \mid \beta \text{ is a } (p, q) \text{ ballot sequence}\}.$$

The following result is well known.

Theorem 4.2.1 ([And87],[Ber87])**.** *For nonnegative integers $p > q$ we have*

$$\#\mathcal{B}_{p,q} = \frac{p-q}{p+q}\binom{p+q}{q}. \qquad \square$$

Note that if we let $p = d + 1$ and $q = d$ then the previous result gives get

$$\#\mathcal{B}_{d+1,d} = \frac{1}{2d+1}\binom{2d+1}{d} = C_d$$

where C_d is the dth Catalan number.

Our first proof that the $\mathfrak{p}(m, d)$ are ballot numbers will use the theory of finite differences. If $f(m)$ is a function of a nonnegative integer m then its *forward difference* is the function Δf defined by

$$\Delta f(m) = f(m + 1) - f(m).$$

For a fixed $d \in \mathbb{P}$, define the following polynomial in m of degree $d - 1$

$$p_d(m) = \frac{m - 2d + 1}{(d-1)!}\prod_{i=2}^{d-1}(m - i).$$

Lemma 4.2.2. *The polynomial $p_d(m)$ satisfies*

$$\Delta p_d(m) = p_{d-1}(m)$$

and

$$p_d(2d+1) = C_d.$$

Proof. To prove the first equality, we compute

$$\begin{aligned}
\Delta p_d(m) &= p_d(m+1) - p_d(m) \\
&= \frac{m-2d+2}{(d-1)!} \prod_{i=2}^{d-1}(m+1-i) - \frac{m-2d+1}{(d-1)!} \prod_{i=2}^{d-1}(m-i) \\
&= \frac{(m-2d+2)(m-1) - (m-2d+1)(m-d+1)}{(d-1)!} \prod_{i=2}^{d-2}(m-i) \\
&= \frac{(d-1)(m-2d+3)}{(d-1)!} \prod_{i=2}^{d-2}(m-i) \\
&= p_{d-1}(m).
\end{aligned}$$

For the second equality, we have

$$\begin{aligned}
p_d(2d+1) &= \frac{2}{(d-1)!} \prod_{i=2}^{d-1}(2d+1-i) \\
&= \frac{2d}{d!} \cdot \frac{(2d-1)!}{(d+1)!} \\
&= \frac{(2d)!}{d!(d+1)!} \\
&= C_d
\end{aligned}$$

which finishes the proof. □

Note that by the criterion in Theorem 2.0.4, $\mathfrak{p}(m,d)$ can only be nonzero if $m > 2d$. We thank Richard Stanley who, on being shown Lemma 4.2.2, pointed out that $\mathfrak{p}(m,d)$ is a ballot number.

Theorem 4.2.3. *If $m, d \in \mathbb{P}$ with $m > 2d$ then $\mathfrak{p}(m, d) = p_d(m)$. Thus*

$$\mathfrak{p}(m, d) = \frac{m - 2d + 1}{m - 1}\binom{m - 1}{d - 1} = \#\mathcal{B}_{m-d,d-1}.$$

Proof. Induct on d where the base case of $d = 1$ is trivial to verify. To finish the first claim, it suffices to use the previous lemma and show that both $\Delta\mathfrak{p}(m, d) = \mathfrak{p}(m, d - 1)$ and $\mathfrak{p}(2d + 1, d) = C_d$. But these were proved in [DNKPT18, Sections 2.2–2.3]. The first displayed equality now follows from simple manipulation of the definition of $p_d(m)$, while the second comes from Theorem 4.2.1. □

We would like to give a bijective proof of the relationship between admissible pinnacle sets and ballot sequences from the previous theorem. Let

$$\mathfrak{P}(m, d) = \{S \mid S \text{ admissible with } \max S = m \text{ and } \#S = d\}$$

so that $\#\mathfrak{P}(m, d) = \mathfrak{p}(m, d)$. For $m > 2d$, define a map

$$\eta : \mathcal{B}_{m-d,d-1} \to \mathfrak{P}(m, d)$$

by sending ballot sequence $\beta = \beta_1\beta_2 \ldots \beta_{m-1}$ to

$$\eta(\beta) = \{i \mid \beta_i = Y\} \uplus \{m\}.$$

For example, if $m = 9$, $d = 3$ and $\beta = XXXYXXYX$ then

$$\eta(\beta) = \{4, 7\} \uplus \{9\} = \{4, 7, 9\}.$$

Theorem 4.2.4. *The map η is a well-defined bijection.*

Proof. We must first show that η is well defined in that $\eta(\beta) \in \mathfrak{P}(m, d)$. Since $\beta \in \mathcal{B}_{m-d,d-1}$ we see that the set $\{i \mid \beta_i = Y\}$ is contained in $[m - 1]$ and has cardinality $d - 1$. It follows that $S = \eta(\beta)$ has maximum m and cardinality d.

There remains to show that $S = \{s_1 < s_2 < \ldots < s_d\}$ is admissible. By Theorem 2.0.4, it suffices to show that $s_i > 2i$ for all i. But s_i is the index of the ith Y in β. Since β is a ballot sequence, this Y is preceded by i copies of Y (including itself) and at least $i + 1$ copies of X. So $s_i \geq i + (i + 1) = 2i + 1$ which is what we wished to prove.

To show that η is a bijection, we create its inverse. Given $S \in \mathfrak{P}(m, d)$ we define $\eta^{-1}(S) = \beta = \beta_1\beta_2\ldots\beta_{m-1}$ by letting

$$\beta_i = \begin{cases} X & \text{if } i \notin S, \\ Y & \text{if } i \in S. \end{cases}$$

The proof that η^{-1} is well defined is similar to the one for η. And proving that the compositions of η with η^{-1} are identity maps is easy. So we are done. □

4.3 Permutations with a given pinnacle set

Given an admissible set S, there does not seem to be an expression for $p_S(n)$, the number of permutations in $\mathfrak{S}_n$ with S as pinnacle set, analogous to the one in Theorem 2.0.3 for peak sets. In [DNKPT18], they found expressions for $p_S(n)$ when $\#S \leq 2$ as well as bounds for general S, and asked whether an exact formula could be given in the general case. Such an expression was given in [DLHH$^+$21] as a summation. In this section we will give another sum which is asymptotically more efficient. In addition, this method can be extended to count $\#O(S)$, the number of admissible orderings of S.

Since our sum will involve a significant amount of new notation, we will collect it here and then explain its relevance afterwards. Fix $n \in \mathbb{P}$. Suppose we have an admissible pinnacle set $S = \{s_1 < s_2 < \ldots < s_d\}$ for permutations in $\mathfrak{S}_n$. We use the convention $s_0 = 0$ and $s_{d+1} = n + 1$ and let

$$n_i = s_{i+1} - s_i - 1$$

for $0 \leq i \leq d$. Let

$$D = \{1_l, 1_r, 2_l, 2_r, \ldots, d_l, d_r\}$$

and give the following total order to D's elements

$$1_l < 1_r < 2_l < 2_r < \ldots < d_l < d_r.$$

We call i_l and i_r the elements of *rank* i in D. If $B \subseteq D$ then we will let

$$b = \#B$$

and

$$r_j = \text{the rank of the } j\text{th smallest element of } B$$

for $1 \leq j \leq b$. We also define

$$b_i = \text{the number of elements in } B \text{ with rank at least } i.$$

Note that we always have $b_1 = b$ and $b_{d+1} = 0$ since d is the largest rank. For example, if $d = 4$, then $D = \{1_l, 1_r, 2_l, 2_r, 3_l, 3_r, 4_l, 4_r\}$ and one possible B might be $B = \{1_l, 3_l, 3_r, 4_r\}$ which has $r_1 = 1, r_2 = 3, r_3 = 3, r_4 = 4$ and $b_1 = 4, b_2 = 3, b_3 = 3, b_4 = 1, b_5 = 0$. We can now state the first main result of this section.

Theorem 4.3.1. *Given $n \in \mathbb{P}$ and admissible $S = \{s_1 < s_2 < \ldots < s_d\}$ we have*

$$p_S(n) = 2^{n-2d-1} \sum_{B \subseteq D:\ |B| \leq d} (-1)^b (d-b)! \left(\prod_{i=0}^{b-1} (d+1-i-r_{b-i}) \right) \left(\prod_{i=0}^{d} (d+1-i-b_{i+1})^{n_i} \right).$$

To prove this, it will be convenient to convert the linear permutations we have been studying into cyclic ones in order to avoid considering boundary cases. Given a linear permutation $\pi = \pi_1 \pi_2 \ldots \pi_n$ the corresponding *cyclic permutation* is the set of permutations

$$[\pi] = \{\pi_1 \pi_2 \ldots \pi_n, \quad \pi_2 \ldots \pi_n \pi_1, \quad \ldots, \quad \pi_n \pi_1 \ldots \pi_{n-1}\}.$$

Intuitively, we think of $[\pi]$ as the result of arranging the elements of π on a circle. Let

$$[\mathfrak{S}_n] = \{[\pi] \mid \pi \in \mathfrak{S}_n\}.$$

For example if $\pi = 1324$ then

$$[\pi] = \{1324,\ 3241,\ 2413,\ 4132\}.$$

We are also using the bracket notation in $[n]$ where $n \in \mathbb{N}$ but this should not cause any confusion. Cyclic permutations are of interest in part because of their relation with pattern avoidance, standard Young tableaux, quasisymmetric functions, and other mathematical objects [AGRR20, Cal02, DLM$^+$21a, DLM$^+$21b, GLW18, GLW19].

We define the *pinnacle set* of $[\pi] = [\pi_1\pi_2 \dots \pi_n]$ to be

$$\mathrm{Pin}[\pi] = \{\pi_i \mid \pi_{i-1} < \pi_i > \pi_{i+1} \text{ where subscripts are taken modulo } n\}.$$

Continuing our example from the last paragraph

$$\mathrm{Pin}[1324] = \{3, 4\}.$$

Note in particular that $\mathrm{Pin}[12] = \{2\}$ and, more generally, $n \in \mathrm{Pin}[\pi]$ for any $[\pi] \in [\mathfrak{S}_n]$ where $n \geq 2$.

Lemma 4.3.2. *For $n \in \mathbb{P}$, there is a bijection between linear permutations in $\mathfrak{S}_n$ with pinnacle set S and cyclic permutations in $[\mathfrak{S}_{n+1}]$ with pinnacle set $S' = S \cup \{n+1\}$.*

Proof. Given a linear π, append the element $n+1$ to the end of π and take the corresponding equivalence class in $\mathfrak{S}_{n+1}$ to form an element of $[\mathfrak{S}_{n+1}]$. The map is clearly invertible and does not destroy or create any pinnacles for elements in $[n]$. Since $n+1 \geq 2$, we know that $n+1$ will become a pinnacle. Therefore the map has the desired properties concerning the pinnacle set. □

Consider some admissible pinnacle set $S = \{s_1 < s_2 < \ldots < s_d\}$. Given the above lemma, we may count the number of permutations in $\mathfrak{S}_n$ with pinnacle set S by counting the number of cyclic permutations $[\pi] \in [\mathfrak{S}_{n+1}]$ with pinnacle set $S' = S \cup \{n+1\}$ where we let $s_{d+1} = n+1$. Therefore, much of what follows will be in regards to cyclic permutations with pinnacle set S'.

A *factor* of a (cyclic) permutation is a subsequence of consecutive elements. We may attempt to construct a $[\pi]$ with pinnacle set S' by first putting the elements of S' in some cyclic order, and then placing all elements in $\overline{S'} = [n+1] - S'$ into either decreasing factors starting with some s_i, or into increasing factors ending with some s_i. Such a $[\pi]$ will then be completely determined by the increasing/decreasing factors that each element of $\overline{S'}$ falls into, and we will call every such assignment a *placement*. Note that it is possible for multiple placements to result in the same permutation since each *vale* (an element of $[\pi]$ smaller than the elements on either side) can be part of the factor on either side. For example, start with a desired pinnacle set $\{4, 5\}$ and place

non-pinnacles between these elements to form the cyclic permutation $[\pi] = [14325]$. Then $[\pi]$ would be associated with a placement where the decreasing factor starting with 4 is 43 and the increasing factor ending with 5 is 25. But it would also be associated with a placement having these factors be 432 and 5, respectively.

It is also possible, depending on the placement, that $[\pi]$ will not have pinnacle set S' if no sufficiently small elements are placed between two pinnacles. In our example above, this could have happened if we had placed 1, 2 and 3 all in the increasing factor ending in 5, resulting in the cyclic permutation [41235] in which only 5 is a pinnacle. It is true, however, that any $[\pi]$ so constructed will have a pinnacle set that is a subset of S' since every non-pinnacle was placed so that its factor contains an s_i which is the largest element. For our arguments, we will focus on counting placements and then convert them into permutations later.

Fix a cyclic ordering of the pinnacle indices and write it as $[\tau] = [\tau_1 \cdots \tau_{d+1}] \in [\mathfrak{S}_{d+1}]$. An example is shown in Figure 4.2 where $\tau = [7612354]$. Now given a placement consistent with this ordering, for every space between two adjacent elements in $[\tau]$ define the *dale set* of this placement to consist of all elements between the two corresponding pinnacles that are also smaller than both pinnacles. So in Figure 4.2 the dales are outlined by triangles with solid lines as sides. If s_i is the smaller of the two pinnacles, then we say that the dale has *rank* i. Note that the rank is from the index of s_i and not its actual value. We will further denote the rank as either i_l or i_r depending on whether the dale is to the left, or right of the pinnacle s_i. In Figure 4.2 the dale ranks are given along the x-axis. Define the *dale rank set* $D_{[\tau]}$ to be the set of the dale ranks of $[\tau]$. And define the *master dale rank set* to be

$$D = \{1_l, 1_r, 2_l, 2_r, \ldots, d_l, d_r\}$$

so that $D \supseteq D_{[\tau]}$ for all $[\tau]$. In Figure 4.2, we have that $D_{[\tau]} = \{1_l, 1_r, 2_r, 3_r, 4_1, 4_r, 6_l\}$ while $D = \{1_l, 1_r, 2_l, 2_r, \ldots, 6_l, 6_r\}$. Note that, by our definitions, there will be no dales in the case where $d = 0$.

Clearly $D_{[\tau]}$ will be a subset of D consisting of exactly $d + 1$ elements if $d > 0$, and empty otherwise. We can derive further information about $D_{[\tau]}$ if we want, such as how it will always

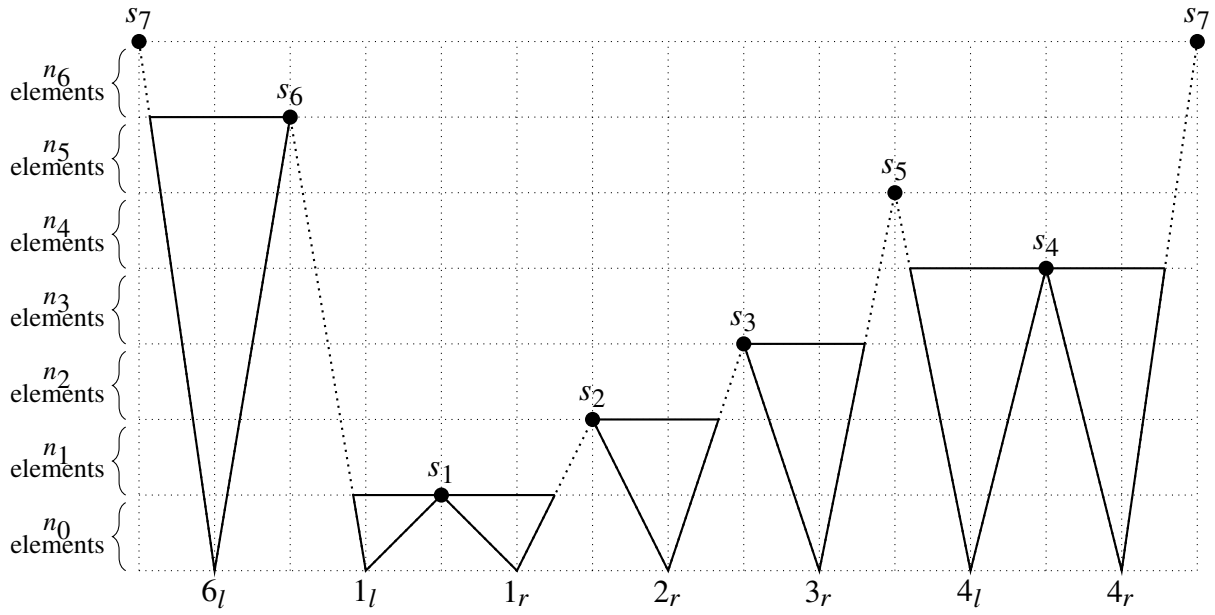

Figure 4.2: Example of a pinnacle set ordering $[\tau] = [7612354]$ with corresponding dales.

contain both 1_l and 1_r if $d > 0$, how it will never contain both d_l and d_r if $d > 1$, and how $D_{[\tau]}$ will never be able to have certain combinations of the higher ranked dales. These facts are not necessary for proving our formula, although further analysis of them might help to improve its efficiency.

Lemma 4.3.3. *For $n \in \mathbb{P}$, a given placement will correspond to a permutation $[\pi] \in [\mathfrak{S}_{n+1}]$ with pinnacle set S' if and only if every dale is non-empty.*

Proof. First, suppose $d = 0$. In this case, the theorem is trivial since there are no dales. And every placement will automatically result in only one pinnacle, namely $n + 1$, as long as $n > 0$.

Now suppose $d > 0$. Clearly if any dale of rank i (whether left or right) is empty, then the pinnacle s_i will have no smaller elements between itself and the higher pinnacle next to it, which will force s_i to not be a pinnacle. On the other hand, if all dales have at least one element, then the space between any two pinnacles will always contain an element smaller than both, and all elements of S' will in fact be pinnacles. □

We can now enumerate all placements corresponding to a given cyclic ordering of the indices of the pinnacle set S'.

Lemma 4.3.4. *Given an admissible pinnacle set S', fix an order $[\tau]$ of the pinnacle indices. The total number of placements with order $[\tau]$ that will result in a permutation with pinnacle set S' is given by*

$$\sum_{B \subset D_{[\tau]}} (-1)^b \prod_{i=0}^{d} 2^{n_i} (d + 1 - i - b_{i+1})^{n_i}$$

where b, d, the b_i, and the n_i are defined above.

Proof. We will use the Principle of Inclusion and Exclusion or PIE. We let our universal set be all possible placements with no restrictions. We then wish to exclude any placement where at least one dale is empty. Therefore, if B is some subset of the dales, we must be able to count the number of placements where all dales in B (and possibly others) are empty.

First consider the case when $B = \emptyset$. There are $2(d + 1)$ factors of which $2i$ only exist below s_i. So each of the n_i non-pinnacles between s_i and s_{i+1} may be placed in any of the $2(d + 1 - i)$ factors that are long enough to extend above s_i. As an example, in fig. 4.2, if we look between the horizontal boundary lines for the elements counted by n_2 we see there are $10 = 2(6 + 1 - 2)$ such factors represented by the diagonal lines (solid or dotted) which intersect the region.

For non-empty B, each dale of rank at least $i + 1$ that we require to be empty will result in a loss of two additional factors, and so there are only $2(d + 1 - i - b_{i+1})$ choices. Therefore, for a given B, the total number of placements guaranteeing the dales in B are empty is

$$\prod_{i=0}^{d} 2^{n_i} (d + 1 - i - b_{i+1})^{n_i}.$$

To use the PIE, we must also attach the sign $(-1)^{|B|} = (-1)^b$ to this term before summing. Therefore, given a fixed order $[\tau]$ of the pinnacle indices of S', we have that the total number of placements that will result in a permutation with pinnacle set S' is

$$\sum_{B \subseteq D_{[\tau]}} (-1)^b \prod_{i=0}^{d} 2^{n_i} (d + 1 - i - b_{i+1})^{n_i}.$$

Finally, when $B = D_{[\tau]}$ then $b_1 = \#B = d + 1$. So we can ignore this term because the product has a factor of $d + 1 - b_1 = 0$. □

The above formula must be summed over all possible $[\tau]$ to give a final count for the number of $[\pi]$ with $\mathrm{Pin}[\pi] = S'$. This results in computationally expensive double sum. Also, note that in the above formula there may be multiple B resulting in the same term. For example, $\{1_l, 2_r, 5_l\}$ is not the same as $\{1_r, 2_r, 5_l\}$ even though both produce the same b_i. We will take care of this redundancy when we optimize our formula below.

To fix the double sum problem, note that each B in Lemma 4.3.4 is a subset of the master dale rank set D. We will fix some subset $B \subseteq D$ and count the number of orderings $[\tau]$ that will produce a $D_{[\tau]}$ which can have B as a subset. This will allow us to just sum over all subsets $B \subseteq D$ without having to keep track of $[\tau]$. Furthermore, we only have to sum over the subsets B of cardinality at most d since requiring more than d dales to be empty is impossible for an admissible pinnacle set.

Lemma 4.3.5. *Fix some $B \subseteq D$ with $|B| \leq d$. The number of orderings $[\tau]$ that will produce a $D_{[\tau]}$ such that $B \subseteq D_{[\tau]}$ is given by*

$$(d-b)! \prod_{i=0}^{b-1} (d+1-i-r_{b-i})$$

where b, d, and the r_i are defined as above.

Proof. We will start by viewing all $d+1$ pinnacles as separate and then adjoin them in pairs in such a way so that the desired dales are formed. Here, "adjoining a pair of pinnacles" means requiring that they be adjacent in $[\tau]$.

We start with the dale of rank r_b the largest rank in B. In that case, the only way to generate such a dale is to order s_{r_b} so that one of the $d + 1 - r_b$ higher pinnacles is directly to its left or right depending on whether the corresponding element of B is a left or right rank, respectively. So select one such pinnacle and adjoin it to the appropriate side of s_{r_b}.

Next we will examine the dale in B with the next highest rank, r_{b-1}. If r_{b-1} is a smaller rank than r_b, we may once again select a taller pinnacle to place next to $s_{r_{b-1}}$, on either the left or right as necessary, in order to produce the desired dale. This time however, although there are $d+1-r_{b-1}$ pinnacles higher than $s_{r_{b-1}}$, one of them is unavailable since we have already adjoined two of the higher-ranked pinnacles together. More specifically, because of adjoining a higher pinnacle with

s_{r_b}, we know that one taller pinnacle cannot be joined to its left and another cannot be joined to its right. So no matter whether r_{b-1} corresponded to a left or right dale, there is one less option. Therefore, the number of ways to append a larger pinnacle is $d+1-r_{b-1}-1$. On the other hand, if $r_{b-1} = r_b$ then we need to adjoin a second pinnacle to s_{r_b} on the side opposite the one used when considering r_b. Again, the pinnacle already adjoined to s_{r_b} removes one option so the number of choices is $d+1-r_{b-1}-1$ as before. So in either case we have the same number of possibilities. Similar consideration show that, in general, each r_{b-i} results in $d+1-i-r_{b-i}$ choices for adjoining pinnacles. Note that for this argument we are using the fact that $b \leq d$ since if $b = d+1$ then the string of pinnacles would wrap into a circle before creating the final dale.

Once all dales have been created by the above process, we only need to count the number of ways to join the resulting strings of pinnacles together. Since we have adjoined pinnacles together b times, we have $d+1-b$ strings which we then must arrange in a circle. This can be done in $(d+1-b-1)! = (d-b)!$ ways. Therefore,

$$(d-b)! \prod_{i=0}^{b-1} (d+1-i-r_{b-i})$$

is the number of orderings $[\tau]$ that will allow for a given B to be a subset of $D_{[\tau]}$. □

We are now in a position to prove Theorem 4.3.1 which we restate here for ease of reference.

Theorem 4.3.6. *Given $n \in \mathbb{P}$ and admissible $S = \{s_1 < s_2 < \ldots < s_d\}$ we have*

$$p_S(n) = 2^{n-2d-1} \sum_{B \subseteq D:\ |B| \leq d} (-1)^b (d-b)! \left(\prod_{i=0}^{b-1} (d+1-i-r_{b-i}) \right) \left(\prod_{i=0}^{d} (d+1-i-b_{i+1})^{n_i} \right).$$

Proof. It is easy to verify the formula if $d = 0$, so we assume $d > 0$. From Lemma 4.3.3, the number of permutations $\pi \in \mathfrak{S}_{n+1}$ with pinnacle set S equals the number of cyclic permutations $[\pi] \in [\mathfrak{S}_{n+1}]$ with pinnacle set $S' = S \cup \{n+1\}$. So we will count the latter. From Lemma 4.3.4, the number of placements which correspond to a cyclic permutation with pinnacle set S' is given by

$$\sum_{[\tau]} \sum_{B \subseteq D_{[\tau]}} (-1)^b \prod_{i=0}^{d} 2^{n_i} (d+1-i-b_{i+1})^{n_i}$$

where the outer sum is over all possible cyclic orderings $[\tau]$ of the index set of S'. We now wish to swap the summations so that the outer sum is over all $B \subseteq D$ with $|B| \leq d$. We may restrict to size at most d since any larger B will either consist of a combination of dales that cannot exist, or will require all $d+1$ dales to be empty which is impossible because of the assumption that $d > 0$. In order to interchange the summations we must multiply the term corresponding to each B by the number of distinct permutations $[\tau]$ that could have generated it. This was counted in Lemma 4.3.5, and so we get the formula

$$\sum_{B \subset D:\ |B| \leq d} (-1)^b (d-b)! \left(\prod_{i=0}^{b-1} (d+1-i-r_{b-i}) \right) \left(\prod_{i=0}^{d} 2^{n_i} (d+1-i-b_{i+1})^{n_i} \right)$$

for the number of placements.

Now we seek to turn the placements into permutations. Since all dales are guaranteed to be non-empty, we have that every permutation corresponding to one of these placements will have $d+1$ non-pinnacle elements that are part of both a decreasing factor and an increasing factor. This means that every such corresponding $[\pi]$ has been counted by 2^{d+1} placements. Dividing by this, and also pulling some common factors of two out from the second product, we have

$$\begin{aligned} p_S(n) &= 2^{-d-1} \prod_{i=0}^{d} 2^{n_i} \sum_{B \subseteq D:\ |B| \leq d} (-1)^b (d-b)! \left(\prod_{i=0}^{b-1} (d+1-i-r_{b-i}) \right) \left(\prod_{i=0}^{d} (d+1-i-b_{i+1})^{n_i} \right) \\ &= 2^{n-2d-1} \sum_{B \subseteq D:\ |B| \leq d} (-1)^b (d-b)! \left(\prod_{i=0}^{b-1} (d+1-i-r_{b-i}) \right) \left(\prod_{i=0}^{d} (d+1-i-b_{i+1})^{n_i} \right) \end{aligned}$$

where this is the formula we set out to prove. □

In [DNKPT18], explicit formulas were given for $p_S(n)$ when $|S| \leq 3$. These expressions follow easily from the previous reslt.

Corollary 4.3.7. *We have the following values for $p_S(n)$.*

(1) If $S = \emptyset$ then

$$p_S(n) = 2^{n-1}.$$

(2) If $S = \{l\}$ where $3 \leq l \leq n$ then

$$p_S(n) = 2^{n-2}(2^{l-2} - 1).$$

(3) If $S = \{l, m\}$ where $l \geq 3$, $m \geq 5$, and $l < m \leq n$, then

$$p_S(n) = 2^{n+m-l-5}(3^{l-1} - 2^l + 1) - 2^{n-3}(2^{l-2} - 1).$$

Proof. In each of the results we apply Theorem 4.3.6.

(1) When $d = 0$, the first product in Theorem 4.3.6 is always empty and the second always equals one. Therefore, everything reduces immediately to $p_S(n) = 2^{n-1}$, as desired.

(2) When $d = 1$ we have $n_0 = l - 1, n_1 = n - l$. Therefore, we have the following possibilities for B, and the corresponding terms in the summation

- $B = \emptyset : 2^{l-1}$
- $B = \{1_l\}$ or $\{1_r\}$: -1

which when substituted into the formula gives

$$p_S(n) = 2^{n-3}(2^{l-1} - 2) = 2^{n-2}(2^{l-2} - 1).$$

(3) When $d = 2$ we have $n_0 = l - 1, n_1 = m - l - 1$, and $n_2 = n - m$. Additionally, the first inner product will always zero out if $2_r, 2_l$ are both in B. Therefore, we have the following possibilities for B, and the corresponding terms in the summation:

- $B = \emptyset : (2)3^{l-1}2^{m-l-1}$
- $B = \{1_l\}$ or $\{1_r\}$: $(-2)2^{l-1}2^{m-l-1}$
- $B = \{2_l\}$ or $\{2_r\}$: $(-1)2^{l-1}$
- $B = \{1_l, 1_r\}$: 2
- $B = \{1_l, 2_r\}$ or $\{1_r, 2_r\}$ or $\{1_l, 2_l\}$ or $\{1_r, 2_l\}$: 1.

When we substitute all these into the formula, we get

$$\begin{aligned} p_S(n) &= 2^{n-5}[(2)3^{l-1}2^{m-l-1} - (4)2^{l-1}2^{m-l-1} - (2)2^{l-1} + (2)2^{m-l-1} + 4] \\ &= 2^{n-5}[(2)3^{l-1}2^{m-l-1} - (4)2^{l-1}2^{m-l-1} + (2)2^{m-l-1}] - 2^{n-5}[(2)2^{l-1} - 4] \\ &= 2^{n+m-l-5}(3^{l-1} - 2^l + 1) - 2^{n-3}(2^{l-2} - 1) \end{aligned}$$

as desired. □

We can make Theorem 4.3.6 more efficient by summing over certain weak compositions rather than subsets. A *weak composition* of $n \in \mathbb{N}$ is a sequence $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_k]$ of nonnegative integers called *parts* such that $\sum_i \alpha_i = n$. In this case we write $\alpha \models n$ or $|\alpha| = n$ where $|\alpha| = \sum_i \alpha_i$. To $B \subseteq D$ we associate the composition $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_d]$ where α_i is the number of dales in B of rank i. To illustrate, for the example in Figure 4.2 the corresponding composition is $\alpha = [2, 1, 1, 2, 0, 1]$. Note that all the necessary parameters for D can be read off of α. In particular

$$r_j = \min\{i \mid \alpha_1 + \alpha_2 + \cdots + \alpha_i \geq j\},$$

and

$$b_i = \alpha_i + \alpha_{i+1} + \cdots + \alpha_d.$$

Note that

$$b = b_1 = |\alpha|.$$

Thus we will be able to sum over the following set

$$C(d) = \{\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_d] \mid \alpha_i \in [0, 2] \text{ for all } i \text{ and } |\alpha| \leq d\}.$$

We must find how many B correspond to a given α. If $\alpha_i = 0$ then B contains no dales of rank i. If $\alpha_i = 2$ then B contains both dales of rank i. So the only choice comes if $\alpha_i = 1$ in which case B could contain either i_l or i_r. Letting

$$o = \text{the number of } \alpha_i = 1$$

we see that the number of B represented by α is 2^o. Thus we have proved the following result.

Corollary 4.3.8. *Given $n \in \mathbb{P}$ and admissible $S = \{s_1 < s_2 < \ldots < s_d\}$ we have*

$$p_S(n) = 2^{n-2d-1} \sum_{\alpha \in C(d)} (-1)^b 2^o (d-b)! \left(\prod_{i=0}^{b-1} (d+1-i-r_{b-i}) \right) \left(\prod_{i=0}^{d} (d+1-i-b_{i+1})^{n_i} \right). \quad \square$$

In order to compare this formula to the one in [DLHH$^+$21], we need to introduce some notation. The *vale set* of a permutation π is

$$\operatorname{Val} \pi = \{\pi_i \mid \pi_{i-1} > \pi_i < \pi_{i+1}\}.$$

Call a pair (S, T) *n-admissible* if there is a permutation $\pi \in \mathfrak{S}_n$ with $\operatorname{Pin} \pi = S$ and $\operatorname{Val} \pi = T$. Define

$$\mathcal{V}_n(S) = \{T \mid (S, T) \text{ is } n\text{-admissible}\}.$$

Theorem 4.3.9 ([DLHH$^+$21])**.** *Given $n \in \mathbb{P}$ and admissible S with $\#S = d$ we have*

$$p_S(n) = 2^{n-2d-1} \sum_{T \in \mathcal{V}_n(S)} \prod_{s \in S} \binom{N_{ST}(s)}{2} \prod_{t \in [n]-(S \uplus T)} N_{ST}(t)$$

where $S_i = \{s \in S \mid s < i\}$, $T_i = \{t \in T \mid t < i\}$, and $N_{ST}(i) = \#T_i - \#S_i$. $\square$

In order to estimate the number of terms in this sum, we need a formula for $\#\mathcal{V}_n(S)$. Let

$$K(d) = \{\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_d] \models d \mid \alpha_1 + \alpha_2 + \cdots + \alpha_k \geq k \text{ for all } k \in [d]\}.$$

Theorem 4.3.10 ([DLHH$^+$21])**.** *Given $n \in \mathbb{P}$ and admissible $S = \{s_1 < s_2 < \ldots < s_d\}$ we have*

$$\#\mathcal{V}_n(S) = \sum_{\alpha \in K(d)} \binom{n_0 - 1}{\alpha_1} \prod_{i=2}^{d} \binom{n_{i-1}}{\alpha_i}. \quad \square$$

We can now compare the number of terms in the sums of Corollary 4.3.8 and Theorem 4.3.9. In the former we have $c(d) := \#C(d) \leq 3^d$ terms, where the inequality comes from the fact that every $\alpha_i \in \{0, 1, 2\}$. In the latter, we have $v_n(S) := \#\mathcal{V}_n(S)$ terms which depends on n and S, and not just d as seen in Theorem 4.3.10. If $n_1 \leq 4$ and $n_i \leq 3$ for $i \geq 2$ then each of the binomial coefficients in the sum is a most 3 and so $v_n(S)$ could be significantly smaller than $c(d)$. But if

S	DLHHIN	DLMSSS
$\{3,5,7,9,11,13,15,17,19,21\}$	9.2×10^{-5}	0.72
$\{3,6,9,12,15,18,21,24,27,30\}$	0.11	0.73
$\{3,7,11,15,19,23,27,31,35,39\}$	9.5	0.73
$\{3,8,13,18,23,28,33,38,43,48\}$	210	0.78

Table 4.1: Run times in seconds compared when most n_i are equal

even one of the n_i is large, then the inequality will be reversed. For example, suppose $n_1 \geq 2d+1$ and take $\alpha = [d, 0, 0, \ldots, 0] \in K(d)$. Then, by Stirling's approximation,

$$v_n(S) \geq \binom{2d}{d} \sim \frac{4^d}{\sqrt{\pi d}}$$

which will eventually be greater than 3^d. So, for fixed d, there are only finitely many n such that $v_n(S) \leq c(d)$. Thus, in most cases, Corollary 4.3.8 will be more efficient. We should mention that Diaz-Lopez, Insko, and Nilsen [DLIN21] have come up with a refinement of the ideas in [DLHH$^+$21] which permits the product of binomial coefficients in Theorem 4.3.10 to be replaced by 2^d.

The observations of the previous paragraph are borne out by actual computer computations. In Tables 4.1 and 4.2 we show the results of computing $p_S(1000)$ for various sets S (first column) with constant d by the algorithm in [DLHH$^+$21] (second column) and our algorithm (third column). The run times are in seconds and are the average over 10 trials for each set using a 15-inch 2017 MacBook Pro with a 3.1 GHz Quad-Core Intel Core i7 processor. In Table 4.1 the n_i for $0 < i < d$ are constant in each set, but allowed to increase as one goes down the table. As expected, the algorithm using vales starts out orders of magnitude faster than the one using dales but quickly becomes orders of magnitude slower, with the latter's times being virtually constant. Similar behaviour is shown in the two parts of Table 4.2 which keep all of the n_i for $0 \leq i < d$ constant except for one which is allowed to grow. Note the difference in growth rate of the vale algorithm between increasing n_4 (upper chart) and n_0 (lower chart).

Increase n_4 with other n_i constant

S	DLHHIN	DLMSSS
$\{3,5,7,9,11\}$	2.9×10^{-5}	0.0014
$\{3,5,7,9,21\}$	7.1×10^{-5}	0.0014
$\{3,5,7,9,31\}$	0.00012	0.0015
$\{3,5,7,9,41\}$	0.00017	0.0015

Increase n_0 with other n_i constant

S	DLHHIN	DLMSSS
$\{3,5,7,9,11\}$	2.9×10^{-5}	0.0014
$\{13,15,17,19,21\}$	0.012	0.0015
$\{23,25,27,29,31\}$	0.26	0.0015
$\{33,35,37,39,41\}$	1.8	0.0015

Table 4.2: Run times in seconds compared when most n_i are constant

Another advantage to this approach is that it can be modified to count $\#\mathcal{O}(S)$, the number of admissible orderings of an admissible pinnacle set S. First, if we fix $n > 0$ we have that Lemma 4.3.2 will again allow us to reduce to the case of cyclic orderings of the pinnacle set S' for permutations in $\mathfrak{S}_{n+1}$. We now prove the following intermediate result.

Lemma 4.3.11. *Consider a cyclic ordering $[\tau]$ with dale set $D_{[\tau]}$ and corresponding r_j. The ordering is admissible if and only if*

$$j \leq n_0 + n_1 + \cdots + n_{r_j - 1}$$

for all $j \in [d+1]$.

Proof. Note that, by definition of the n_i and r_i, the right hand side of the inequality is simply the number of non-pinnacles small enough to be placed in any of the dales having rank at least r_j. So if for any j we have $j > n_0 + n_1 + \cdots n_{r_j - 1}$, then there will be at least $j+1$ dales having rank at most r_j. This means there would not be enough small non-pinnacle elements to fill them all. Therefore,

any such ordering is not admissible. On the other hand, if we have that $j \leq n_0 + n_1 + \cdots n_{r_j - 1}$ for all j, then we may always fill all the dales by placing the smallest non-pinnacle in the lowest ranked dale, and proceeding upwards. The inequalities guarantee that we will always have enough non-pinnacles to do this at every step, and so we are done. □

Since the problem is trivial if $d = 0$, so we may also assume that $d > 0$. We also define for the master dale rank set D

$$D' = D - \{1_l, 1_r\}$$

and for any subset B

$$\delta_B = \begin{cases} 1 & \text{if } j \leq n_0 + n_1 + \cdots + n_{r_j - 1} \text{ for all } j \in [b], \\ 0 & \text{otherwise.} \end{cases}$$

With this notation, we can count admissible orderings.

Theorem 4.3.12. *If $d \in \mathbb{P}$ and S is admissible then*

$$\#\mathcal{O}(S) = \sum_{B \subseteq D' \colon |B| = d-1} \delta_{B \cup \{1_l, 1_r\}} \prod_{i=0}^{d-2} (d + 1 - i - r_{d-1-i}) \, .$$

Proof. We first wish to sum over all possible orderings, partitioned by their dales. Since every dale set for $d > 0$ is guaranteed to have $d + 1$ elements and contain $\{1_l, 1_r\}$, we may index the dales by taking $B \subseteq D'$ where $|B| = d - 1$. We then consider the following summation

$$\sum_{B \subseteq D' \colon |B| = d-1} \prod_{i=0}^{d-2} (d + 1 - i - r_{d-1-i}) \, .$$

Clearly this sums over every possible dale set once, and the expression inside comes from Lemma 4.3.5, which counts the number of cyclic orderings of S' which have dales containing those in B. However, due to the restrictions placed on the size of B and the comments above, this expression will count those cyclic orderings of S' which have dales equal to $B \cup \{1_l, 1_r\}$ instead of just a subset. Therefore, no ordering can be counted twice by two different B's and so every ordering is accounted for exactly once in the above summation, making the total $d!$.

Finally, using Lemma 4.3.11, we may exclude from this sum precisely those orderings which are not admissible by writing it as

$$\sum_{B\subseteq D':\,|B|=d-1} \delta_{B\cup\{1_l,1_r\}} \prod_{i=0}^{d-2} (d+1-i-r_{d-1-i})\,.$$

This completes the proof. □

We may also rewrite our result in terms of compositions for a faster summation. Lemma 4.3.11 still holds as the r_j are the same whether or not the dales sets are represented as compositions, but now we will need make some new definitions. Let

$$C'(d) = \{\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_{d-1}] \models [d-1] \;\mid\; \alpha_i \in [0,2] \text{ for all } i\},$$

and if $\alpha \models b$

$$\delta_\alpha = \begin{cases} 1 & \text{if } j \le n_0 + n_1 + \cdots + n_{r_j-1} \text{ for all } j \in [b], \\ 0 & \text{otherwise.} \end{cases}$$

Also, if $\alpha = [\alpha_1, \alpha_2, \ldots, \alpha_{d-1}]$ then define

$$2 \oplus \alpha = [2, \alpha_1, \alpha_2, \ldots, \alpha_{d-1}].$$

The following result follows from Theorem 4.3.12 in much the same way that Corollary 4.3.8 followed from Theorem 4.3.6. So the proof is omitted.

Corollary 4.3.13. *If $d \in \mathbb{P}$ and S is admissible then*

$$\#O(S) = \sum_{\alpha\in C'(d)} \delta_{2\oplus\alpha} 2^o \prod_{i=0}^{d-2} (d+1-i-r_{d-1-i}).$$ □

4.4 Open problems and concluding remarks

Others have also been working on finding a fast formula for computing the number of permutations with a given pinnacle set. Recently, Falque, Novelli, and Thibon [FNT21] have constructed an efficient recursion to compute $p_S(n)$. This formula is a low degree polynomial in both m, the

maximum of the pinnacle set S and d, the cardinality of the set, and has complexity $O(md^2)$. While the result was originally stated in terms of n instead of m, we can simplify in the following way. A permutation in $\mathfrak{S}_m$ with pinnacle set S can be extended to a permutation in $\mathfrak{S}_{m+1}$ by placing $m+1$ at either the far left or the far right of the permutation. Any other way of inserting $m+1$ into the permutation would make $m+1$ a pinnacle. Recursively applying this procedure to $\pi \in \mathfrak{S}_m$ with the elements $\{m+1, m+2, \ldots, n\}$ will extend it to a permutation $\pi' \in \mathfrak{S}_n$. Since there were two possible positions to place each of the elements $\{m+1, m+2, \ldots, n\}$, we have $p_n(S) = 2^{n-m} p_m(S)$ and thus the result can be stated in terms of m. In addition, they provide a conjectured formula for the weighted sum introduced in [DNKPT18]:

$$q_S(n) := \sum_{I \subset S} 2^{|I|} p_I(n). \tag{4.3}$$

Following this work, Fang [Fan21] provided another recurrence to compute $p_S(n)$ with complexity $O(d^4 + d \log n)$. He also proved an expression for eq. (4.3) which is simpler than the earlier conjecture and which is very combinatorial in nature. Quinn Minnich has recently found a simpler proof of this result.

CHAPTER 5

BACKGROUND ON BACKBONE EXTRACTION

Bipartite or two-mode networks are composed of two types of nodes, which we call *agents* and *artifacts*, and edges between nodes of one type and nodes of the other type. They can be used to represent a wide range of phenomena and therefore are studied in a diverse range of disciplines. For example, natural selection unfolds as species (the agents) compete over sites (the artifacts), commerce is possible as traders exchange resources, scientific advances are reported as scholars write papers, and laws are adopted as legislators sponsor bills. Although bipartite networks are useful in their own right, they can also be useful for inferring unipartite (i.e., one-mode) networks that would otherwise be difficult or impossible to measure directly. A bipartite projection transforms a bipartite network into a unipartite co-occurrence network in which agents are connected to the extent that they share artifacts. For example, competitive interaction networks can be inferred from species' co-occurrence in sites [Dia75], trade networks can be inferred from firm co-location [TCW02] or product co-exchange [SDCGS15], scholarly collaboration networks can be inferred from paper co-authorship [New01], and political alliance networks can be inferred from bill co-sponsorship [Nea20]. Throughout this thesis, we use these applications to offer concrete examples, however the models we discuss are perfectly general and can be applied to derive unipartite backbones in a range of contexts [AABB11, Tol21, ZH05]. Indeed, in principle any unipartite network can be represented as the projection of some bipartite network [VFO20, GL04, NP03] .

Despite their promise, bipartite projections (i.e., co-occurrence networks) are challenging to analyse because they are typically dense and weighted, and because the edge weights do not necessarily capture the strength of the relationship between nodes [Nea14]. In particular, when transforming a bipartite graph into a unipartite graph via projection, information about the artifacts responsible for edges between vertices is lost [LMDV08], specifically, one no longer knows *which* artifact(s) gave rise to a given edge and therefore no longer knows whether the artifact(s) are large

or small (i.e. the column sums of the bipartite matrix). This is important because co-participation in small artifacts provides more information about the relationship between two vertices than co-participation in large artifacts [Nea14]. For example, observing two people attending the same small party provides more information about a potential social relationship between them than observing these individuals attending the same large gathering. Similarly, observing two legislators co-sponsoring the same unpopular bill (i.e. one that is co-sponsored by no one else) provides more information about a potential political relationship between them than observing these legislators co-sponsoring the same popular bill (i.e. one that is co-sponsored by many others also).

Bipartite projection also involves the loss of information about the individual vertices, one no longer knows *how many* artifacts a given vertex participated in (i.e. the row sums of the bipartite matrix). This information is important to consider because the scale of each edge weight in a bipartite projection is driven by the number of artifacts participated in by the two vertices it connects [Nea14]. For example, on average the number of events co-attended by two people who each attend many events will be larger (on average) than the number of events co-attended by two people who each attend few events. Similarly, on average the number of bills co-sponsored by two legislators who each sponsor many bills will be larger (on average) than the number of bills co-sponsored by two legislators who each sponsor few bills. Therefore, what counts as a 'large' or 'small' number of co-attendances or co-sponsorships depends in part on the total number of attendances or sponsorships of both members of a dyad. As we will see, the backbone extraction methods we consider cope with these challenges by controlling for the row and column sums of the bipartite matrix associated with the bipartite graph in question.

As a result of these challenges, it is often useful to analyze the *backbone* of a bipartite projection, which is an unweighted and typically sparser network that retains only the most 'important' edges. Although well-known methods exist for extracting the backbone of weighted networks that are not bipartite projections [SBV09, Dia16], methods designed specifically for bipartite projections have recently been developed [Nea14, ZK11, SSDC$^+$17, TML$^+$11].

To begin, we'll define notation and language for discussing bipartite projections and backbones.

Throughout this chapter, we use the ecological case of Darwin's Finches to provide a concrete example [San00, Got00]. On his voyage to the Galapagos Islands on the H.M.S. Beagle, Darwin observed that only some species of finches lived on each island. These patterns can be represented as a bipartite network in which finch species (the agent nodes) are connected to the islands (the artifact nodes) where they are found [NN20]. A bipartite network can be represented as a binary matrix in which the agents are arrayed as rows, and the artifacts are arrayed as columns. We use $\mathbf{B}$ to denote a bipartite network's representation as a matrix, where $B_{ik} = 1$ if agent i is connected to artifact k, and otherwise is 0. The sequence of row sums and the sequence of column sums of $\mathbf{B}$ are called the agent and artifact degrees sequences, respectively. These sequences are among the bipartite network's most significant features and are known to have implications for bipartite projections and backbones [VFO20, DNS21, NDS21a]. In the ecological case, the agent degree sequence captures the number of islands where each species is found, while the artifact degree sequence captures the number of species found on each island.

The *projection* of a bipartite network is a weighted unipartite co-occurrence network in which a pair of agents is connected by an edge with a weight equal to their number of shared artifacts. For example, the bipartite projection of Darwin's species location network is a species co-occurrence network in which a pair of species is connected by an edge with a weight equal to the number of islands where they are both found. We use $\mathbf{P}$ to denote the matrix representation of a bipartite projection, which is computed as $\mathbf{B}\mathbf{B}^T$, where $\mathbf{B}^T$ indicates the transpose of $\mathbf{B}$. In a projection $\mathbf{P}$, P_{ij} indicates the number of times both i and j were connected to the same artifact k in $\mathbf{B}$. The diagonal entries of $\mathbf{P}$, P_{ii}, are equal to the agent degrees. Typically the backbone of $\mathbf{P}$ will discard these diagonal entries, though their values are used in deciding which other edges are deemed important.

As the reader may have inferred, bipartite networks and their weighted projections are equivalent to bipartite and weighted graphs. This equivalence helps in the visualization and analysis techniques in the network sciences. A *graph* G is a set of objects called *vertices*, together with a set of 2-element subsets of the vertices which are called *edges*. An edge between vertices i and j can be

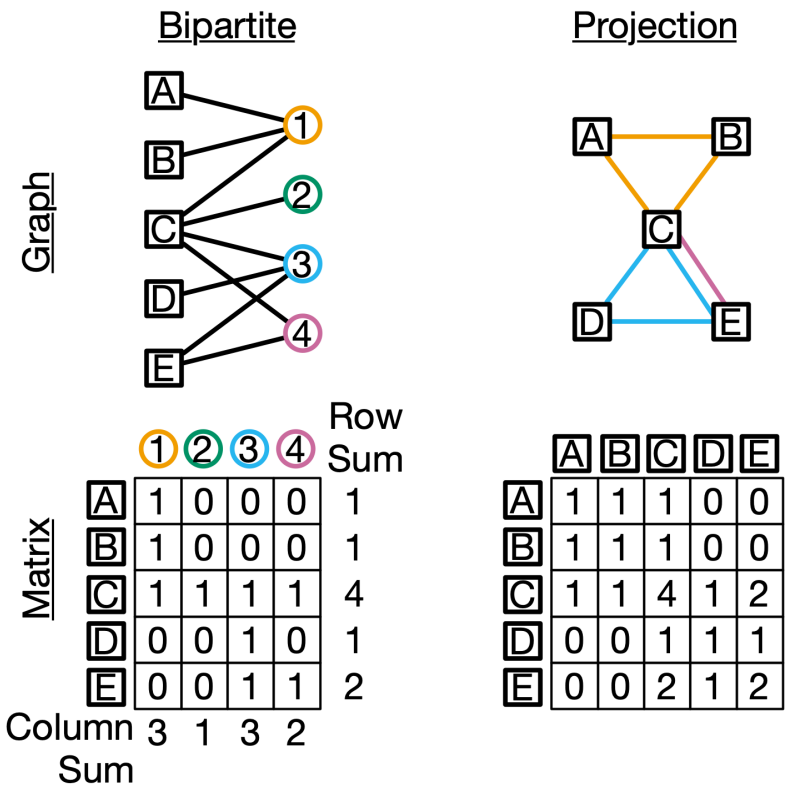

Figure 5.1: Bipartite and bipartite projection networks

denoted as $e = ij$. If there exists an edge $e = ij$ between vertices i and j, we say that i and j are *adjacent*. We call a graph *weighted* if each edge has an associated numeric value, and unweighted otherwise. The weight of edge $e = ij$ is denoted $w(ij)$; in unweighted graphs, we set $w(ij) = 1$ for all present edges. The *degree* of vertex i is the number of edges of the form ij for some j. Graphs are often discussed by viewing their *adjacency matrices* $\mathbf{G}$, where $G_{ij} = w(ij)$. As mentioned above, the matrix representation of a bipartite network $\mathbf{B}$ is the graph's *bipartite adjacency matrix*, while the matrix $\mathbf{P}$ is the adjacency matrix of the weighted graph. See fig. 5.1 for an example of this connection.

The *backbone* of a bipartite projection is a binary representation of $\mathbf{P}$ that contains only the most 'important' or 'significant' edges. For example, the backbone of a species co-occurrence network connects pairs of species if they are found on a significant number of the same islands, which might be interpreted as evidence that the two species do not compete for resources and perhaps are

symbiotic. We use $\mathbf{P}'$ to denote the matrix representation of the backbone of $\mathbf{P}$. Because multiple methods exist for deciding when an edge is significant and thus should occur in the backbone, we use $\mathbf{P}'^{\mathrm{M}}$ to denote a backbone extracted using method M.

Backbone extraction methods that were originally developed for non-projection weighted networks are often also applied to weighted bipartite projections. One simple method preserves an edge in the backbone if its weight in the projection exceeds some *universal threshold* T. However, when $T = 0$ is chosen (which is common), since each artifact of degree d induces $d(d-1)/2$ edges in the backbone, this leads to a very dense backbone with a high clustering coefficient [LMDV08]. Here, density refers to the number of edges present in the network divided by the maximum possible number of edges. A network clustering coefficient measures how many 'triangles', three pairwise adjacent vertices, are present in the network compared to all triples. Backbones with high density and clustering coefficient may not elucidate any interesting information regarding the network. Using $T > 0$ can yield a sparser and less clustered backbone [DT05, Fon20, BR11], but the choice of a particular threshold value is arbitrary, and applying the same threshold to all edges yields backbones that overlook agents with low degree in the projection [SBV09]. More sophisticated methods, including the *disparity filter* [SBV09] and *likelihood filter* [Dia16], aim to overcome these limitations of the universal threshold method by using a different threshold for each edge based on a null model. However, all methods that can be applied to non-projection weighted networks have the same shortcoming when applied to weighted bipartite projections: they ignore information about the artifacts [Nea14]. In the ecological case, the universal threshold, disparity filter, and likelihood filter methods all decide whether two species should be connected in the backbone only by examining how many islands they are both found on, but do not consider the characteristics of those islands, including how many other species are found there, or even how many islands there are. Therefore, although these methods are promising for extracting the backbone from non-projection weighted networks, different methods are required for extracting the backbone from a bipartite projection.

CHAPTER 6

BACKBONE MODELS AND THEIR PROBABILITY MASS FUNCTIONS

This chapter contains material from Neal, Domagalski, and Sagan [NDS21b]. All results in this chapter are from this manuscript unless otherwise noted.

6.1 Bipartite ensemble backbone models

Bipartite ensemble backbone models decide whether an edge's observed weight P_{ij} is significantly large, and thus whether a corresponding edge should be included in the backbone, in the following way. Let $\mathcal{B}$ be the set of all bipartite networks $\mathbf{B}^*$ having the same number of agents and artifacts as $\mathbf{B}$. In the ecological case, $\mathbf{B}^*$ might be viewed as representing a possible world containing the same species and islands, but in which locations of species on islands is different, and likewise $\mathcal{B}$ is the set of all such possible worlds. We will create our ensembles by taking a subset $\mathcal{B}^{\mathrm{M}}$ of $\mathcal{B}$ subject to certain constraints M and imposing a probability distribution on it. In all our models except the SDSM, we impose the uniform probability distribution on $\mathcal{B}^{\mathrm{M}}$, that is, each element of the ensemble is equally likely. We will then extract the backbone from the projection of $\mathbf{B}$ by using the distribution of edge weights arising from projections of members of the ensemble under consideration.

We use P^*_{ij} to denote a random variable equal to $(\mathbf{B}^*\mathbf{B}^{*T})_{ij}$ for $\mathbf{B}^* \in \mathcal{B}^{\mathrm{M}}$. That is, P^*_{ij} is the number of artifacts shared by i and j in a bipartite network randomly drawn from $\mathcal{B}^{\mathrm{M}}$. In the ecological case, P^*_{ij} represents the number of islands that are home to both species i and j in a possible world, while the distribution of P^*_{ij} is the distribution of the number of islands shared by species i and j in all possible worlds.

Decisions about which edges should appear in a backbone extracted at the two-tailed statistical

significance level α are made by comparing P_{ij} to P^*_{ij}

$$P'_{ij} = \begin{cases} 1 & \text{if } \Pr(P^*_{ij} \geq P_{ij}) < \frac{\alpha}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

This test preserves an edge in the backbone if its weight in the observed projection is uncommonly large compared to its weight in projections of members of the ensemble. A two-tailed significance test is used because, in principle, an edge's weight in the observed projection could be uncommonly *larger* or uncommonly *smaller* than its weight in projections of members of the ensemble. One can use the same principles to obtain a *signed* backbone by comparing P_{ij} to P^*_{ij} with

$$P'_{ij} = \begin{cases} 1 & \text{if } \Pr(P^*_{ij} \geq P_{ij}) < \frac{\alpha}{2}, \\ -1 & \text{if } \Pr(P^*_{ij} \leq P_{ij}) < \frac{\alpha}{2}, \\ 0 & \text{otherwise.} \end{cases}$$

In the ecological case, two species are connected in the backbone if their number of shared islands in the observed world is uncommonly large compared to their number of shared islands in all possible worlds.

There are many ways that $\mathcal{B}$ can be constrained [SUG18], with each set of constraints describing a different ensemble $\mathcal{B}^{\mathrm{M}}$ and different ensemble backbone model; however, in this work we focus on five possibilities. We describe each of these models and their meaning in the context of Darwin's species and islands, and derive their probability mass functions for the respective edge weight distributions. These probability mass functions of P^*_{ij} are used by ensemble backbone models to evaluate the statistical significance of the weight of edge P_{ij} in a bipartite projection. We use the following notation:

- Let $\mathbf{B}$ be an $m \times n$ bipartite matrix, with a vector of row sums $R = (r_1, \ldots, r_m)$, a vector of column sums $C = (c_1, \ldots, c_n)$, and f cells containing a 1. So

$$f = \sum_{i=1}^{m} r_i = \sum_{j=1}^{n} c_j.$$

- Let $\mathcal{B}^{\mathrm{M}}$ be the ensemble of all $m \times n$ matrices $\mathbf{B}^* = (B^*_{ij})$ that obey the constraints of the respective model. In all models, the probability distribution on $\mathcal{B}^{\mathrm{M}}$ is uniform except in the stochastic case.
- Let P^*_{ij} be a random variable equal to $(\mathbf{B}^*\mathbf{B}^{*T})_{ij}$ for all $\mathbf{B}^* \in \mathcal{B}^{\mathrm{M}}$. Note that we have

$$P^*_{ij} = B^*_{i1}B^*_{j1} + B^*_{i2}B^*_{j2} + \cdots + B^*_{in}B^*_{jn}. \tag{6.1}$$

6.2 Fixed degree sequence model (FDSM)

In the *fixed degree sequence model* (FDSM) $\mathbf{B}^* \in \mathcal{B}^{\mathrm{FDSM}}$ are constrained to have the same agent and artifact degree sequences as $\mathbf{B}$. Adopting the FDSM implies, for example, that in all possible worlds a given species is found on exactly the same number of islands, and a given island is home to exactly the same number of species. The distribution of P^*_{ij} arising from $\mathcal{B}^{\mathrm{FDSM}}$ is unknown, but can be approximated by uniformly sampling $\mathbf{B}^*$ from $\mathcal{B}^{\mathrm{FDSM}}$, constructing $\mathbf{P}^*$, and saving the values P^*_{ij}. In the studies below, we use 1000 samples of $\mathbf{B}^*$ generated using the 'curveball' algorithm, which is among the fastest methods to sample $\mathcal{B}^{\mathrm{FDSM}}$ uniformly at random [SNB+14, Car15]. The FDSM has been used to extract the backbone of bipartite projections of, for example, movies co-liked by viewers [ZK11] and conference panel co-participation by scholars [SR12, DL16]. In this paper, we use the FDSM as the reference model to which other ensemble models are compared because it fully controls for both degree sequences.

The primary limitation of the FDSM is its computational cost. First, constructing each $\mathbf{P}^*$ requires matrix multiplication, which must be performed repeatedly and has complexity $O(n^{2.37})$ for two $n \times n$ matrices using the fast Coppersmith-Winograd algorithm [CW90]. Second, computing $\Pr(P^*_{ij} \geq P_{ij})$ with sufficient precision to achieve a two-tailed familywise error rate of α requires at least $\frac{.5m^2 - .5m}{\alpha/2} + 1$ samples, where m is the number of rows (i.e., agents) in $\mathbf{B}$ and $\mathbf{P}$. Thus, for example, extracting the backbone of a bipartite projection with 1000 agents at a family-wise error rate of 0.05 would require performing at least 20 million matrix multiplications. Therefore, the tightly-constrained FDSM is frequently impractical for backbone extraction. However, models that rely on ensembles with more relaxed constraints offer computationally-feasible alternatives.

6.3 Fixed fill model (FFM)

In the highly relaxed *fixed fill model* (FFM), $\mathbf{B}^* \in \mathcal{B}^{\text{FFM}}$ are simply constrained to contain the same number of 1s as $\mathbf{B}$. Adopting the FFM implies, for example, that in all possible worlds only the total number of species-habitat pairs is fixed, but any given species may be found on a different number of islands and any given island may be home to a different number of species. The distribution of P^*_{ij} arising from B^{FFM} has not been described before. We derive it and call it a *Jacobi distribution* because it is related to Jacobi polynomials.

Let the *fixed fill model* constrain all $\mathbf{B}^* \in \mathcal{B}^{\text{FFM}}$ to contain the same number of 1s (i.e. fill) as $\mathbf{B}$.

Theorem 6.3.1. *Under the fixed fill model, the distribution of P^*_{ij} for $i \neq j$ satisfies*

$$\Pr(P^*_{ij} = k) = \frac{\binom{n}{k} \sum_r 2^{n-k-r} \binom{n-k}{r} \binom{(m-2)n}{f-n-k+r}}{\binom{mn}{f}}. \tag{6.2}$$

Proof. For the denominator we need to compute the cardinality $\#\mathcal{B}^{\text{FFM}}$. If $\mathbf{B}^* \in \mathcal{B}^{\text{FFM}}$ then $\mathbf{B}^*$ has mn entries of which f must be chosen to be ones. So

$$\#\mathcal{B}^{\text{FFM}} = \binom{mn}{f}.$$

For the numerator, suppose $P^*_{ij} = k$. We see from equation (6.1) that there are exactly k columns c where $B^*_{ic} = B^*_{jc} = 1$. There are $\binom{n}{k}$ ways to choose these columns. Now define the following parameters:

$$\begin{aligned} p &= \text{number of columns } c \text{ where } B^*_{ic} = 1 \text{ and } B^*_{jc} = 0, \\ q &= \text{number of columns } c \text{ where } B^*_{ic} = 0 \text{ and } B^*_{jc} = 1, \\ r &= \text{number of columns } c \text{ where } B^*_{ic} = 0 \text{ and } B^*_{jc} = 0. \end{aligned}$$

The number of ways to pick the columns counted by these parameters from the $n - k$ columns which do not contains ones in both rows is the trinomial coefficients $\binom{n-k}{p,q,r}$. Now we have used

$2k+p+q$ ones in rows i and j. So there are $f-2k-p-q$ left to distribute to the remaining $m-2$ rows. And these rows have $(m-2)n$ entries. So the number of possibilities for these remaining ones is $\binom{(m-2)n}{f-2k-p-q}$. Thus the total number of choices from this and the previous paragraph is

$$\begin{aligned}\binom{n}{k}\sum_{p+q+r=n-k}\binom{n-k}{p,q,r}\binom{(m-2)n}{f-2k-p-q} &= \binom{n}{k}\sum_{p+q+r=n-k}\binom{n-k}{r}\binom{n-k-r}{p}\binom{(m-2)n}{f-n-k+r}\\ &= \binom{n}{k}\sum_{r}\binom{n-k}{r}\binom{(m-2)n}{f-n-k+r}\sum_{p}\binom{n-k-r}{p}\\ &= \binom{n}{k}\sum_{r}2^{n-k-r}\binom{n-k}{r}\binom{(m-2)n}{f-n-k+r}\end{aligned}$$

as desired. □

For even modestly large **B**, computing equation (6.2) involves values larger than can be handled by some programs. In practice, we use logs to make these computations practical.

We now show that the sum in the numerator of this probability is related to the famous Jacobi orthogonal polynomials. This sum is a terminating hypergeometric series. Given a real number a and a nonnegative integer r the corresponding *Pochhammer symbol* or *rising factorial* is

$$(a)_r = a(a+1)(a+2)\cdots(a+r-1).$$

Note that if a is an integer with $-r < a \leq 0$ then $(a)_r = 0$ because the product contains 0 as a factor. Given real numbers $a_1, a_2, \ldots, a_p$ and $b_1, b_2, \ldots, b_q$ as well as a variable z, the corresponding *hypergeometric series* is

$${}_pF_q\left[\begin{matrix}a_1 & a_2 & \ldots & a_p\\ b_1 & b_2 & \ldots & b_q\end{matrix};z\right] = \sum_{r\geq 0}\frac{(a_1)_r(a_2)_r\cdots(a_p)_r}{(b_1)_r(b_2)_r\cdots(b_q)_r}\frac{z^r}{r!}.$$

Note that if any of the a_i are negative integers then, because of the remark above, this series will terminate and become a polynomial in z.

To convert a binomial coefficient into Pochhammer symbols, we write

$$\begin{aligned}\binom{n}{r} &= \frac{(n)(n-1)\cdots(n-r+1)}{r!} \\ &= \frac{(-1)^r(-n)(-n+1)\cdots(-n+r-1)}{(1)_r} \\ &= \frac{(-1)^r(-n)_r}{(1)_r}.\end{aligned}$$

The following identity will also be useful

$$\begin{aligned}(a)_{b+r} &= (a)(a+1)\cdots(a+b-1)\times(a+b)(a+b+1)\cdots(a+b+r-1) \\ &= (a)_b(a+b)_r.\end{aligned}$$

We now return to the sum in the numerator of equation (6.2). We will ignore the factor of 2^{n-k} since it is constant with respect to the sum and so can be pulled outside. For simplicity of calculation we will also use the substitutions

$$s = (m-2)n, \qquad t = f-n-k.$$

Thus we have

$$\begin{aligned}\sum_r 2^{-r}\binom{n-k}{r}\binom{(m-2)n}{f-n-k+r} &= \sum_r \binom{n-k}{r}\binom{s}{t+r}(1/2)^r \\ &= \sum_r \frac{(-1)^r(k-n)_r}{(1)_r}\cdot\frac{(-1)^{t+r}(-s)_{t+r}}{(1)_{t+r}}(1/2)^r \\ &= (-1)^t\sum_r \frac{(k-n)_r(-s)_t(-s+t)_r}{(1)_t(t+1)_r}\frac{(1/2)^r}{(1)_r} \\ &= \frac{(-1)^t(-s)_t}{(1)_t}\sum_r \frac{(k-n)_r(-s+t)_r}{(t+1)_r}\frac{(1/2)^r}{r!} \\ &= \binom{s}{t}\,{}_2F_1\left[\begin{matrix}k-n & -s+t\\ & t+1\end{matrix};\frac{1}{2}\right]\end{aligned}$$

We are indebted to Marko Petkovšek [personal communication] for pointing out that this ${}_2F_1$ is, up to a factor, a specialization of a Jacobi polynomial. Given a nonnegative integer ℓ and real

numbers α, β the associated *Jacobi polynomial* is

$$P_\ell^{(\alpha,\beta)}(z) = \binom{\alpha+\ell}{\ell} {}_2F_1\left[\begin{matrix} -\ell & \ell+\alpha+\beta+1 \\ & \alpha+1 \end{matrix}; \frac{1-z}{2}\right]$$

To make these ${}_2F_1$ polynomials agree we can let $\ell = n-k$, $\alpha = t = f-n-k$,

$$\beta = -s + t - (\ell + \alpha + 1) = k - (m-1)n - 1$$

and $z = 0$. With these substitutions we get

$$\sum_r 2^{-r} \binom{n-k}{r} \binom{(m-2)n}{f-n-k+r} = \frac{\binom{(m-2)n}{f-n-k}}{\binom{f-2k}{n-k}} P_{n-k}^{(f-n-k,\ k-(m-1)n-1)}(0).$$

6.4 Fixed row model (FRM)

In the more constrained *fixed row model* (FRM), $\mathbf{B}^* \in \mathcal{B}^{\mathrm{FRM}}$ are constrained to have the same agent degree sequence as $\mathbf{B}$, but have unconstrained artifact degree sequences. Adopting the FRM for backbone extraction implies, for example, that in all possible worlds a given species is found on the same number of islands, but a given island may be home to a different number of species. The distribution of P^*_{ij} arising from $\mathcal{B}^{\mathrm{FRM}}$ is hypergeometric [TML+11, Nea13]. The FRM has been used to extract the backbone of bipartite projections of, for example, movies co-starring actors [TML+11], papers co-written by authors [TML+11], parties co-attended by women [Nea13], majority opinions joined by Supreme Court justices [Nea13], and microRNAs co-associated with diseases [CXW+18].

Let the *fixed row model* constrain all $\mathbf{B}^* \in \mathcal{B}^{\mathrm{FRM}}$ to have the same row sums as $\mathbf{B}$.

Theorem 6.4.1. *Under the fixed row model, the distribution of P^*_{ij} for $i \neq j$ is hypergeometric and satisfies*

$$\Pr(P^*_{ij} = k) = \frac{\binom{r_j}{k}\binom{n-r_j}{r_i-k}}{\binom{n}{r_i}}.$$

Proof. The total number of ways to pick r_i of the n columns for ones in the ith row and r_j of the n columns for ones in the jth row is

$$\binom{n}{r_i}\binom{n}{r_j} = \binom{n}{r_i}\frac{n!}{r_j!(n-r_j)!}. \tag{6.3}$$

So that will go in the denominator of the desired probability.

For the numerator we follow the same line of reasoning as in the previous proof, where the parameters therein can be expressed as

$$\begin{aligned} p &= r_i - k, \\ q &= r_j - k, \\ r &= n - r_i - r_j + k. \end{aligned}$$

So we have a total of

$$\binom{n}{k}\binom{n-k}{p,q,r} = \frac{n!}{k!(r_i-k)!(r_j-k)!(n-r_i-r_j+k)!} \tag{6.4}$$

choices.

Dividing equation (6.4) by (6.3) and cancelling $n!$ gives

$$\Pr(P^*_{ij} = k) = \frac{\frac{r_j!}{k!(r_j-k)!}\cdot\frac{(n-r_j)!}{(r_i-k)!(n-r_i-r_j+k)!}}{\binom{n}{r_i}} = \frac{\binom{r_j}{k}\binom{n-r_j}{r_i-k}}{\binom{n}{r_i}}.$$

as desired. □

6.5 Fixed column model (FCM)

In the closely related *fixed column model* (FCM), $\mathbf{B}^* \in \mathcal{B}^{\mathrm{FCM}}$ are constrained to have the same artifact degree sequence as $\mathbf{B}$, but have unconstrained agent degree sequences. Adopting the FCM for backbone extraction implies, for example, that in all possible worlds a given species may be found on a different number of islands, but a given island is home to the same number of species.

The distribution of P^*_{ij} arising from $\mathcal{B}^{\mathrm{FCM}}$ has not been described before, but we derive it here to show it is Poisson-binomial.

Let the *fixed column model* constrain all $\mathbf{B}^* \in \mathcal{B}^{\mathrm{FCM}}$ to have the same column sums as $\mathbf{B}$.

Let $X_1, \ldots, X_n$ be independent Bernoulli random variables. Let the probability of success for X_i be

$$\Pr(X_i = 1) = p_i.$$

The random variable

$$X = X_1 + \cdots + X_n \tag{6.5}$$

is said to have the *Poisson binomial distribution* with parameters $p_1, \ldots, p_n$.

Theorem 6.5.1. *Under the fixed column model, the distribution of P^*_{ij} for $i \neq j$ is Poisson binomial with parameters*

$$p_1 = \frac{c_1(c_1-1)}{m(m-1)},\ p_2 = \frac{c_2(c_2-1)}{m(m-1)},\ \ldots,\ p_n = \frac{c_n(c_n-1)}{m(m-1)}.$$

Proof. The B^*_{ik} are all either zero or one and are independent in different columns when only the column sums are fixed. So as k varies, the products $B^*_{ik}B^*_{jk}$ are independent Bernoulli random variables. Comparing equations (6.1) and (6.5), we see that the distribution of P^*_{ij} is Poisson binomial.

If column k has column sum $c = c_k$ then all zero-one vectors with sum c are equally likely for that column of $\mathbf{B}^*$. So there are $\binom{m}{c}$ possible kth columns. The number of ways to have a success is the number of possible columns which have ones in both positions i and j where $i \neq j$. So the number of choices is the number of ways to choose the remaining $c-2$ ones in that column from the other $m-2$ positions, that is, $\binom{m-2}{c-2}$. Thus

$$p_k = \Pr(B^*_{ik}B^*_{jk} = 1) = \frac{\binom{m-2}{c-2}}{\binom{m}{c}} = \frac{c(c-1)}{m(m-1)}$$

which finishes the demonstration. □

6.6 Stochastic degree sequence model (SDSM)

Finally, the *stochastic degree sequence model* (SDSM) takes $\mathcal{B}^{\text{SDSM}}$ to be all binary $m \times n$ matrices, but also gives a process for generating these matrices with different probabilities. Each $\mathbf{B}^*$ is generated by filling the cells B^*_{ik} with a 0 or 1 depending on the outcome of an independent Bernoulli trial with probability p^*_{ik}. The distribution of the random variable P^*_{ij} arising from $\mathcal{B}^{\text{SDSM}}$ is Poisson-binomial with parameters which can be computed using the p^*_{ik} [DNS21, LR16]. There are many ways to choose p^*_{ik}, but in the studies in chapter 8, we choose p^*_{ik} so that it approximates $\Pr(B^*_{ik} = 1)$ for $\mathbf{B}^* \in \mathcal{B}^{\text{FDSM}}$, with the goal of ensuring that the *expected* agent and artifact degree sequences of $\mathbf{B}^* \in \mathcal{B}^{\text{SDSM}}$ match those of $\mathbf{B}$. Adopting such a version of SDSM implies, for example, that in each possible world a given species may be found on many or few islands and a given island may be home to many or few species, but the *average* number of islands on which a given species lives in all possible worlds and the *average* number of species that live on an given island in all possible worlds matches these values the observed world. The SDSM has been used to extract the backbone of bipartite projections of, for example, legislators co-sponsoring bills [Nea20, Nea14, SB20], zebrafish (*Danio rerio*) sharing operational taxonomic units [BDS+20], countries sharing exports [SDCGS15], and genes expressed in genesets [MLLS21].

In the *stochastic degree sequence model*, $\mathcal{B}^{\text{SDSM}}$ consists of all binary $m \times n$ matrices. A method is then chosen to generate probabilities p^*_{ik}. Finally, matrices $\mathbf{B}^* \in \mathcal{B}^{\text{SDSM}}$ are generated using these probabilities for independent Bernoulli trials, where B^*_{ik} is filled with a one with probability p^*_{ik} and zero otherwise.

Theorem 6.6.1. *Under the stochastic degree sequence model, the distribution of P^*_{ij} for $i \neq j$ is Poisson binomial with parameters*

$$p_1 = p^*_{i1}p^*_{j1}, \ \ldots, \ p_n = p^*_{in}p^*_{jn}.$$

Proof. The fact that the distribution is Poisson binomial follows immediately from the independence assumption on the $\Pr(B^*_{ik})$ and equation (6.1). Furthermore, the probability that the kth variable is

one is

$$p_k = \Pr(B^*_{ik} B^*_{jk} = 1) = \Pr(B^*_{ik} = 1) \Pr(B^*_{jk} = 1) = p^*_{ik} p^*_{jk}.$$

So we are done. □

In the following chapter, we will implement these emsemble methods in the R package backbone.

CHAPTER 7

BACKBONE: AN R PACKAGE FOR EXTRACTING THE BACKBONE OF WEIGHTED GRAPHS

This chapter contains material from Domagalski, Neal, and Sagan [DNS21, NDS21a], and background from Neal, Domagalski, and Yan [NDY22]. Replication materials are available at https://www.github.com/domagal9/dissertation.

We now introduce the R package `backbone` that implements these five models, fixed degree sequence model (FDSM), fixed fill model (FFM), fixed column model (FCM), fixed row model (FRM), and the stochastic degree sequence model (SDSM). The `backbone` package provides these methods in a common framework making them both accessible and easy to use for scientists and researchers. It can be installed in R [R C18] from The Comprehensive R Archive Network (CRAN) via `install.packages("backbone")` and used with `library(backbone)` [DNS20]. Information regarding the CRAN distribution is found at https://CRAN.R-project.org/package=backbone. Additional materials relating to `backbone` including papers, presentations, workshop materials, and datasets are available at https://rbackbone.net.

7.1 Two Illuminating Data Sets

We illustrate the use of the R `backbone` package to extract the backbone of two networks: the first is a network of bill co-sponsorship relations among Senators in the 114^{th} session of the United States Senate, the second is a network of world city firm co-locations amongst large cities in the year 2000. Both of these networks, legislative and spatial, are used as templates for network research in their corresponding fields.

7.1.1 Legislative Networks

For more than a decade, legislative networks have shed new light on understanding legislative behavior [Fow06a, Fow06b]. Although legislative networks clarify that governance is an interactive

and interdependent process, they are most useful if they help us explain or predict key parts of this process. The most consequential action a legislator can take is voting, and several studies have shown that a legislator's position in a legislative network helps explain their voting behavior. For example, [Fow06a] found that US legislators were more likely to vote in favor of bills sponsored by well-connected legislators, even after controlling for shared party membership, and therefore that well-connected legislators were more effective at advancing their legislative agendas. Similarly, [RNH13] found that social ties among European legislators exacerbated ideological voting patterns: friendship increased the likelihood of political allies voting the same way, but decreased the likelihood of political adversaries voting the same way. [Fon20] offers one potential explanation for the network's influence over voting behavior: "When legislators are called on to vote on a question that they do not understand, they take cues from experts who are nearby in the legislative network" (p. 270). Although voting is particularly consequential, legislative networks have also been used to explain how the coalitions that shape voting outcomes change over time. For example, [Nea20] demonstrated that the US Congress has become substantially more partisan since 1973 with legislators increasingly collaborating only with members of the same party, and opposing members of the other party. However, [KMN16] and [AN20a] clarified that these coalitions are not strictly partisan and frequently include members from both parties.

Directly measuring legislative networks (e.g., simply asking legislators who they work with) is challenging because legislators are busy and may have motivations to conceal or misrepresent their true collaborations. As a result, most studies of legislative networks rely on more indirect measurements derived from bill sponsorship [e.g., Nea20], committee memberships [e.g., PMNW05], attendance at press events [e.g., DMSK15], and roll call votes [e.g., ALH$^+$15a]. What do such indirectly measured legislative networks measure? Different source data provides information about different types of relations among legislators. For example, voting similarly in roll call votes provides information about ideological alignment, whereas sharing membership on a committee provides information about alignment on prioritized issues. The majority of legislative networks are derived from patterns of bill sponsorship, which also provides information about ideological

and issue alignment, but more directly provides information about collaboration as legislators join together in lending their collective support to bills [Kir11, KK96].

All but the most popular legislative measures require collaboration to cultivate support and ensure their eventual passage. Past studies have identified many factors that influence when legislators choose to collaborate, consistently finding support for homophily [MSLC01]: similar legislators are more likely to collaborate [CP87]. In the context of legislative collaboration, homophily with respect to political party is known as partisanship, which when particularly intense leads to partisan polarization. Both research [e.g., Nea20, LCH06, MM13] and media reports [e.g., Ing15] confirm that polarization has become a hallmark of legislative relations in the US Congress, so observing party homophily in networks of legislative collaboration is expected.

To demonstrate how the `backbone` package works, we employ its use on a co-sponsorship network of the United States Senate during the 114$^{\text{th}}$ session. Since both prior research [LCH06, Nea20, SB20, ALH$^+$15b, AN20b] and media accounts [Dru16] of the current US political climate provide us with *a priori* expectations about what structure a properly extracted backbone should have, we expect positive relationships to form primarily between those in the same political party, and accordingly a relatively large modularity statistic computed from a partition of the nodes by political party. Modularity measures the strength of division within the network. Specifically, for a network $\mathbf{G}$ with vertex degree sequence $(d_1, \ldots, d_n)$, it is given by the quantity

$$Q = \frac{1}{2\sum \mathbf{G}} \sum_{i,j} \frac{G_{ij} - d_i d_j}{2\sum \mathbf{G}} \delta(c_i, c_j),$$

where c_i and c_j represent the communities (in this case political party) that vertices i and j belong to, and $\delta(c_i, c_j)$ is the Kronecker delta function. In visualizations of the extracted backbones, we depict Republican senators by red vertices, and both Democratic and Independent senators who are left-leaning and caucused with Democrats by blue vertices. Although we discuss signed backbones in the text, for visual clarity we only provide figures for binary backbones which contain positive edges. Positive relations of collaboration between two Republicans are depicted in red, between two Democrats are blue, and for all other pairs are purple. For an example, see fig. 7.1

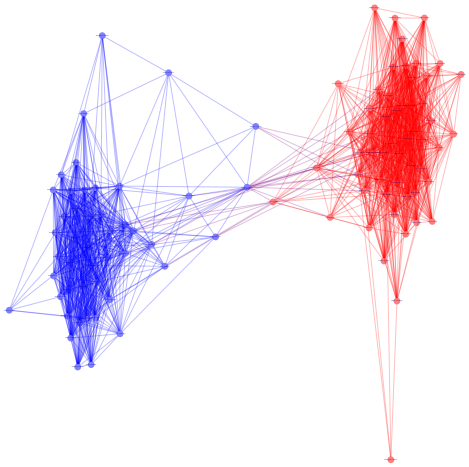

Figure 7.1: An example of an extracted backbone, with Democratic senators represented by blue vertices, and Republican senators represented by red vertices.

The data set consists of 100 senators and the 3589 bills that they have sponsored or co-sponsored in the 114^{th} session of Congress [USG20]. This data takes the form of a bipartite network $\mathbf{B}$, where the agents are the senators (rows) and the artifacts are the bills (columns). Here, $B_{ik} = 1$ if senator i sponsored or co-sponsored bill k, and otherwise is 0. Below we examine the data set. Notice that the row names correspond to each senator (including their party affiliation and the state they represent) and the column names refer to the bill number.

```
> set.seed(19)
> library(backbone)
> senate <- read.csv("S114.csv", row.names = 1, header = TRUE)
> senate <- as.matrix(senate)
> dim(senate)
[1]  100 3589


> senate[1:5, 1:5]
                     sj9 sj8 sj7 sj6 sj5
Alexander, L. (TN-R)   0   1   0   1   0
```

```
Boxer, B. (CA-D)       0   0   0   0   1
Cantwell, M. (WA-D)    0   0   0   0   1
Carper, T. (DE-D)      0   0   0   0   1
Cochran, T. (MS-R)     0   1   0   1   0
```

A weighted network $\mathbf{P}$ can be constructed from $\mathbf{B}$ via bipartite projection, where $\mathbf{P} = \mathbf{B}\mathbf{B}^T$ and P_{ij} contains the number of bills that both senator i and senator j sponsored. Notice the network is now 100 rows by 100 columns.

```
> G <- senate%*%t(senate)
> dim(G)
[1] 100 100


> G[1:5, 1:2]
                     Alexander, L. (TN-R) Boxer, B. (CA-D)
Alexander, L. (TN-R)                  141               10
Boxer, B. (CA-D)                       10              303
Cantwell, M. (WA-D)                    15               82
Carper, T. (DE-D)                      12               55
Cochran, T. (MS-R)                     40               25
```

The projected network $\mathbf{P}$ now indicates that Senator Lamar Alexander sponsored a total of 141 bills in the 114^{th} session. Among these 141 bills, 10 were co-sponsored with Senator Barbara Boxer, and 15 were co-sponsored with Senator Maria Cantwell.

We can use the values of graph $\mathbf{P}$ to observe differences between those with similar or dissimilar ideology. Below, we compare the number of bills co-sponsored by two individuals with similar political ideology, Senators Cory Booker and Elizabeth Warren, versus those with dissimilar ideology, Senators Ted Cruz and Bernie Sanders. The results are consistent with the expectation that legislators sharing a similar ideology engage in more co-sponsorships.

```
> G["Booker, C. (NJ-D)", "Warren, E. (MA-D)"]
[1] 98

> G["Cruz, T. (TX-R)", "Sanders, B. (VT-I)"]
[1] 5
```

The differences in the number of bills co-sponsored prompts an important underlying question: how many bills do two senators have to co-sponsor before we would be justified in concluding they are political collaborators? Similarly, how few bills do they have to co-sponsor before we would be justified in concluding they are political opponents? These questions are what the `backbone` package seeks to answer.

7.1.2 Spatial Networks

The second type of network we will examine with the `backbone` package is a spatial network. Bipartite projections appear in spatial analysis, where they can take two distinct forms depending on whether the agents or artifacts are spatial entities (i.e., locations). In the *locations-as-agents* approach, a spatial bipartite projection is a network of locations, such that a pair of locations is connected to the extent that they share artifacts. Calling it the "interlocking world city network model," this is the approach that [Tay01] proposed and which launched a wave of research on world city networks: major cities (the agents, which are locations) are connected to the extent that they house branch offices of the same advanced producer services firms (e.g., finance, accounting, consulting; the artifacts). It rests on the logic that offices of the same firm must communicate and interact with one another, and therefore that when two cities have an office of the same firm, there is likely interaction between them. Spatial networks adopting the locations-as-agents approach to measurement via bipartite projection are quite common at multiple spatial scales, and have been used to measure networks among urban locations connected by twitter users [Poo18], bus routes [LD20], networks among cities connected by patents [BR17], banking syndicates [PWK19], networks among countries connected by treaties [HBKM09], trade [SCS17], and corporate executives [HFC16].

In the *locations-as-artifacts* approach, a spatial bipartite projection is a network of agents (often people or other social actors), such that a pair of agents is connected to the extent that they share locations. The locations-as-artifacts approach is less common in geography because the spatial units play only an instrumental role in the network, forging the links between agents, but do not appear in the bipartite projection network itself. However, it is common in sociological research, where the focus is on social networks emerging from spatial interactions. For example, [BCS+17] and [XCB20] use this approach to measure and study the social network among households in Los Angeles: households (the agents) are connected to the extent that they visit the same routine activity locations (e.g., school, work; the artifacts). This rests on the logic that places offer opportunities for casual encounters which lead to the formation of social bonds, and therefore when two households frequent the same places, they are more likely to interact with each other [Jac61]. [HKBH07] adopted a similar locations-as-artifacts approach to derive a 'product space' in which export products were connected to the extent that they were exported by the same countries. This follows the logic that "if [the production of] two goods...require similar institutions, infrastructure, physical factors, technology, or some combination thereof, they will tend to be produced [in the same location]," and therefore the spatial co-production of products indirectly captures their production technology similarity [HKBH07, p. 484].

There is an important link between these two approaches. When $\mathbf{B}$ is a bipartite network where the rows represent locations, then $\mathbf{BB'}$ will yield a locations-as-agents bipartite projection, while $\mathbf{B'B}$ will yield a locations-as-artifacts bipartite projection. Therefore, a single bipartite network can be studied from both perspectives. For example, although the world cities literature usually focuses on cities linked by sharing firms, some have simultaneously examined a network of firms linked by their co-location in cities [e.g., Nea08, VMND16]. Similarly, [SCS17] examined not only a network of countries linked by trading the same products, but also a network of products that are traded by the same countries.

The key advantage to measuring spatial networks using bipartite projections lies in the relative ease of data collection. For example, data about economic exchanges between cities may not be

available from official government sources, and collecting such data directly is often impractical. However, data about where firms' offices are located is readily available, usually on the firms' own websites. Accordingly, bipartite projections offer a practical way for researchers to indirectly approximate a city-level economic network. Similarly, because social network analysis requires data from a population (not a sample) and is sensitive to missingness, it is often impractical to collect data on the social network among residents of a large city. However, data about the places residents visit or tweet about can be collected using routine surveys, remote sensing, and digital trace measures. Accordingly, bipartite projections also offer a practical way for researchers to indirectly approximate social networks in large geographic areas.

In the context of spatial analysis, it can be used for research adopting a locations-as-agents approach, to infer the spatial network among a set of locations from data on their shared characteristics. However, it can also be used for research adopting a locations-as-artifacts approach, to infer a social network among a set of actors from data on their shared locations. To illustrate `backbone`'s application in one specific spatial analytic context, we will demonstrate its use to examine the world city network and identify the most central cities in it.

The Globalization and World Cities (GaWC) "Data Set 11" was originally collected in 2000, and records the extent of 100 advanced producer services firms' presence in each of 315 large cities [TCW02]. These data served as the foundation for one of the earliest and most comprehensive empirical studies of the world city network [Tay04], and as a template for a substantial body of empirical research conducted by those associated with the GaWC research network. Formally, the data set takes the form of a rectangular 315×100 bipartite matrix $\mathbf{B}$, in which B_{ik} contains the 'service value' of firm k's presence in city i. The service values are an ordinal scale intended to capture the importance or extent of a firm's presence in a city, and ranged from 0 (no presence) to 5 (global headquarters), with a value of 2 representing an presence that provides "the 'normal' or 'typical' service level of the given firm in a city" [TCW02, p. 2370]. These publicly available data can be loaded into R directly from the GaWC website (as of July 2021) and converted to matrix form. This data set is also included in the replication materials.

```
> cities <- read.csv(file="https://www.lboro.ac.
                uk/gawc/datasets/da11.csv",
                header = TRUE,
                row.names = 1)
> cities <- as.matrix(cities)
```

The backbone package is designed for use with binary bipartite data, so for this illustration we transform the original ordinal **B** into a binary **B′** such that

$$B'_{ij} = \begin{cases} 1 & \text{if } B_{ij} \geq 3 \\ 0 & \text{if } B_{ij} \leq 2 \end{cases}.$$

This transformation can be achieved, and the cities that contain no firms with a larger-than-typical presence can be excluded, by typing:

```
> cities[cities <= 2] <- 0
> cities[cities >= 3] <- 1
> cities <- cities[rowSums(cities) != 0,]
```

This transformation allows us to focus only on firms that maintain a larger-than-typical presence in a given city, and only on the 196 cities that contain at least one such firm. For convenience, we use **B** to refer to this binary matrix in the remainder of this section. Once the bipartite data has been loaded and transformed, it is possible to examine some of its features. For example, it is possible to look at the pattern of firms' presence in cities.

```
> cities[114:117,8:11]
            Horwath KPMG Summit...Baker RSM
MELBOURNE         0    1              0   1
MEXICO CITY       0    1              0   0
MIAMI             1    1              0   1
MILAN             0    0              0   1
```

This command shows the portion of **B** that includes the 114^{th} to 117^{th} cities, and 8^{th} to 11^{th} firms. The output shows that while the accounting firms of KPMG and RSM maintained offices in several of these cities, Horwath and Summit International+Baker Tilley did not.

Two key characteristics of any bipartite data are the row sums and column sums. In these data, the row sums indicate the number of firms located in a city, while the column sums indicate the number of cities in which a firm maintains a presence.

```
> rowSums(cities)["AMSTERDAM"]
AMSTERDAM
       29
> rowSums(cities)["NEW YORK"]
NEW YORK
      74
> colSums(cities)["KPMG"]
KPMG
  76
> colSums(cities)["HSBC"]
HSBC
  43
```

For example, there are 74 firms that maintain a larger-than-typical presence in New York, but only 29 firms that maintain a larger-than-typical presence in Amsterdam. Likewise, KPMG maintains a larger-than-typical presence in 76 cities, while HSBC maintains a larger-than-typical presence in only 43 cities. Figure 7.2 illustrates these values for all cities and firms in these data. Specifically, Figure 7.2A shows that while most cities contain fewer than 20 firms, some cities contain many more firms. Similarly, Figure 7.2B shows that while most firms maintain a presence fewer than 40 cities, some firms maintain a presence of many more cities.

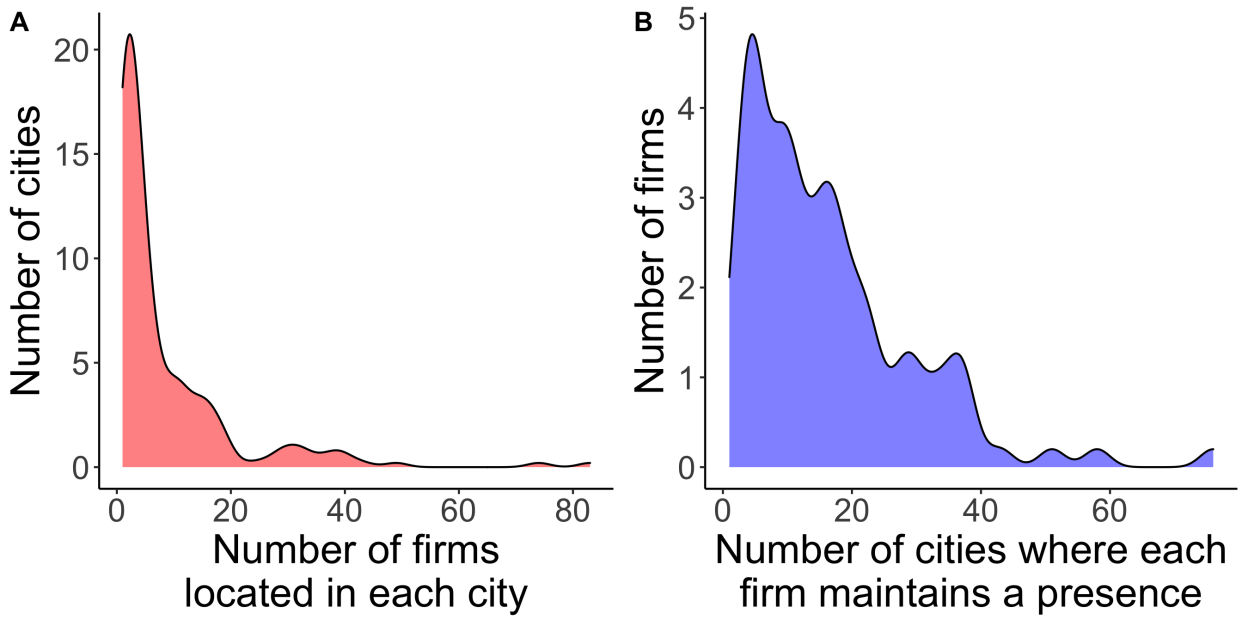

Figure 7.2: The distribution of (A) row sums and (B) column sums in the GaWC Dataset 11.

The conventional "specification of the world city network" used in GaWC research involves computing a weighted bipartite projection **P** from the original bipartite data **B** [Tay01].

```
> P <- cities %*% t(cities)
```

Following this specification, the cities are treated as agents and the firms are treated as artifacts. The resulting square matrix **P** is treated as a weighted world city network in which the strength of the connection between a pair of cities is measured by their number of co-located firms. For example, examining the matrix cell corresponding to the connection between Amsterdam and New York

```
> P["AMSTERDAM","NEW YORK"]
[1] 26
```

indicates that 26 firms maintain a presence in both cities, and might be interpreted as evidence that they interact economically.

Many analyses of the world city network focus on cities' degree centrality, or what is sometimes called a city's "global network centrality" (GNC). This value measures a city's total number or

strength of connections in the network, and is interpreted as an indicator of a city's status or importance in the network.

```
> sort(rowSums(P), decreasing = TRUE)[1:5]
LONDON NEW YORK PARIS HONG KONG SINGAPORE
  1496     1403  1043      1032       913
```

In these data, London and New York have the greatest centrality, occupying the top tier of the urban hierarchy as what GaWC research calls *Alpha*++ cities [BST99]. They are followed by a second tier of *Alpha*+ cities that include Paris, Hong Kong, and Singapore. This approach appears to successfully identify what nearly any scholar of globalization would regard as the cities "used by global capital as basing points in the spatial organization and articulation of production and markets" [Fri86, p. 71].

However, these values and this weighted spatial network are less informative than they might seem. The centrality values derived from this network are almost perfectly correlated with the number of firms located in each city (i.e. the row sums of **B**).

```
> cor(rowSums(P), rowSums(cities))
[1] 0.9767704
```

The high correlation indicates that this approach to identifying central cities in a world city network is actually just identifying cities that contain many firms. This occurs because measuring a world city network using a weighted bipartite projection of firm locations guarantees that cities with many firms will have stronger connections and larger centrality values [Nea12]. If world city researchers were simply interested in finding cities with many firms, there are much simpler ways achieve this (e.g., counting a city's number of firms).

In practice, world city researchers are interested in something more nuanced: studying cities that are central in a network of economic interactions. The challenge is that although firm co-location may provide information about which cities interact economically, firm co-location is not the same

as economic interaction. The backbone package can be used to make inferences about which cities are engaged in economic interaction based on firm co-location patterns. Specifically, it can be used to estimate whether the number of firms co-located in two cities is large enough to warrant concluding that the two cities are engaged in meaningful economic interaction. The *backbone* of the world city network is a binary network in which pairs of cities are connected only if their number of co-located firms suggests they are engaged in meaningful economic interaction, and therefore provides a simplified and potentially more focused depiction of the world city network.

We'll now examine how the backbone package's functionality provides insights on both the spatial and legislative networks described.

7.2 Universal Threshold universal()

The simplest approach to backbone extraction applies a single threshold value T to all edges. As mentioned previously, often $T = 0$ is used which leads to very dense and highly clustered backbones. While we do not recommend using a universal threshold method, this is included in the backbone package for comparison purposes. The function, universal() allows the user to extract a single threshold T, or extract a signed backbone by selecting upper and lower thresholds T^+ and T^-.

For both the senate and the world cities data sets, we'll use the universal() function to compute a backbone with a single threshold of 0. Thus in the legislative network, if two senators have co-sponsored one or more bills, there will be an edge between them. Similarly, *any* number of firm co-locations is interpreted as evidence of economic interaction between a pair of cities. Notice that our backbone graph is represented by a square adjacency matrix with 0-1 entries.

```
> universalbb <- universal(senate, upper = 0, bipartite = TRUE)
> universalbb$backbone[1:5, 1:2]
                     Alexander, L. (TN-R) Boxer, B. (CA-D)
Alexander, L. (TN-R)                    0                1
Boxer, B. (CA-D)                        1                0
```

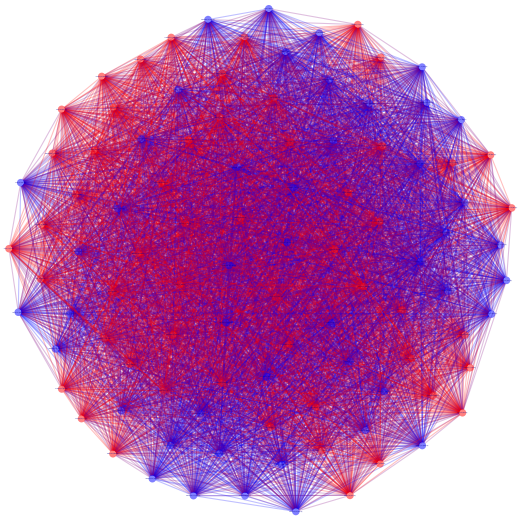

Figure 7.3: The positive backbone of the US Senate co-sponsorship network with edges retained between two senators if they sponsored at least 1 bill together.

```
Cantwell, M. (WA-D)          1          1
Carper, T. (DE-D)            1          1
Cochran, T. (MS-R)           1          1
```

The density of a network is the number of edges in the network, divided by the number of possible edges in the network. Plotting this backbone using the `igraph` package [CN06] reveals that it is extremely dense as only 1 pair of senators out of the total 4950 unique pairs have not sponsored at least one bill together (see fig. 7.3). Accordingly, this universal threshold backbone is uninformative about the underlying structure of the network. Moreover, partitioning this backbone into two groups by political party yields a modularity near zero, which indicates that this backbone does not reflect the partisan polarization known to exist in the US Senate.

We see a similar density problem occur in the world cities network.

```
> universal0 <- universal(cities, upper = 0, bipartite = TRUE)
```

```
> table(universal0$backbone)

    0     1
21506 16910
> mean(universal0$backbone)
[1] 0.4401812
> sort(rowSums(universal0$backbone), decreasing = TRUE)[1:5]
LONDON NEW YORK PARIS HONG KONG LOS ANGELES
   191      185   175       171         171
> cor(rowSums(universal0$backbone), rowSums(cities))
[1] 0.7407175
```

A backbone extracted using $T = 0$ is quite dense (44% of possible inter-city connections are present) because it treats even small numbers of firm co-locations as evidence of economic interaction between cities. As a result, the most central cities are still obviously large cities that contain many firms, and indeed, cities' centrality in this network remains highly correlated ($r = 0.74$) with their total number of firms.

A sparser network containing fewer inter-city connections can be obtained using a higher (i.e. more stringent) threshold that retains only particularly strong connections [e.g., DT05]. For example, the `universal()` function can be used to extract a backbone where $T = 25$, and therefore only cities with more than 25 co-located firms are counted as connected:

```
> universal25 <- universal(cities, upper = 25, bipartite = TRUE)
> mean(universal25$backbone)
[1] 0.001665973
> sort(rowSums(universal25$backbone), decreasing = TRUE)[1:5]
LONDON  NEW YORK HONG KONG PARIS CHICAGO
    15        12         5     5       3
> cor(rowSums(universal25$backbone), rowSums(cities))
```

```
[1] 0.8381523
```

This more stringent universal threshold is indeed much less dense (only 0.16% of possible edges are present). However, it still remains focused on the largest cities, whose centrality is highly correlated ($r = 0.84$) with the total number of firms.

These approaches involve an arbitrarily-selected threshold, however the `universal()` function can also be used to apply a universal threshold that is based on characteristics of the weighted bipartite projection **P**. For example, it is possible to extract a backbone in which cities are connected if they have more than two standard deviations above the average number of co-located firms.

```
> universal.meansd <- universal(B, upper =  function(x)mean(x)+2*sd(x),
            bipartite = TRUE)
> mean(universal.meansd$backbone)
[1] 0.03092461
> sort(rowSums(universal.meansd$backbone), decreasing = TRUE)[1:5]
LONDON NEW YORK HONG KONG PARIS SINGAPORE
    64       61        51    49        42
> cor(rowSums(universal.meansd$backbone), rowSums(cities))
[1] 0.9655334
```

This backbone is also lower density (3% of possible edges are present), but once again it focuses only on large cities, whose centrality is nearly identical to their total number of firms ($r = 0.97$).

To create a signed backbone, we can apply both an upper and lower threshold value. The following code will return a backbone where the positive edges indicate two senators co-sponsored more than 1 standard deviation above the mean number of co-sponsored bills and negative edges indicate two senators co-sponsored less than 1 standard deviation below the mean number of co-sponsored bills. The graph of the positive edges of this backbone can be seen in fig. 7.4.

```
> universalbb2 <- universal(senate, upper = function(x) mean(x)+sd(x),
```

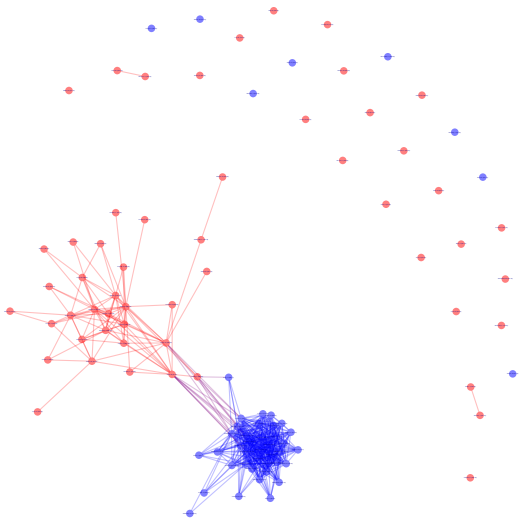

Figure 7.4: The positive backbone of the US Senate co-sponsorship network with edges retained between two senators if they sponsored more bills together than one standard deviation above the mean.

```
lower = function(x) mean(x)-sd(x), bipartite = TRUE)
```

The resulting graph in fig. 7.4 is much less dense than when using an upper threshold of 0. Additionally, the polarized structure of the Senate by political party is visible, and is confirmed by a larger modularity ($Q = 0.277$). However, it still does not necessarily reveal the underlying structure of the network among legislators. In this case, "the application of a threshold to the global weight distribution...belittles nodes with a small [degree]," resulting in a backbone that preserves edges only among legislators who sponsor many bills, and treating legislators who sponsor few bills as isolates [SBV09, p. 6484]. Similarly in the world cities network, the universal threshold backbone extraction does not take into account variations in the number of firms located in each city. By not controlling for these variations (which are substantial in this data, see 7.2A) when deciding whether two cities are connected, it privileges cities that contain many firms. In these data, because there are large variations in the number of firms located in each city that must be controlled for, a

universal threshold backbone is not appropriate.

To obtain meaningfully sparse graphs that do not ignore the multi-scalar character of node degrees we must allow the threshold to vary for different edges. To improve our backbone results, we move to methods of bipartite projection backbones that rely on a distinct threshold value for each pair of vertices.

Extracting a null model backbone: backbone.extract( )

Instead of using a universal threshold to determine a backbone, the `backbone` package incorporates the five different ensemble methods previously mentioned in chapter 6: FFM, FRM, FCM, SDSM, and FDSM. These models M do control for variation in the row and column degree sequences of $\mathbf{B}^* \in \mathcal{B}^M$. To use these methods in `backbone`, one first calls to an ensemble model function (`fixedfill()`, `fixedrow()`, `fixedcol()`, `sdsm()`, or `fdsm()`), which finds the probability of observing an edge with the observed weight in a corresponding null model, returning an object of class 'backbone.' This object contains the following: a `positive` matrix with (i, j) entry equal to the probability that G^*_{ij} is equal to or above the corresponding entry in G, and a `negative` matrix with (i, j) entry equal to the probability that G^*_{ij} is equal to or below the corresponding entry in G, and `summary`, a data frame summary of the inputted matrix and model including the class, model name, number of rows, and number of columns.

This 'backbone' object is then supplied to `backbone.extract()`, which performs the hypothesis test for a given significance value and returns a backbone graph. The user can input bipartite graph objects of class 'matrix', 'sparseMatrix', 'Matrix', 'igraph', 'network', and 'edgelist' (a matrix of two columns), and can choose the type of backbone returned by specifying the desired class in `backbone.extract()`. The `backbone.extract()` function allows the user to input the backbone class object and obtain either a signed or positive backbone. This `backbone.extract()` function has five arguments: `matrix`, `signed`, `alpha`, `class`, `narrative`, and `fwer`. The `matrix` argument takes a backbone object generated by `fixedfill()`, `fixedrow()`, `fixedcol()`, `sdsm()`, or `fdsm()` and returns a backbone graph of class = `class` using a two-tailed significance test with

significance value α = `alpha`. If the `signed` parameter is set to `TRUE` then a signed backbone is returned, if it is set to `FALSE` then a positive backbone is returned. If the `narrative` parameter is set to `TRUE` then suggested narrative text for a manuscript, including possible citations, is displayed.

Extracting the backbone of a bipartite projection involves conducting an independent statistical test on $\ell = m(m-1)/2$ edges in the projection, where m is the number of vertices in the bipartite projection. Because each of these tests is independent, this can inflate the familywise error rate beyond the desired `alpha`. The `fwer` parameter offers two ways to correct for this: the classical Bonferroni correction is applied when `fwer = 'bonferroni'`, and the more powerful Holm-Bonferroni correction is applied when `fwer = 'holm'` [Hol79].

7.3 Fixed fill model `fixedfill()`

The `fixedfill()` function will apply the fixed fill ensemble model to the bipartite network. Due to the large binomial coeffients in the probability distribution, this model as currently implemented in `backbone v1.5.0` is infeasible on large networks like the Senate data set. However, we can still apply it to the world cities network and do so below. Regardless, as we'll see in Chapter chapter 8, FFM is not the recommended model for bipartite backbone extraction when there is concern regarding the degree sequences.

```
> fixedprobs <- fixedfill(cities)
> fixedbb <- backbone.extract(fixedprobs)
```

In this null model, the number of edges in the network is held constant, that is, our observed world cities network is compared to all other possible networks with the same density. Specifically in this instance, the number of firms present in cities remains fixed, but the number of firms per company and number of firms per city may vary. Notice above we've applied the `backbone.extract()` function here after choosing the `fixedfill()` function which determined the ensemble method. Under the default settings, `backbone.extract()` has extracted a positive backbone under an alpha value of 0.05. Since all statistical tests are two-tailed tests, an edge is retained in the cities network

if the probability of two cities having the observed number of co-located firms is greater than or equal to 0.025, i.e., the upper tail of the Jacobi distribution.

```
> mean(fixedbb)
[1] 0.07418784
> sort(rowSums(fixedbb), decreasing = TRUE)[1:5]
   LONDON  NEW YORK     PARIS HONG KONG     TOKYO
       94        87        76        71        71
> cor(rowSums(fixedbb), rowSums(cities))
[1] 0.9293961
```

This FFM backbone network has a low density but again provides information focused around the largest cities. The centrality is highly correlated with number of firms. Instead of this model which compares a bipartite **B** with other networks of the same density, we'll now apply the remaining models which are based upon the degree sequences.

7.4 Fixed row model `fixedrow()`

To apply the fixed row distribution to a bipartite graph, one uses the `fixedrow()` function. The FRM is also often called hypergeometric as it estimates a hypergeometric probability distribution for each pair of nodes in the network. As an example,

```
> rowprobs <- fixedrow(senate)
> rowbb <- backbone.extract(rowprobs, alpha = .01)
```

We can now examine how this method has changed the appearance of our network, focusing only on the positive edges of the signed backbone in fig. 7.5. We can see that the FRM has reduced the density of our network and that we begin to see some of the two party structure that is inherent in the United States Senate. The known polarized structure is also apparent, which is reflected in this network's modularity ($Q = 0.215$).

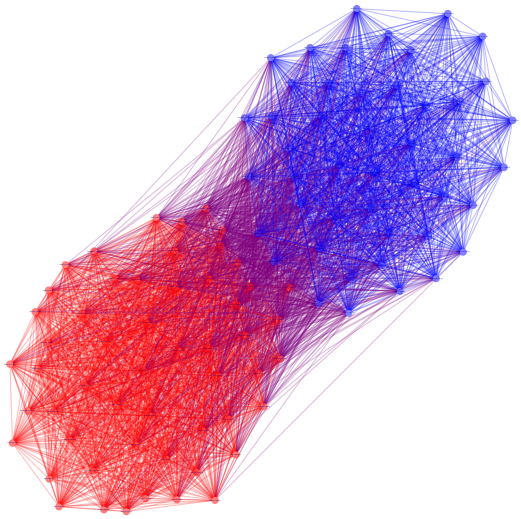

Figure 7.5: The positive backbone of the US Senate co-sponsorship network under the fixed row model.

Specifically, for our example, the fixed row function will fix the number of bills that each senator sponsors, while allowing each bill to be sponsored by a varying number of senators. The function will compute the probability of each senator sponsoring at least (or at most) the observed number of bills when the bills which they sponsor were chosen randomly.

Similarly, we can see how the fixed row model affects the world cities network.

```
> rowprobs2 <- fixedrow(cities)
> rowbb2 <- backbone.extract(rowprobs2, alpha = .1)
> mean(rowbb2)
[1] 0.09225323
> sort(rowSums(rowbb2), decreasing = TRUE)[1:5]
INDIANAPOLIS PORTLAND MELBOURNE LYON AUCKLAND
          60       54         52   49       44
```

```
> cor(rowSums(rowbb2), rowSums(cities))
[1] 0.3039028
```

First, it is less dense than the $T = 0$ universal threshold backbone, but denser than the 25-threshold or mean-threshold backbones, containing 9.2% of possible edges. That is, this model does reduce the complexity of the original network, but still preserves many intercity connections. Second, and perhaps more notably, because the FRM controls for the number of firms in each city when deciding which intercity connections to keep, it does not simply focus on cities that are large and contain many firms. Indeed, while the most central cities are major financial centers, they are not the obvious ones typically highlighted in world cities research. Moreover, cities' centrality and total firm count are only modestly correlated ($r = 0.30$), indicating that cities' centrality in this network provides information that is unique from what could have been learned from simply counting their number of firms.

Although the FRM does control for the number of firms in each city (i.e. the row sums of $\mathbf{B}^* \in \mathcal{B}^{FRM}$), it does not control for the number of cities where each firm maintains a presence (i.e. the column sums of $\mathbf{B}^* \in \mathcal{B}^{FRM}$). However, there is substantial variation in the number of cities where each firm maintains a presence (see Figure 7.2B), and not controlling for this variation can distort decisions about whether a particular city pair's number of co-located firms is significant. For example, if Firm X maintains a presence in *every city*, then observing that it is co-located in Amsterdam and New York is trivial. In contrast, if Firm Y maintains a presence in *only two cities* then observing that it is co-located in Amsterdam and New York is quite noteworthy. Because these data contain not only large variations in the number of firms in each city (see figure 7.2A) but also large variations in the number of cities where each firm maintains a presence (see figure 7.2B), the FRM is not appropriate. More generally, a FRM backbone and the `fixedrow()` function are appropriate only when there is variation in the row sums of $\mathbf{B}$, but limited variation in the column sums of $\mathbf{B}$.

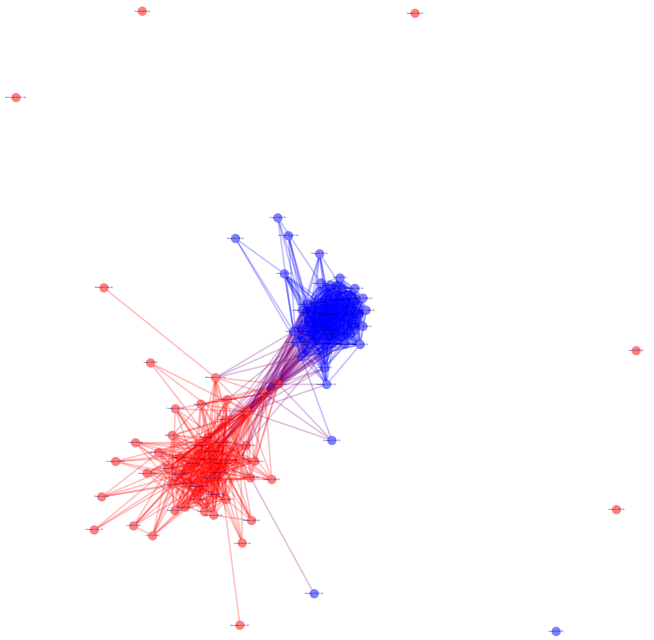

Figure 7.6: The positive backbone of the US Senate co-sponsorship network under the fixed column model.

7.5 Fixed column model `fixedcol()`

The fixed column distribution can be used through the `fixedcol()` function. In this scenario, the fixed column function fixes the number of senators that sponsor each bill, while allowing each senator to sponsor a varying number of bills.

```
> colprobs<- fixedcol(senate)
> colbb <- backbone.extract(colprobs, alpha = .01)
```

We can now examine how the fixed column model (also called Poisson binomial) has changed the appearance of our co-sponsorship network, again examining the positive edges in fig. 7.6. We can see that the fixed column function has again reduced the density of our network and the two party structure is more apparent. The known polarized structure is reflected in this network's even higher modularity ($Q = 0.424$).

We mentioned the FRM is not a good choice for the world cities network because of the substantial variation in the column sums. Here, the FCM would control for this variation in number

of cities where each firm maintains a presence, but introduces a similar problem in that the row sums are now also not controlled for. The high correlation with total number of firms exemplifies this issue.

```
> colprobs2 <- fixedcol(cities)
> colbb2 <- backbone.extract(colprobs2, alpha = 0.1, signed = FALSE)
> mean(colbb2)
[1] 0.07418784
> sort(rowSums(colbb2), decreasing = TRUE)[1:5]
LONDON  NEW YORK     PARIS HONG KONG     TOKYO
    94        87        76        71        71
> cor(rowSums(fixedcol_bb2), rowSums(cities))
[1] 0.9293961
```

We'll now attempt to approach our 'gold-standard' model, where we compare our observed data set to all other bipartite networks with the exact same degree sequences. The `backbone` package provides two ways to do this, SDSM where the degree sequences are approximately fixed and the probability mass function is known, and FDSM where the probability mass function is unknown and thus the distribution is constructed through sampling.

7.6 Stochastic degree sequence model `sdsm()`

When describing the Stochastic degree sequence model in chapter 6, we choose probabilities p^*_{ik} so that it approximates $\Pr(B^*_{ik} = 1)$ for $\mathbf{B}^* \in \mathcal{B}^{SDSM}$. Here we use the Bipartite Configuration Model or BiCM to compute those probabilities for the Poisson binomial distribution, which is used in the SDSM. In the following chapter 8, we will demonstrate why BiCM is the right choice for computing these probabilities.

In the context of the senate co-sponsorship matrix, the stochastic degree sequence model will compare our observed values to a distribution where each senator sponsors roughly the same number of bills, and each bill is sponsored by roughly the same number of people. Also demonstrated is

the 'narrative' parameter which prints out information regarding the backbone network and the citations for the model used.

```
> sdsm <- sdsm(senate)
> sdsmbb <- backbone.extract(sdsm, narrative = TRUE, alpha = .01)

Suggested manuscript text and citations:

From a bipartite graph containing 100 agents and 3589 artifacts, we obtained
the weighted bipartite projection, then extracted its binary backbone using
the backbone package (Domagalski, Neal, & Sagan, 2021). Edges were retained
in the backbone if their weights were statistically significant
(alpha = 0.01) by comparison to a null Stochastic Degree Sequence Model
(SDSM; Neal, 2014).

Domagalski, R., Neal, Z. P., and Sagan, B. (2021). backbone: An R Package
for Backbone Extraction of Weighted Graphs. PLoS ONE.
https://doi.org/10.1371/journal.pone.0244363

Neal, Z. P. (2014). The backbone of bipartite projections: Inferring
relationships from co-authorship, co-sponsorship, co-attendance and other
co-behaviors. Social Networks, 39, 84-97.
https://doi.org/10.1016/j.socnet.2014.06.001
```

We are able to see more of the partisan structure that is suggested to be present in the US Senate in fig. 7.7, and this visualization provides more information than the extremely dense graph found using a universal threshold. Moreover, the known polarized structure of the US Senate is particularly evident, and confirmed by the much larger modularity ($Q = 0.471$).

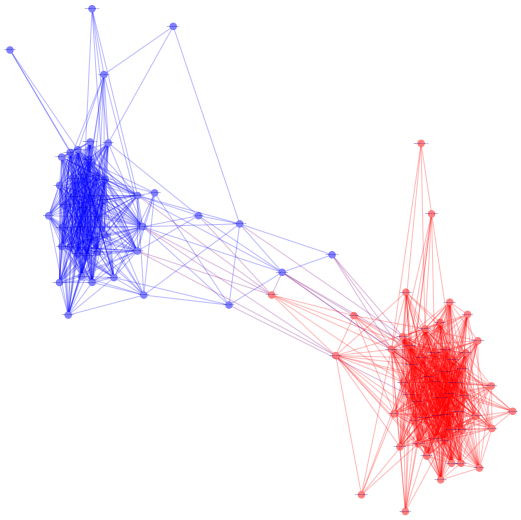

Figure 7.7: The positive backbone of the US Senate co-sponsorship network under the stochastic degree sequence model.

Before examining the entire SDSM world cities backbone, consider how it determines whether the number of co-located firms is statistically significant for a single city-pair. In fig. 7.8, three of our ensemble models are drawn. The blue curve shows the number of firms that would be co-located in Amsterdam and New York if all firms located in cities randomly, but *on average* the number of firms in each city did not change and *on average* the number of cities where each firm maintains a presence did not change. The SDSM distribution is wider and flatter than the FDSM distribution, but has nearly the same midpoint. These differences arise because the SDSM distribution is an approximation of the more targeted FDSM distribution. As an approximation with a wider distribution, the SDSM is less statistically powerful, therefore we use a more liberal threshold of statistical significance so that it will more closely mirror the FDSM. The 26 co-located firms actually observed in Amsterdam and New York is in the middle of the SDSM distribution, which indicates that this value is about what might be expected even under random conditions (i.e. not statistically significant). Therefore, the SDSM backbone does not include a link between

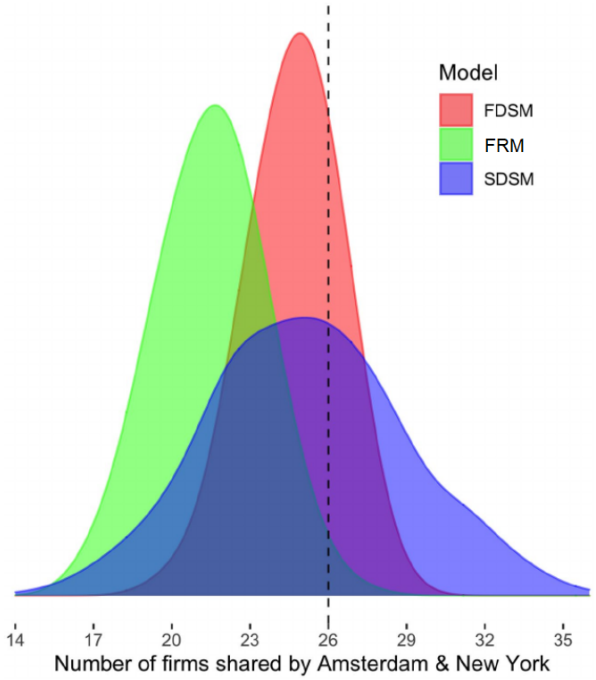

Figure 7.8: Null weight distributions generated using the backbone package on from the GaWC Dataset 11

Amsterdam and New York.

```
> sdsm2 <- sdsm(cities)
> sdsmbb2 <- sdsm(sdsm2)
> mean(sdsmbb2)
[1] 0.01973136
> sort(rowSums(sdsmbb2), decreasing = TRUE)[1:5]
KANSAS CITY CHARLOTTE RICHMOND INDIANAPOLIS BORDEAUX
         24        21       20           18       17
> cor(rowSums(sdsmbb2), rowSums(cities))
[1] -0.1062661
```

The SDSM backbone is a sparse network, in which medium-sized regional centers are the most central cities, and cities' centrality and total firm count are uncorrelated ($r = -0.11$).

7.7 Fixed degree sequence model `fdsm()`

As mentioned in the previous chapter, the fixed degree sequence model first samples random bipartite networks $B^* \in \mathcal{B}^{FDSM}$ that preserves both degree sequences using the curveball algorithm [SUG18]. These bipartite graphs B^* are then projected to obtain random weighted bipartite projection $\mathbf{P}^* = \mathbf{B}^*\mathbf{B}^{*\top}$. These two steps are repeated a number of times to sample the space of possible P^*_{ij}. At each iteration, we compare P_{ij} to the value of P^*_{ij} and keep a record of how often it was above, below, or equal to the generated value. The `fdsm()` function returns a backbone object containing a matrix object `positive` of the proportion of times P^*_{ij} is equal to or above the corresponding entry in $\mathbf{P}$, and a matrix object `negative` containing the proportion of times P^*_{ij} is equal to or below the corresponding entry in $\mathbf{P}$. This differs from the previous ensemble methods where the exact probability mass function is known and a probability can be given.

The `fdsm()` function can also save each value of P^*_{ij} for a given i, j. This is useful for visualizing an example of the empirical null edge weight distribution generated by the model. The values i, j correspond to the row and column indices of a cell in the projected matrix and can be input as either numeric values or a string containing the row names. These values are returned in the list `dyad_values`.

Using the fixed degree sequence model on the senate data set will allow us to compare our observed values to a distribution where each senator sponsors the exact same number of bills and each bill is sponsored by the exact same number of people. We can find the backbone using the fixed degree sequence model as follows:

```
> fdsm <- fdsm(senate, trials = 1000, dyad = c("Booker, C. (NJ-D)",
          "Warren, E. (MA-D)"))
```

The `dyad_values` output is a list of the G^*_{ij} values for each of the 1000 trials, where $i =$ "Booker, C. (NJ-D)" and $j =$ "Warren, E. (MA-D)". These values correspond to the number of bills Senators Booker and Warren would be expected to co-sponsor when we create a random bipartite graph with the curveball algorithm where: (a) the number of bills sponsored by Senator Booker, by

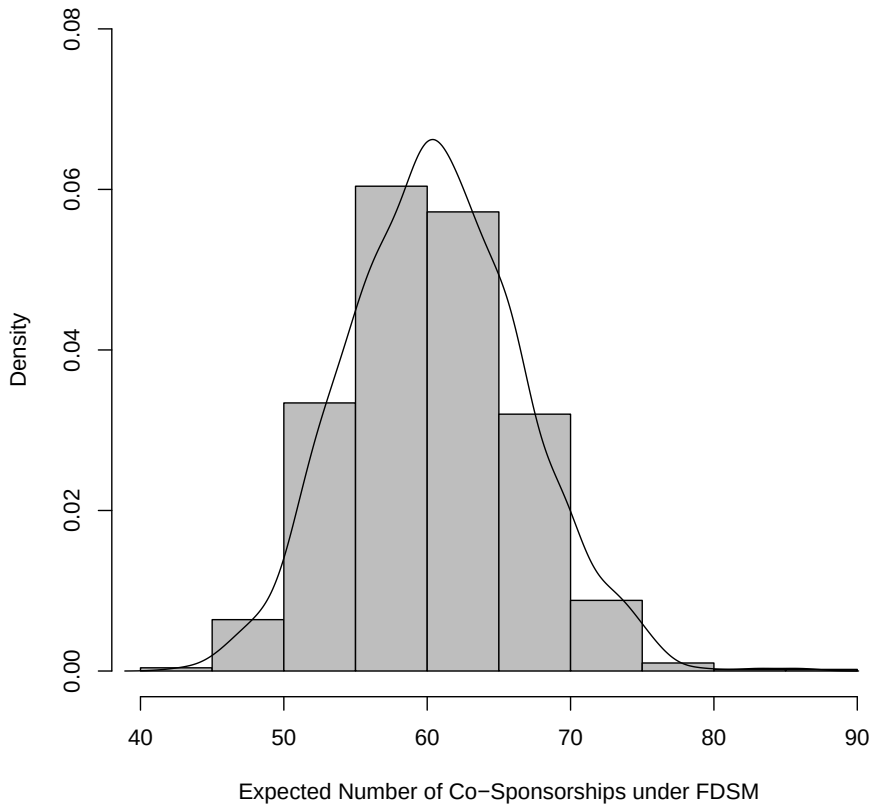

Figure 7.9: A histogram of the expected co-sponsorships between Senators Cory Booker and Elizabeth Warren under the fixed degree sequence model (1000 samples). A positive edge between Booker and Warren would be preserved in the FDSM backbone because their actual number of co-sponsorships (98) is statistically significantly larger.

Senator Warren, and all other Senators was fixed, and (b) the number of senators sponsoring each bill was fixed. We can compare their actual number of co-sponsorships, 98, to what is generated under our null model. We can view a histogram of the expected co-sponsorships generated in each of the 1000 trials as follows (see fig. 7.9):

```
> hist(fdsm$dyadvalues, freq = FALSE, xlab = "Number of Co-Sponsorships")
> lines(density(fdsm$dyadvalues))
> fdsmbb <- backbone.extract(fdsm, alpha = 0.01, signed = TRUE)
```

The FDSM backbone, based on 1000 Monte Carlo samples, requires approximately 81 seconds

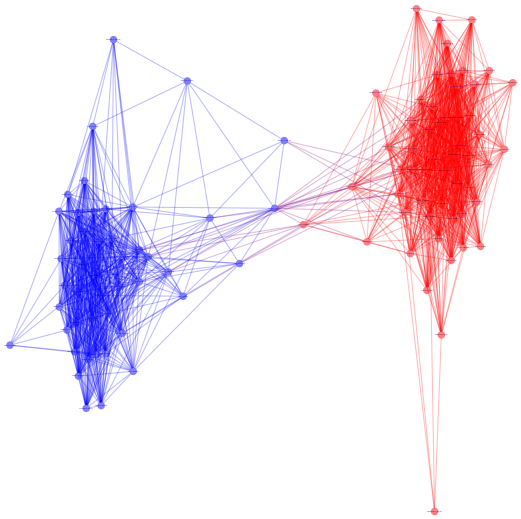

Figure 7.10: The positive backbone of the US Senate co-sponsorship network under the fixed degree sequence model.

to extract. Using the fixed degree sequence model allows us to see more of the partisan structure we assume to be present in the United States Senate in fig. 7.10. This expected partisan structure is confirmed by the backbone's high modularity ($Q = 0.468$).

The spatial network backbone extracted using FDSM is noticeably different from the other networks extracted using FFM, FRM, FCM, and SDSM.

```
> fdsm2 <- fdsm(cities, trials = 10000)
> fdsmbb2 <- backbone.extract(fdsm2, alpha = 0.1, signed = FALSE)
> mean(fdsmbb2)
[1] 0.02207414
> sort(rowSums(fdsmbb2), decreasing = TRUE)[1:5]
KANSAS CITY CHARLOTTE INDIANAPOLIS RICHMOND BORDEAUX
         24        21           20       20       17
```

```
> cor(rowSums(fdsmbb2), rowSums(cities))
[1] -0.001015871
> cor(as.vector(fdsmbb2),as.vector(sdsmbb2))
[1] 0.9315762
```

First, it has a very low density, containing only 2.2% of possible edges. Second, the cities with the highest centrality are medium-sized regional centers. Moreover, cities' centrality and total firm count are uncorrelated ($r = -0.001$), indicating that the FDSM backbone is detecting interaction patterns unrelated to a city's number of firms. Importantly, the pattern of intercity links in the SDSM and FDSM backbones are highly correlated ($r = 0.93$).

The original bipartite firm location data are known to contain substantial variation in both number of firms in each city (see figure 7.2A) but also large variations in the number of cities where each firm maintains a presence (see figure 7.2B). Because the FDSM controls for variation in these two characteristics, it is an appropriate model to use for backbone extraction in this case. Using it yields a world city network backbone that contains only those intercity links that are not simply the product of these characteristics. That is, the FDSM backbone allows world city researchers to look beyond these characteristics to identify pairs of cities with unexpectedly-large numbers of firm co-locations, which are potentially indicative of unexpectedly-strong economic interaction. More generally, the FDSM and `fdsm()` function are appropriate when there is variation in both the row sums of **B** and the column sums of **B**, which is likely to occur in most empirical bipartite data. However, although FDSM may often be the most suitable model for many empirical data, its simulation-based approach can be impractically slow when applied to bipartite data containing many agents and artifacts. As we'll see in the following chapter 8, in such cases, the SDSM model is the recommended alternative. Additionally, we'll investigate the relationship between the alpha values used in SDSM and those used in FDSM. The `backbone` R package in the future will also be home to additional backbone extraction methods, adding functionality for weighted networks that are *not* bipartite projections.

CHAPTER 8

COMPARING MODELS FOR BACKBONE EXTRACTION

All results in this chapter are from Neal, Domagalski, and Sagan [NDS21b]. Replication materials are available at https://www.github.com/domagal9/dissertation.

In this chapter we will compare the different bipartite ensemble backbone models. We begin by examining different methods for choosing the cell-filling probabilities in SDSM. As mentioned in chapter 7, this study will eventually conclude with deciding that the Bipartite Configuration Model is the best choice for these values. After having a defined SDSM to work with, we study its statistical power as compared to the FDSM backbones. Again, we'll use the world cities network for this analysis. Following this comparison, we can evaluate each of the five different models under varying degree distributions, looking to examine their speed, accuracy, similarity, and community detection. The culmination of these studies allows us to make a recommendation that in general, SDSM is the correct backbone extraction method to use for most bipartite projections.

8.1 Study 1: Choosing cell-filling probabilities for the SDSM

The SDSM requires choosing p^*_{ik}, which we want to approximate $\Pr(B^*_{ik} = 1)$ for $\mathbf{B}^* \in \mathcal{B}^{\text{FDSM}}$. There are three types of methods that might be used for doing so: arithmetic, general linear models, and entropy maximization. First, we can choose $p^*_{ik} = (r_i \times c_k)/f$, where r_i is the sum of entries in row i of $\mathbf{B}$, c_k is the sum of entries in column k of $\mathbf{B}$, and f is the sum of all entries in $\mathbf{B}$. When p^*_{ik} falls outside the $[0, 1]$ range, it is truncated toward 0 or 1, respectively [Got00]. We call this method RCF because the value is chosen based on a row sum, a column sum, and the number of entries of $\mathbf{B}$ that are filled with a one. Second, an estimate can be obtained by fitting a general linear model of the form:

$$B_{ik} = \beta_0 + \beta_1 r_i + \beta_2 c_k + \epsilon, \text{ or}$$

$$B_{ik} = \beta_0 + \beta_1 r_i + \beta_2 c_k + \beta_3 r_i c_k + \epsilon,$$

where the β's are estimated coefficients and ϵ is an error term. If the model is treated as a linear regression and the coefficients are estimated using ordinary least squares, then the predicted value of B_{ik} is chosen for p^*_{ik}, either truncating values outside the required $[0, 1]$ range (linear probability model; LPM) or transforming them into the required range using a linear discriminant model (LDM) [AWvH20]. If the model is treated as a logistic regression and the coefficients are estimated using maximum likelihood, then the predicted probability that $B_{ik} = 1$ is chosen for p^*_{ik}. In prior work, the logistic regression approach has used a scobit or logit link function, with or without an interaction term (β_3) [Nea14, SB20, Nea20]. Finally, an estimate can be obtained by entropy maximization methods, including the polytope method (Poly) [DNS21, NDY22] or bipartite configuration model (BiCM) [SSDC$^+$17]. In this study, we evaluate the accuracy and speed of these methods for choosing p^*_{ik} that approximate $\Pr(B^*_{ik} = 1)$ for $\mathbf{B}^* \in \mathcal{B}^{\text{FDSM}}$.

8.1.1 Methods

To evaluate accuracy, we begin by enumerating all the members of a small $\mathcal{B}^{\text{FDSM}}$. For example, given an agent degree sequence of $[1, 1, 2]$ and an artifact degree sequence of $[1, 1, 2]$, $\mathcal{B}^{\text{FDSM}}$ contains 5 members (see Table 8.1A). Second, from this complete enumeration, we compute the probabilities we wish p^*_{ik} to approximate (i.e., $\Pr(B^*_{ik} = 1)$ for $\mathbf{B}^* \in \mathcal{B}^{\text{FDSM}}$, see Table 8.1B). Third, we compute p^*_{ik} using each of nine methods (see Table 8.1C for values obtained using the BiCM method). Finally, we quantify the accuracy with which p^*_{ik} approximates the desired probabilities using the absolute mean difference for all i, k. In the example shown in Table 8.1, BiCM's accuracy for these degree sequences is 0.028. That is, on average p^*_{ik} chosen using BiCM deviates from the desired probabilities by $\pm$ 0.028. Because evaluating accuracy in this way requires enumerating all members of $\mathcal{B}^{\text{FDSM}}$, it is possible only for short degree sequences that define $\mathcal{B}^{\text{FDSM}}$ with small cardinality. We focus on degree sequences ranging in length from 2 to 5, which define 384 unique $\mathcal{B}^{\text{FDSM}}$ ranging in cardinality from 4 to 2040.

After identifying each method's accuracy, we evaluate the computational running time of the four most accurate methods by using them to choose p^*_{ik} for bipartite graphs defined by up to 3162

(**A**) Members of $\mathcal{B}^{\text{FDSM}}$

1	0	0
0	0	1
0	1	1

0	0	1
1	0	0
0	1	1

0	0	1
0	0	1
1	1	0

0	0	1
0	1	0
1	0	1

0	1	0
0	0	1
1	0	1

(**B**) Desired probabilities

0.2	0.2	0.6
0.2	0.2	0.6
0.6	0.6	0.8

(**C**) p^*_{ik} computed using BiCM

0.216	0.216	0.568
0.216	0.216	0.568
0.568	0.568	0.863

Table 8.1: SDSM probabilities given agent and artifact degree sequences [1,1,2]

agents and up to 3162 artifacts, and thus requiring choosing up to 10,000,000 probabilities.

8.1.2 Results

Figure 8.1A shows the accuracy of each method's computation of p^*_{ik}. Each gray line plots the accuracy of each method for a single $\mathcal{B}^{\text{FDSM}}$, while the red line plots the mean accuracy of each method over all 384 $\mathcal{B}^{\text{FDSM}}$. We find that choosing p^*_{ik} using a logistic regression with an interaction term (i.e., (Scobit-I and Logit-I)) is on average least accurate [Nea14, Nea20], while choosing p^*_{ik} using entropy maximization (i.e., BiCM and Poly) is on average most accurate [DNS21, SDCGS15].

Figure 8.1B shows the number of seconds required to compute p^*_{ik} using a 2.3 GhZ Intel i7 processor. Among the two most accurate methods, BiCM is several orders of magnitude faster than Polytope. When computing more than 10^4 probabilities, BiCM is also faster than the two slightly less accurate Logit and LDM methods. In the largest case we evaluated, computing 10^7 probabilities, BiCM took only about 0.3 seconds. Therefore, we use BiCM for choosing p^*_{ik} when extracting SDSM backbones in the remaining studies because it is both the most accurate and fastest. In previous versions of the R package `backbone`, different methods for determining these probabilities were included. However, based on these results, the `sdsm()` function uses the BiCM method.

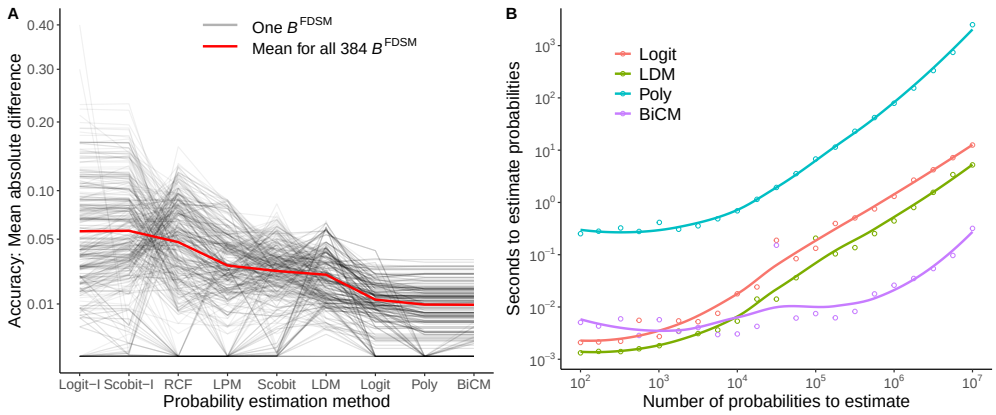

Figure 8.1: (A) Accuracy and (B) speed computing p^*_{ik} using different methods.

8.2 Study 2: Statistical power of SDSM

Ensemble backbone models require the specification of a statistical significance level α, which determines how uncommonly large an observed edge weight P_{ij} must be when compared to edge weights P^*_{ij} arising from an ensemble in order for a corresponding edge to be included in the backbone. For a given model, smaller values of α represent more stringent criteria for retaining edges, and therefore yield sparser backbones. Although FDSM and SDSM define their respective ensembles by constraining both agent and artifact degree sequences, and thus aim to yield similar backbones, a given α does not necessarily represent the same level of stringency in these two models. Because the SDSM allows variation in the degree sequences of $\mathbf{B}^* \in \mathcal{B}^{\text{SDSM}}$, the distribution of P^*_{ij} is wider. These wider distributions mean that the SDSM provides a more conservative test of edge weight significance than FDSM, or alternatively the SDSM has less statistical power to detect significant edges than FDSM.

A concrete example serves to illustrate this difference. As in chapter 7, we study the world city network using a bipartite projection where two cities are linked to the extent that firms maintain locations in both cities. Recall the Globalization and World Cities (GaWC) data set takes the form of a bipartite network recording the presence or absence of 100 firms (artifacts) in 196 cities (agents) in the year 2000 [TCW02, NDS21a]. In this bipartite network, the agent degrees are right-tailed because most cities contain only a few firms, while a few cities such as New York

contain many (see fig. 7.2). Likewise, the artifact degrees are also right tailed because most firms maintain locations in only a few cities, while a few firms such as the accounting firm KPMG maintain locations in many.

Figure 8.2A illustrates the distribution of the Milan-Paris edge weight in projections arising from $\mathcal{B}^{\text{FDSM}}$ and $\mathcal{B}^{\text{SDSM}}$ of which the observed bipartite network is a member (i.e., the random variable P^*_{ij}). These distributions allow a researcher to decide whether Milan and Paris's observed number of co-located firms is significantly large, and therefore whether Milan and Paris should be connected in a world city network backbone. The SDSM distribution is wider than the FDSM distribution, which has implications for whether the Milan-Paris edge will be included in a backbone extracted at a given significance level using each model. In the observed data, there are 26 firms co-located in Milan and Paris (i.e., $P_{ij} = 26$). The probability of observing the same or larger edge weight in projections from the FDSM ensemble is 0.0033, which is less than $\frac{0.05}{2}$, and therefore a Milan-Paris edge is deemed significant by the FDSM and is included in the FDSM backbone extracted at $\alpha = 0.05$. In contrast, the probability of observing the same or larger edge weight in projections from the SDSM ensemble is 0.0275, which is *not* less than $\frac{0.05}{2}$, and therefore a Milan-Paris edge is *not* deemed significant by the SDSM and is *not* included in the SDSM backbone extracted at $\alpha = 0.05$. For a given level of significance α, this difference in statistical power leads the SDSM backbone to be sparser than the FDSM backbone (density = 0.004 vs. 0.012), and means that these two backbones are dissimilar (Jaccard = 0.36).

In this study, we investigate SDSM's statistical power relative to FDSM, and specifically whether extracting an SDSM backbone using a more liberal (i.e., larger) α makes it more similar to an FDSM backbone extracted at $\alpha = 0.05$.

8.2.1 Methods

To evaluate SDSM's statistical power and the effect of significance levels on the similarity of SDSM and FDSM backbones, we first extracted the FDSM backbone from the GaWC bipartite network at $\alpha = 0.05$. We then extracted several SDSM backbones from the GaWC bipartite network at

$0.01 \leq \alpha \leq 0.3$ in 0.001 increments, each time computing the Jaccard index (J) to measure the similarity between the SDSM and FDSM backbones. The Jaccard index is the ratio of the edges the $\mathbf{P}'^{SDSM}$ and $\mathbf{P}'^{FDSM}$ have in common to their total edges. After comparing SDSM and FDSM backbones extracted from the empirical GaWC bipartite network, we repeat this process using 100 synthetic bipartite networks with the same dimensions (196×100), density (0.08) and right-tailed agent and artifact degree distributions.

8.2.2 Results

The green line in Figure 8.2B shows the Jaccard similarity between an FDSM backbone extracted from the empirical GaWC network at $\alpha = 0.05$ and SDSM backbones extracted at the significance levels shown on the x-axis. We find that an SDSM backbone achieves its maximum similarity to the FDSM backbone ($J = 0.81$) when it is extracted using the more liberal significance level of $\alpha = 0.12$. Returning to the example in Figure 8.2A, using this more liberal significance level would result in the Milan-Paris edge being deemed significant and included in the SDSM backbone because its SDSM p-value $0.0275 < \frac{0.12}{2}$. Because this more liberal significance level results in the inclusion of additional edges, the new SDSM backbone extracted at $\alpha = 0.12$ has a density (0.01), which is closer to that of the FDSM backbone extracted at $\alpha = 0.05$ (0.012).

The purple line in Figure 8.2B shows the mean Jaccard similarity between an FDSM backbone extracted using $\alpha = 0.05$ and SDSM backbones extracted using $0.01 \leq \alpha \leq 0.3$ from 100 bipartite networks generated to resemble the empirical GaWC network. The shaded purple region shows the 10^{th} and 90^{th} percentile of Jaccard similarities of these backbones. We find that these synthetic networks behave similarly to the empirical network. Specifically, SDSM and FDSM backbones extracted from a low-density 196×100 bipartite network with right-tailed degree distributions achieve a maximum similarity of $0.49 < J < 0.76$ when the FDSM backbone is extracted using $\alpha = 0.05$ and the SDSM backbone is extracted using $\alpha = 0.14$. This is promising because it suggests that, given the characteristics of an empirical bipartite network, it may be possible to select a significance level for extracting a computationally-efficient SDSM backbone that closely

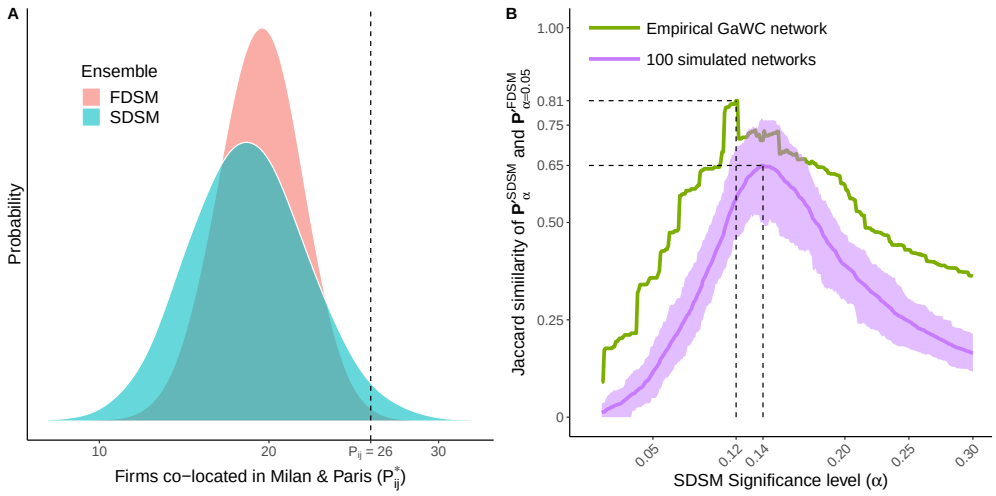

Figure 8.2: Statistical power of SDSM. (A) Distribution of weights for the Paris-Milan edge in projections derived from FDSM and SDSM ensembles. (B) Similarity of an FDSM backbone extracted at $\alpha = 0.05$ to SDSM backbones extracted at various α from an empirical bipartite network (green line) and from 100 synthetic bipartite networks (purple line = mean, purple region = 10^{th}–90^{th} percentile).

resembles a computationally-infeasible FDSM backbone.

8.3 Study 3: Backbone equivalence under varying degree distributions

Agent and artifact degree distributions are a key feature of a bipartite network, and are known to have implications for bipartite projections [VFO20, DNS21, NDS21a]. The FDSM is particularly appealing because it allows decisions about the significance of edges in a projection to be conditioned on both bipartite degree sequences, thereby taking into account these important features. However, because the computational requirements of the FDSM make it impractical for extracting the backbone from most bipartite projections, it is often necessary to use a different backbone model. In this study, we evaluate the equivalence of an FDSM backbone and backbones extracted using more computationally efficient models. We perform this comparison for backbones extracted from bipartite networks characterized by five types of degree distributions: right-tailed, left-tailed, normal, constant, and uniform.

For the sake of concreteness, in this section we use the example of a bipartite network in which authors (agents) are linked to the papers they have written (artifacts). The projection of

Degree Distribution	Authors (agents)	Papers (artifacts)
Right-tailed $\sim \beta(1, 10)$	Most write some papers, but a few are prolific (most departments).	Most papers are sole-authored, but some are written by large teams (e.g., sociology).
Left-tailed $\sim \beta(10, 1)$	Most are prolific, but some are inactive (elite departments).	Most papers are written by large teams, but some are sole-authored (e.g., physics).
Uniform $\sim \beta(1, 1)$	There is substantial diversity in scholarly output (e.g., interdisciplinary departments).	There is substantial diversity in the size of authorship teams (e.g., an entire university).
Constant $\sim \beta(10000, 10000)$	There are strong norms about how many papers an author should have (e.g., for performance evaluations).	There are strong norms about how many authors a paper should have (e.g., a senior author & a junior author)
Normal $\sim \beta(10, 10)$	Scholarly output varies around some typical level.	Authorship teams vary around some typical size.

Table 8.2: Bipartite degree distributions, with examples in the context of a scholarly authorship bipartite network

such a network yields a co-authorship network in which the edge weight between a pair of authors indicates their number of co-authored papers [New01]. These edge weight values will depend heavily on the distribution of papers written by authors (i.e., the agent degree sequence), and on the distribution of authors on each paper (i.e., the artifact degree sequence). Different degree distributions describe different kinds of scholarly environments as shown in Table 8.2. The choice of a backbone model affects whether these distributions are considered, and in this example affects whether decisions about the significance of two authors' number of co-authored papers consider the scholarly environment. The FDSM compares their observed number of co-authored papers to the number that might be observed in alternative realizations *of the same environment*, while other backbone models relax the extent to which the environment is held constant.

8.3.1 Methods

We evaluate similarities among the backbones extracted using different models by comparing backbones extracted from synthetic 100×100 bipartite networks with a density of 0.1, and with a combination of agent and artifact degree distributions shown in Table 8.2. Following our example,

these synthetic bipartite networks might represent a college of 100 faculty who collectively wrote 100 papers, in a particular type of scholarly environment where each individual had a 10% chance of being an author on each paper. After generating a bipartite network with a given size, density, and degree distributions, we extract five different backbones from the generated bipartite network, using the fixed fill model, fixed row model, fixed column model, stochastic degree sequence model, and fixed degree sequence model; in all cases we use $\alpha = 0.05$. We compute the similarity of the first four backbones to the FDSM backbone using a Jaccard index, repeating this process 100 times for each of the 25 possible combinations of agent and artifact degree distributions.

8.3.2 Results

The heatmaps in Figure 8.3 illustrate the similarity between an FDSM backbone and a backbone extracted using an alternative model. The rows of each heat map correspond to different agent degree distributions, and the columns correspond to different artifact degree distributions, in the synthetic bipartite networks from which the backbones were extracted. The lightest patches identify conditions under which a given backbone model yields a backbone that is similar to what would be obtained using the computationally costly FDSM, while darker patches identify conditions under which these two backbones differ. We find that when agent degrees are constant (i.e., every agent has the same degree) and artifact degrees are constant or left-tailed, all backbone models yield the same backbone as FDSM (Mean $J = 1$). However, beyond this special case, which is likely to be rare in empirical data, similarity to FDSM-extracted backbones varies.

As expected, the similarity of backbones extracted using FRM and FDSM depends primarily on the distribution of artifact degrees, not agent degrees (see Figure 8.3B). For example, for any agent degree distribution, these two models yield very different backbones when artifact degrees follow a right-tailed distribution (Mean $J = 0.186$), but very similar backbones when artifact degrees follow a normal distribution (Mean $J = 0.863$). This occurs because both models exactly control for agent degrees, however FDSM also controls for artifact degrees, while FRM does not.

A similar but rotated pattern emerges when considering the FCM: the similarity of backbones

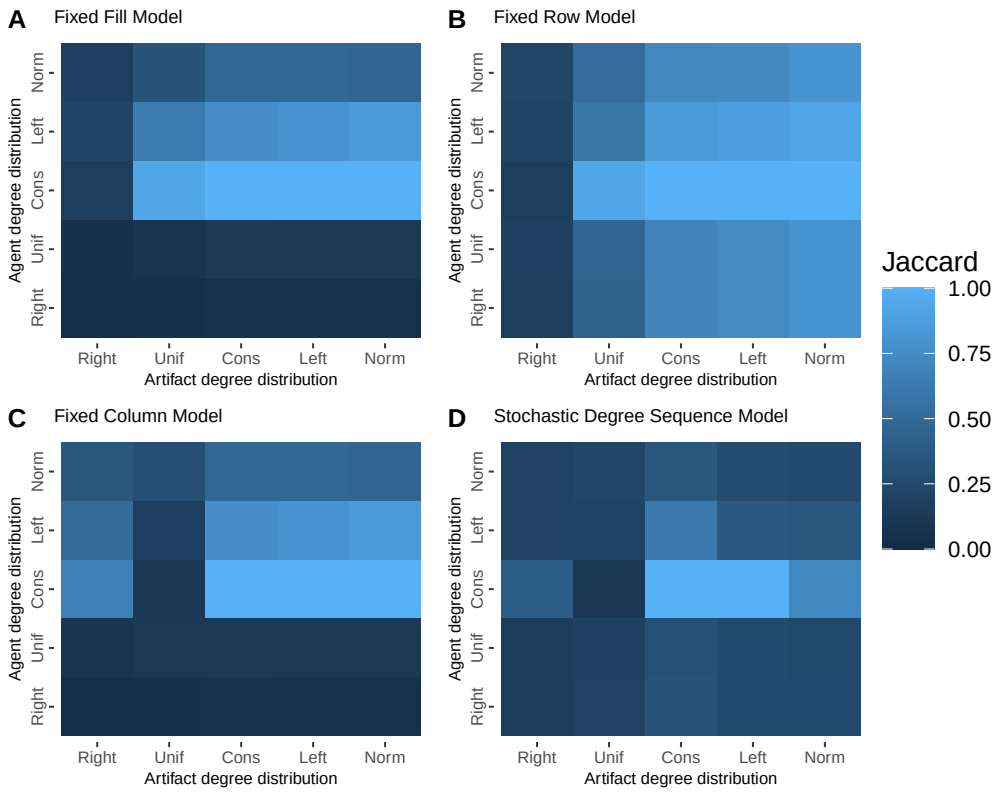

Figure 8.3: Jaccard similarity of a backbone extracted at $\alpha = 0.05$ using the Fixed Degree Sequence Model and a backbone extracted using (A) the Fixed Fill Model, (B) Fixed Row Model, (C) Fixed Column Model, (D) Stochastic Degree Sequence Model. Each cell represents the mean over 100 instances of a 100×100 bipartite network with given agent and artifact degree distributions.

extracted using FCM and FDSM depends primarily on the distribution of agent degrees, not artifact degrees (see Figure 8.3C). For any artifact degree distribution, these two models yield very different backbones when agent degrees follow a right-tailed or uniform (Mean $J = 0.084$) distribution , but more similar backbones when agent degrees follow a left-tailed distribution or are constant (Mean $J = 0.617$). This occurs because both models exactly control for artifact degrees, however FDSM also controls for agent degrees, while FRM does not. However, there is a notable exception to this general pattern: when artifact degrees follow a uniform distribution, FCM and FDSM always yield different backbones (Mean $J = 0.151$).

The conditions under which the FFM yields FDSM-similar backbones occur at the intersection of the conditions under which the FRM and FCM both yield FDSM-equivalent backbones (see

Figure 8.3A). When artifact degrees follow a right-tailed distribution and/or the agent degrees follow a right-tailed or uniform distribution, then FFM and FDSM backbones differ (Mean $J = 0.1$). In contrast, for other combinations of degree distributions, FFM and FDSM backbones are more similar (Mean $J = 0.724$).

Finally, as expected based on the findings from study 2, we observe that the SDSM generally yields different backbones than FDSM when both are extracted at $\alpha = 0.05$ (see Figure 8.3D). Specifically, except in the narrow case where agent degrees are constant and artifact degrees are constant or left-tailed (Mean $J = 1$), SDSM and FDSM backbones exhibit only modest similarity (Mean $J = 0.314$). This lack of similarity or equivalence occurs because SDSM offers a less statistically powerful (or more conservative) test of edges statistical significance than FDSM, and therefore retains fewer edges in the backbone. However, findings from study 2 also suggested that careful selection of the significance level used for extracting an SDSM backbone can yield results more similar to FDSM.

To explore this possibility, we expanded the analysis reported in figure 8.3D by extracting SDSM backbones at different significance levels. We find that when a suitably more liberal (i.e., larger) significance level α is used to extract an SDSM backbone, the resulting SDSM backbone is very similar to an FDSM backbone extracted at $\alpha = 0.05$ (see Figure 8.4A). Specifically, for backbones extracted from bipartite networks with *any* agent or artifact degree distributions, these two backbones tend to be nearly equivalent (Mean $J = 0.865$). This suggests that in principle the fast SDSM can be used to obtain a close approximation of a computationally-infeasible FDSM backbone from any bipartite network.

In practice, using SDSM to obtain an FDSM-like backbone requires selecting an α value for the SDSM that corresponds to $\alpha = 0.05$ in the FDSM. We observe that there are three distinct values of such an 'optimal' α that depend on agent and artifact degree distributions (see Figure 8.4B). First, when agent degrees are constant, a value only slightly higher than 0.05 (Mean = 0.062, SD = 0.021) achieves the best approximation of an FDSM backbone. Second, when artifact degrees are constant, a value roughly double (Mean = 0.09, SD = 0.022) achieves the best approximation

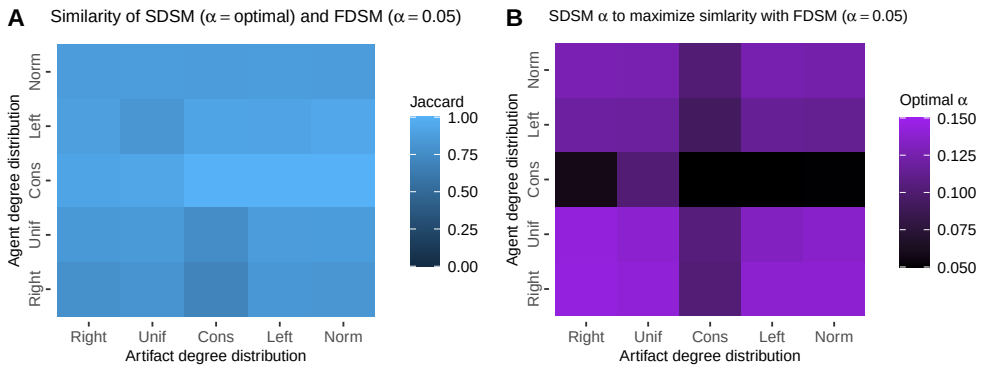

Figure 8.4: (A) Given agent and artifact degree distributions, there exists a statistical significance level α that maximizes the similarity between an SDSM backbone extracted at this level and an FDSM backbone extracted at $\alpha = 0.05$, and (B) when used yields an SDSM backbone that is very similar to the corresponding FDSM backbone.

of an FDSM backbone. Finally, when neither agent nor artifact degrees are constant, which is likely in most empirical bipartite networks, a value roughly 2.5 times larger (Mean = 0.13, SD = 0.014) achieves the best approximation of an FDSM backbone. Although further work is needed to facilitate the *a priori* selection of an α that allows an SDSM backbone to closely approximate an FDSM backbone, these results suggest that under the most common circumstances (i.e., when there is variation in degrees) $\alpha \approx 0.13$ may be appropriate.

8.4 Study 4: Recovery of community structure

Studies 1-3 examine the backbones extracted from synthetic random bipartite networks; however, empirical bipartite networks are generally not random, but instead have a clustered or blocked structure. In this study, we evaluate the extent to which backbones extracted using different models reflect a known community structure that is encoded in the bipartite data from which they are extracted [CWW18]. As shown in chapter 7, SDSM and FDSM backbones extracted from a bipartite network representing bill co-sponsorship in the 114^{th} session of the US Senate more clearly captured the known partisan community structure than an FRM backbone [DNS21]. For the sake of concreteness, we use this legislative network context as an example in this section, but we extend this prior work by considering a broader range of backbone models, and by examining their ability

to recover community structures from bipartite data containing varying levels of evidence for this structure.

8.4.1 Methods

We investigate the ability for backbones to recover a known community structure in three steps. First, we simulate a 200×1000 bipartite network with a density of 0.1 and right-tailed agent and artifact degree distributions. We focus on a bipartite network with more artifacts than agents to ensure that these data contain sufficient information to encode potential community memberships. We focus on a bipartite network with right-tailed degree distributions because they are common in many empirical unipartite [BC19] and bipartite networks [Nea20, NDS21a, AABB11]. Similar to the Senate data set we examined in chapter 7, this synthetic bipartite network could represent a legislative body composed of 200 legislators casting votes on 1000 bills, where any given legislator had a 10% chance of voting in favor of any given bill. The right-tailed degree distributions capture the fact that most legislators vote in favor of only a few bills, and that most bills receive the support of only a few legislators, which is typical of legislative bodies. The backbone of a projection of such a bipartite network would represent a network of collaboration or ideological alignment among legislators [Nea20].

Second, we incorporate evidence of communities in this bipartite network by randomly assigning each agent and each artifact to one of two groups. We then perform checkerboard swaps, which preserve the degree distributions, until a given fraction of edges W are within-group, connecting an agent and artifact from the same group [GSPA07]. Figure 8.5A provides graphical depictions of the matrices describing synthetic bipartite networks at two values of W. In each plot, the rows represent agents assigned to group A or B, the columns represent artifacts assigned to group A or B, and a cell is shaded black if the row agent is connected to the column artifact. When $W = 0.5$, agents in a given group are equally likely to associate with artifacts in either group, placing ≈ 0.5 of the edges (i.e., shaded cells) in the diagonal blocks and ≈ 0.5 of the edges in the off-diagonal blocks. In contrast, when $W = 0.8$, agents in a given group are much more likely to associate with

artifacts from their own group than artifacts in the other group, placing ≈ 0.8 of the edges in the diagonal blocks and ≈ 0.2 of the edges in the off-diagonal blocks. Returning to our example, the groups could represent political parties: each legislator belongs to one of two parties (i.e., there are conservative and liberal legislators), and each bill advances the agenda of one of these parties (i.e., there are conservative and liberal bills). When $W = 0.5$, a conservative legislator is equally likely to vote for conservative and liberal bills, while when $W = 0.8$, a conservative legislator is four-times more likely to vote for a conservative bill than a liberal bill.

Finally, we extract a backbone from the bipartite network using a given model and compute the backbone's modularity Q with respect to the agents' group assignments [NG04]. If a backbone model is able to recover the community structure from evidence in the bipartite network, then we expect a positive association between W and Q. In the legislative example, if legislators are bipartisan in their voting patterns (i.e., $W = 0.5$), then legislators should not be clustered by party in the backbone (i.e., $Q \approx 0$). In contrast, if legislators are strongly partisan in their voting patterns (i.e., $W = 0.8$), then legislators should be clustered by party in the backbone (i.e., $Q \approx 0.5$).

We repeat these three steps 10 times for $0.5 \leq W \leq 0.8$ in 0.05 increments. When evaluating the SDSM backbone, we consider both a backbone extracted using the conventional significance level of $\alpha = 0.05$ and one extracted at the more liberal $\alpha = 0.13$, which study 3 suggests yields a backbone similar to FDSM.

8.4.2 Results

Figure 8.5B shows the modularity (y-axis; with respect to known community memberships) of backbones extracted using different models from bipartite networks with different fractions of within-community edges (x-axis). All six lines increase monotonically, confirming that all backbone models yield backbones that can recover a known community structure. However, there is notable variation among the models. As evidence of community structure grows stronger in the bipartite network, the modularity of backbones extracted using the FFM and FCM slowly increase, but even when the evidence of such a structure is quite strong (i.e., when $W = 0.8$) they only achieve

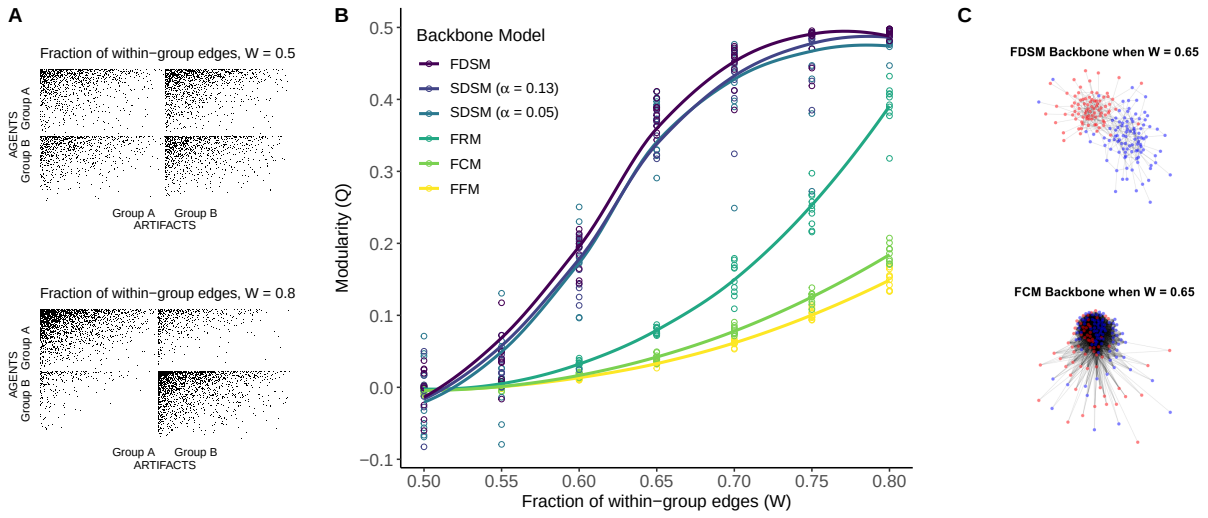

Figure 8.5: (A) Synthetic bipartite networks with varying levels of block structure, from which (B) backbones extracted using different models exhibit varying modularity. (C) When 65% of bipartite edges are within-block, a backbone extracted using FDSM shows a clear group structure (top) while a backbone extracted using FCM does not (bottom).

average values of $Q = 0.15$ and 0.18, respectively. Backbones extracted using the FRM display a similar pattern, but achieve a higher average modularity ($Q = 0.39$) value when W is large.

In contrast, backbones extracted using FDSM and SDSM are virtually indistinguishable in their ability to recover the known community structure, and do so very well. As evidence of a community structure grows stronger in the bipartite network, the modularity of backbones extracted using these models rapidly increases. When the evidence of community structure is strong (i.e., when $W = 0.8$), these backbones have very high modularity (mean $Q = 0.49$). However, even when there is only modest evidence of community structure in the bipartite network (e.g., when $W = 0.65$), these backbones are still able to identify the community structure and have a distinctively high modularity (mean $Q = 0.37$).

Figure 8.5C illustrates the difference between two backbone models' abilities to recover a known community structure, when evidence of that structure is modest in the bipartite data from which the backbone is extracted ($W = 0.65$). In both plots, agents from group A (e.g., conservatives, in the legislative example) are colored red, while agents from group B (e.g., liberals, in the legislative example) are colored blue. The FDSM-extracted backbone clearly places agents from different

groups in separate clusters. In contrast, the FCM-extracted backbone is unable to distinguish this group structure and fails to cluster agents according to their known group memberships. These findings suggest that although all backbone models can yield backbones that recover a known community structure, SDSM and FDSM backbones are able to detect this structure more clearly and from a weaker signal.

8.5 Recommendations for Backbone Selection

Bipartite networks can be used to represent a wide range of phenomena in the social and natural worlds including interspecies competition, global trade, scientific advances, and legislative deliberation. Likewise, projections of bipartite networks, which take the form of co-occurrence networks, can be useful for inferring unipartite networks that would otherwise be difficult to measure directly. Several models have been proposed for extracting the backbone of bipartite projections, and thus for making such inferences, including the fixed fill model (FFM), fixed row model (FRM), fixed column model (FCM), fixed degree sequence model (FDSM), and stochastic degree sequence model (SDSM). We have introduced each of these models and found their probability mass functions in chapter 6. To facilitate their use, we have described the `R` package `backbone` where we have implemented each model in chapter 7. We then systematically compared these models in terms of their relative accuracy, speed, statistical power, similarity, and ability to recover a known community structure in chapter 8.

In study 1, we examined several methods for choosing the probabilities necessary for applying the stochastic degree sequence model (SDSM), finding that the bipartite configuration model (BiCM) is both the fastest and most accurate. In study 2, we examined the statistical power of the SDSM relative to the fixed degree sequence model (FDSM), finding that the SDSM can be viewed as a statistically less powerful (or more conservative) variant of the FDSM. In study 3, we examined the similarity of an FDSM-extracted backbone to backbones extracted using other models, finding that the SDSM and FDSM extract very similar backbones from bipartite networks with a wide range of possible degree distributions when an appropriate significance level α is chosen. Finally,

in study 4, we examined the ability for backbones extracted using different models to recover a known community structure, finding that although all models can recover the structure, SDSM and FDSM can detect a community structure more clearly and from a weaker signal.

Based on these findings, and with the goal of offering researchers some guidance in extracting the backbones of bipartite projections, we offer three recommendations. First, *we recommend the stochastic degree sequence model (SDSM) for extracting the backbones of bipartite projections* because it is fast, controls for both agent and artifact degree sequences, and yields modular backbones when the bipartite data contains even modest evidence of within-community clustering. Second, when the SDSM is used, *we recommend that the cell-filling probabilities p^*_{ik} be chosen using the Bipartite Configuration Model (BiCM)* because it is faster and more accurate than any other currently available method. Third, when an FDSM backbone extracted at the $\alpha = 0.05$ significance level is desired but computationally infeasible, *we recommend extracting an SDSM backbone at the $\alpha = 0.13$ significance level*, which we observe is very similar when there is variation in the agent and artifact degree sequences. The models and options necessary to adopt these recommendations are implemented in the `backbone` package for **R** [DNS21].

These findings and recommendations must be viewed in light of the fact that, due to the computational requirements of the FDSM and of extracting a large number of backbones across the four studies, these studies have relied on small synthetic bipartite networks ranging in size from 3×3 (study 1) to 200×1000 (study 4). However, in practice bipartite networks may be several orders of magnitude larger. For example, a bipartite network used to infer collaborations in the US House of Representatives includes 435 agents (representatives) and over 6000 artifacts (bills) [Nea20, NDY22], while a bipartite network used to infer movie recommendations includes 17,770 agents (films) and nearly 500,000 artifacts (viewers) [ZK11]. Future research should explore whether these findings extend to backbones extracted from such large bipartite networks. Limitations of existing backbone models also point to directions for future research. First, using the FDSM will generally be computationally infeasible in practice because the distribution of P^*_{ij} arising from $\mathcal{B}^{\text{FDSM}}$ must be estimated via numerical simulation. Identifying this distribution's

probability mass function, which is known for the other ensembles (as discussed in chapter 6), would facilitate the use of this otherwise attractive model; however, this is a well-studied problem and so is probably very hard to solve. Second, all the ensemble models we have considered impose constraints on the degree sequences, but other types of constraints may also be useful. For example, in some contexts it may be necessary to constrain all members of an ensemble to contain a 0 in a particular cell (e.g., to represent that an author was not alive to co-author a paper, or a legislator was not present to co-sponsor a bill). These limitations and future directions notwithstanding, the results presented above provide a starting point for further development of backbone models, and provide applied researchers with some practical guidance on model selection.

BIBLIOGRAPHY

BIBLIOGRAPHY

[AABB11] Yong-Yeol Ahn, Sebastian E Ahnert, James P Bagrow, and Albert-László Barabási. Flavor network and the principles of food pairing. *Scientific Reports*, 1(1):1–7, 2011.

[AGRR20] Ron M. Adin, Ira M. Gessel, Victor Reiner, and Yuval Roichman. Cyclic quasi-symmetric functions. *Sém. Lothar. Combin.*, 82B:Art. 67, 12, 2020.

[ALH+15a] C. Andris, D Lee, M. J. Hamilton, M. Martino, C. E. Gunning, and J. A. Selden. The rise of partisanship and super-cooperators in the us house of representatives. *PloS One*, 10:e0123507, 2015.

[ALH+15b] Clio Andris, David Lee, Marcus J Hamilton, Mauro Martino, Christian E Gunning, and John Armistead Selden. The rise of partisanship and super-cooperators in the us house of representatives. *PloS One*, 10(4):e0123507, 2015.

[AN20a] S. Aref and Z. P. Neal. Detecting coalitions by optimally partitioning signed networks of political collaboration. *Scientific Reports*, 10:1506, 2020.

[AN20b] Samin Aref and Zachary P. Neal. Detecting coalitions by optimally partitioning signed networks of political collaboration. *Scientific reports*, 10(1):1–10, 2020.

[And87] Désiré André. Solution directe du probleme résolu par m. bertrand. *CR Acad. Sci. Paris*, 105(436):7, 1887.

[AWvH20] Paul Allison, R. A. Williams, and P. von Hippel. Better predicted probabilities from linear probability models with applications to multiple imputation. 2020 Stata Conference 1, Stata Users Group, 2020.

[BBMD+02] Cyril Banderier, Mireille Bousquet-Mélou, Alain Denise, Philippe Flajolet, Danièle Gardy, and Dominique Gouyou-Beauchamps. Generating functions for generating trees. *Discrete Math.*, 246:29–55, 2002. Formal power series and algebraic combinatorics (Barcelona, 1999).

[BBPS15] Sara Billey, Krzysztof Burdzy, Soumik Pal, and Bruce E. Sagan. On meteors, earthworms and WIMPs. *Ann. Appl. Probab.*, 25(4):1729–1779, 2015.

[BBS13] Sara Billey, Krzysztof Burdzy, and Bruce E. Sagan. Permutations with given peak set. *J. Integer Seq.*, 16(6):Article 13.6.1, 18, 2013.

[BC19] Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 10(1):1–10, 2019.

[BCS+17] Christopher R Browning, Catherine A Calder, Brian Soller, Aubrey L Jackson, and Jonathan Dirlam. Ecological networks and neighborhood social organization. *American Journal of Sociology*, 122(6):1939–1988, 2017.

[BDS+20] Amanda N Buerger, David T Dillon, Jordan Schmidt, Tao Yang, Jasenka Zubcevic, Christopher J Martyniuk, and Joseph H Bisesi Jr. Gastrointestinal dysbiosis following diethylhexyl phthalate exposure in zebrafish (danio rerio): Altered microbial diversity, functionality, and network connectivity. *Environmental Pollution*, 265:114496, 2020.

[Ber87] J. Bertrand. Solution d'un problème. *CR Acad. Sci. Paris*, 105:369, 1887.

[BFT16] Sara Billey, Matthew Fahrbach, and Alan Talmage. Coefficients and roots of peak polynomials. *Exp. Math.*, 25(2):165–175, 2016.

[BM03] Mireille Bousquet-Mélou. Four classes of pattern-avoiding permutations under one roof: generating trees with two labels. *Electron. J. Combin.*, 9(2):Research paper 19, 31, 2002/03. Permutation patterns (Otago, 2003).

[Bón04] Miklós Bóna. *Combinatorics of permutations*. Discrete Mathematics and its Applications (Boca Raton). Chapman & Hall/CRC, Boca Raton, FL, 2004.

[BR11] K. A. Bratton and S. M. Rouse. Networks in the legislative arena: How group dynamics affect cosponsorship. *Legislative Studies Quarterly*, 36:423–460, 2011.

[BR17] Pierre-Alexandre Balland and David Rigby. The geography of complex knowledge. *Economic Geography*, 93(1):1–23, 2017.

[BS00] Eric Babson and Einar Steingrímsson. Generalized permutation patterns and a classification of the Mahonian statistics. *Sém. Lothar. Combin.*, 44:Art. B44b, 18, 2000.

[BST99] Jonathan V Beaverstock, Richard G Smith, and Peter J Taylor. A roster of world cities. *cities*, 16(6):445–458, 1999.

[Cal02] David Callan. Pattern avoidance in circular permutations. Preprint `arXiv:0210014`, 2002.

[Car15] C. J. Carstens. Proof of uniform sampling of binary matrices with fixed row sums and column sums for the fast curveball algorithm. *Physical Review E*, 91(4), Apr 2015.

[CGHK78] F. R. K. Chung, R. L. Graham, V. E. Hoggatt, Jr., and M. Kleiman. The number of Baxter permutations. *J. Combin. Theory Ser. A*, 24(3):382–394, 1978.

[CN06] Gabor Csardi and Tamas Nepusz. *The `igraph` software package for complex network research*, 2006.

[CP87] G. A. Caldeira and S. C. Patterson. Political friendship in the legislature. *The Journal of Politics*, 49:953–975, 1987.

[CVDLO+17] Francis Castro-Velez, Alexander Diaz-Lopez, Rosa Orellana, José Pastrana, and Rita Zevallos. The number of permutations with the same peak set for signed permutations. *J. Comb.*, 8(4):631–652, 2017.

[CW90] Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. *Journal of Symbolic Computation*, 9:251–280, 1990.

[CWW18] Tristan JB Cann, Iain S Weaver, and Hywel TP Williams. Is it correct to project and detect? assessing performance of community detection on unipartite projections of bipartite networks. In *International Conference on Complex Networks and their Applications*, pages 267–279. Springer, 2018.

[CXW+18] Xing Chen, Di Xie, Lei Wang, Qi Zhao, Zhu-Hong You, and Hongsheng Liu. BNPMDA: Bipartite network projection for mirna–disease association prediction. *Bioinformatics*, 34(18):3178–3186, 2018.

[DDJ+12] Theodore Dokos, Tim Dwyer, Bryan P. Johnson, Bruce E. Sagan, and Kimberly Selsor. Permutation patterns and statistics. *Discrete Math.*, 312(18):2760–2775, 2012.

[Dia75] Jared M Diamond. Assembly of species communities. In M. L. Cody and J. M. Diamond, editors, *Ecology and evolution of communities*, pages 342–444. Harvard University Press, 1975.

[Dia16] Navid Dianati. Unwinding the hairball graph: Pruning algorithms for weighted complex networks. *Physical Review E*, 93(1):012304, 2016.

[DL16] Ben Derudder and Xingjian Liu. How international is the annual meeting of the association of american geographers? A social network analysis perspective. *Environment and Planning A*, 48(2):309–329, 2016.

[DLHH+21] Alexander Diaz-Lopez, Pamela E. Harris, Isabella Huang, Erik Insko, and Lars Nilsen. A formula for enumerating permutations with a fixed pinnacle set. *Discrete Math.*, 344(6):112375, 2021.

[DLHIO17a] Alexander Diaz-Lopez, Pamela E. Harris, Erik Insko, and Mohamed Omar. A proof of the peak polynomial positivity. *Sém. Lothar. Combin.*, 78B:Art. 6, 9, 2017.

[DLHIO17b] Alexander Diaz-Lopez, Pamela E. Harris, Erik Insko, and Mohamed Omar. A proof of the peak polynomial positivity conjecture. *J. Combin. Theory Ser. A*, 149:21–29, 2017.

[DLHIPL17] Alexander Diaz-Lopez, Pamela E. Harris, Erik Insko, and Darleen Perez-Lavin. Peak sets of classical Coxeter groups. *Involve*, 10(2):263–290, 2017.

[DLIN21] Alexander Diaz-Lopez, Erik Insko, and Lars Nilsen. Pinnacle ordering. In preparation, 2021.

[DLM+21a] Rachel Domagalski, Jinting Liang, Quinn Minnich, Bruce E. Sagan, Jamie Schmidt, and Alexander Sietsema. Cyclic pattern containment and avoidance. *arXiv:2106.02534 [math]*, Jun 2021. arXiv: 2106.02534.

[DLM+21b] Rachel Domagalski, Jinting Liang, Quinn Minnich, Bruce E. Sagan, Jamie Schmidt, and Alexander Sietsema. Cyclic shuffle compatibility. *arXiv:2106.10182 [math]*, Jun 2021. arXiv: 2106.10182.

[DLM+21c] Rachel Domagalski, Jinting Liang, Quinn Minnich, Bruce E. Sagan, Jamie Schmidt, and Alexander Sietsema. Pinnacle set properties. *arXiv:2105.10388 [math]*, May 2021. arXiv: 2105.10388.

[DMSK15] B. A. Desmarais, V. G. Moscardelli, B. F. Schaffner, and M. S. Kowal. Measuring legislative collaboration: The senate press events network. *Social Networks*, 40:43–54, 2015.

[DNKPT18] Robert Davis, Sarah A. Nelson, T. Kyle Petersen, and Bridget E. Tenner. The pinnacle set of a permutation. *Discrete Math.*, 341(11):3249–3270, 2018.

[DNS20] Rachel Domagalski, Zachary P. Neal, and Bruce Sagan. *backbone: Extracts the Backbone from Weighted Graphs*, 2020. R package version 1.2.0.

[DNS21] Rachel Domagalski, Zachary P Neal, and Bruce Sagan. Backbone: An R package for extracting the backbone of bipartite projections. *PloS One*, 16(1):e0244363, 2021.

[Dru16] L. Drutman. *American politics has reached peak polarization*, 2016.

[DT05] Ben Derudder and Peter Taylor. The cliquishness of world cities. *Global Networks*, 5(1):71–91, 2005.

[ES35] P. Erdős and G. Szekeres. A combinatorial problem in geometry. *Compositio Math.*, 2:463–470, 1935.

[ES21] Sergi Elizalde and Bruce Sagan. Consecutive patterns in circular permutations. 2021.

[Fan21] Wenjie Fang. Efficient recurrence for the enumeration of permutations with fixed pinnacle set. *arXiv:2106.09147 [math]*, Jun 2021. arXiv: 2106.09147.

[FNT21] Justine Falque, Jean-Christophe Novelli, and Jean-Yves Thibon. Pinnacle sets revisited. *arXiv:2106.05248 [math]*, Jun 2021. arXiv: 2106.05248.

[Fon20] Christian Fong. Expertise, networks, and interpersonal influence in congress. *The Journal of Politics*, 82(1):269–284, 2020.

[Fow06a] J. H. Fowler. Connecting the congress: A study of cosponsorship networks. *Political Analysis*, 14:456–487, 2006.

[Fow06b] J. H. Fowler. Legislative cosponsorship networks in the us house and senate. *Social Networks*, 28:454–465, 2006.

[Fri86] John Friedmann. The world city hypothesis. *Development and change*, 17(1):69–83, 1986.

[GL04] Jean-Loup Guillaume and Matthieu Latapy. Bipartite structure of all complex networks. *Information Processing Letters*, 90(5):215–221, 2004.

[GLW18] Daniel Gray, Charles Lanning, and Hua Wang. Pattern containment in circular permutations. *Integers*, 18B:Paper No. A4, 13, 2018.

[GLW19] Daniel Gray, Charles Lanning, and Hua Wang. Patterns in colored circular permutations. *Involve*, 12(1):157–169, 2019.

[Got00] Nicholas J Gotelli. Null model analysis of species co-occurrence patterns. *Ecology*, 81(9):2606–2621, 2000.

[GSPA07] Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral. Module identification in bipartite and directed networks. *Physical Review E*, 76(3):036102, 2007.

[HBKM09] Emilie M Hafner-Burton, Miles Kahler, and Alexander H Montgomery. Network analysis for international relations. *International organization*, pages 559–592, 2009.

[HFC16] Eelke M Heemskerk, Meindert Fennema, and William K Carroll. The global corporate elite after the financial crisis: evidence from the transnational network of interlocking directorates. *Global Networks*, 16(1):68–88, 2016.

[HKBH07] César A Hidalgo, Bailey Klinger, A-L Barabási, and Ricardo Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007.

[Hol79] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[Ing15] Christopher Ingraham. A stunning visualization of our divided congress. *Washington Post*, Apr 2015.

[Jac61] Jane Jacobs. *The death and life of great American cities*. Random House, 1961.

[Kir11] J. H. Kirkland. The relational determinants of legislative outcomes: Strong and weak ties between legislators. *The Journal of Politics*, 73:887–898, 2011.

[KK96] Daniel Kessler and Keith Krehbiel. Dynamics of cosponsorship. *The American Political Science Review*, 90(3):555–566, 1996.

[KMN16] G. Koger, S. Masket, and H. Noel. No disciplined army: American political parties as networks. In J. N. Victor, A. H. Montgomery, and Lubell M., editors, *The Oxford Handbook of Political Netwokrs*, chapter 18, pages 453–470. Oxford University Press, Oxford, 2016.

[Kre00] Darla Kremer. Permutations with forbidden subsequences and a generalized Schröder number. *Discrete Math.*, 218(1-3):121–130, 2000.

[LCH06] Geoffrey C Layman, Thomas M Carsey, and Juliana Menasce Horowitz. Party polarization in american politics: Characteristics, causes, and consequences. *Annu. Rev. Polit. Sci.*, 9:83–110, 2006.

[LD20] Chengliang Liu and Dezhong Duan. Spatial inequality of bus transit dependence on urban streets and its relationships with socioeconomic intensities: A tale of two megacities in china. *Journal of Transport Geography*, 86:102768, 2020.

[LMDV08] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31–48, 2008.

[LR16] J. Liebig and A. Rao. Fast extraction of the backbone of projected bipartite networks to aid community detection. *Europhysics Letters*, 113(2):28003, 2016.

[MLLS21] Federico Marini, Annekathrin Ludt, Jan Linke, and Konstantin Strauch. Genetonic: an r/bioconductor package for streamlining the interpretation of rna-seq data. *bioRxiv*, 2021.

[MM13] J. Moody and P. J. Mucha. Portrait of political party polarization. *Network Science*, 1:119–121, 2013.

[MSLC01] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[NDS21a] Z. P. Neal, R. Domagalski, and B. Sagan. Analysis of spatial networks from bipartite projections using the R backbone package. *Geographical Analysis*, 2021.

[NDS21b] Zachary P. Neal, Rachel Domagalski, and Bruce Sagan. Comparing models for extracting the backbone of bipartite projections. *arXiv:2105.13396 [cs, stat]*, Jun 2021. arXiv: 2105.13396.

[NDY22] Zachary P Neal, Rachel Domagalski, and Xiaoqin Yan. Homophily in collaborations among us house representatives, 1981–2018. *Social Networks*, 68:97–106, 2022.

[Nea08] Zachary P. Neal. The duality of world cities and firms: comparing networks, hierarchies, and inequalities in the global economy. *Global Networks*, 8(1):94–115, 2008.

[Nea12] Zachary P. Neal. Structural determinism in the interlocking world city network. *Geographical Analysis*, 44(2):162–170, 2012.

[Nea13] Zachary P. Neal. Identifying statistically significant edges in one-mode projections. *Social Network Analysis and Mining*, 3(4):915–924, Dec 2013.

[Nea14] Zachary P. Neal. The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks*, 39:84–97, Oct 2014.

[Nea20] Zachary P. Neal. A sign of the times? Weak and strong polarization in the us congress, 1973–2016. *Social Networks*, 60:103–112, 2020.

[New01] Mark EJ Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1):016131, 2001.

[NG04] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[NN20] Zachary P. Neal and Jennifer W Neal. Out of bounds? The boundary specification problem for centrality in psychological networks, Aug 2020.

[NP03] Mark EJ Newman and Juyong Park. Why social networks are different from other types of networks. *Physical review E*, 68(3):036122, 2003.

[PMNW05] M. A. Porter, P. J. Mucha, M. E. Newman, and C. M. Warmbrand. A network analysis of committees in the us house of representatives. *Proceedings of the National Academy of Sciences*, 102:7057–7062, 2005.

[Poo18] Ate Poorthuis. How to draw a neighborhood? the potential of big data, regionalization, and community detection for understanding the heterogeneous nature of urban neighborhoods. *Geographical Analysis*, 50(2):182–203, 2018.

[PWK19] Vladimír Pažitka, Dariusz Wójcik, and Eric Knight. Critiquing construct validity in world city network research: Moving from office location networks to interorganizational projects in the modeling of intercity business flows. *Geographical Analysis*, 2019.

[R C18] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[RNH13] N. Ringe, Victor J. N., and Gross J. H. Keeping your friends close and your enemies closer? information networks in legislative politics. *British Journal of Political Science*, 43:601–628, 2013.

[RT] Irena Rusu and Bridget E. Tenner. Admissible pinnacle orderings. Preprint `arXiv:2001.08185`.

[Rus20] Irena Rusu. Sorting permutations with fixed Pinnacle set. *Electron. J. Combin.*, 27(3):Paper No. 3.23, 21, 2020.

[Sag20] Bruce E. Sagan. *Combinatorics: The art of counting*, volume 210 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2020.

[San00] James G Sanderson. Testing ecological patterns. *American Scientist*, 88(4):332, 2000.

[SB20] David Schoch and Ulrik Brandes. Legislators' roll-call voting behavior increasingly corresponds to intervals in the political spectrum. *Scientific Reports*, 10(1):1–9, 2020.

[SBV09] M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009.

[SCS17] Mika J Straka, Guido Caldarelli, and Fabio Saracco. Grand canonical validation of the bipartite international trade network. *Physical Review E*, 96(2):022306, 2017.

[SDCGS15] Fabio Saracco, Riccardo Di Clemente, Andrea Gabrielli, and Tiziano Squartini. Randomizing bipartite networks: the case of the world trade web. *Scientific Reports*, 5(1):1–18, 2015.

[SNB+14] Giovanni Strona, Domenico Nappo, Francesco Boccacci, Simone Fattorini, and Jesus San-Miguel-Ayanz. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature Communications*, 5:4114, Jun 2014.

[SR12] Christian Stegbauer and Alexander Rausch. How international are international congresses? *Connections*, 32(1):1–11, 2012.

[SS85] Rodica Simion and Frank W. Schmidt. Restricted permutations. *European J. Combin.*, 6(4):383–406, 1985.

[SSDC+17] Fabio Saracco, Mika J Straka, Riccardo Di Clemente, Andrea Gabrielli, Guido Caldarelli, and Tiziano Squartini. Inferring monopartite projections of bipartite networks: An entropy-based approach. *New Journal of Physics*, 19(5):053022, 2017.

[Sta97] Richard P. Stanley. *Enumerative Combinatorics. Vol. 1*, volume 49 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1997. With a foreword by Gian-Carlo Rota, Corrected reprint of the 1986 original.

[Sta99] Richard P. Stanley. *Enumerative Combinatorics. Vol. 2*, volume 62 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999. With a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin.

[Ste97] John R. Stembridge. Enriched P-partitions. *Trans. Amer. Math. Soc.*, 349(2):763–788, 1997.

[SUG18] Giovanni Strona, Werner Ulrich, and Nicholas J. Gotelli. Bi-dimensional null model analysis of presence-absence binary matrices. *Ecology*, 99(1):103–115, 2018.

[Tay01] Peter J Taylor. Specification of the world city network. *Geographical analysis*, 33(2):181–194, 2001.

[Tay04] Peter J Taylor. *World city network: a global urban analysis*. Routledge, 2004.

[TCW02] Peter J Taylor, Gilda Catalano, and David RF Walker. Measurement of the world city network. *Urban Studies*, 39(13):2367–2376, 2002.

[TML+11] Michele Tumminello, Salvatore Micciché, Fabrizio Lillo, Jyrki Piilo, and Rosario N. Mantegna. Statistically validated networks in bipartite complex systems. *PLoS One*, 6(3):e17994, Mar 2011.

[Tol21] Jeff Tollefson. Tracking QAnon: How Trump turned conspiracy-theory research upside down. *Nature*, 2021.

[USG20] USGPO. govinfo – Bulk Data - Bill Status. *United States Government Publishing Office (GPO)*, 2020.

[VFO20] Demival Vasques Filho and Dion R. J. O'Neale. Transitivity and degree assortativity explained: The bipartite structure of social networks. *Phys. Rev. E*, 101:052305, May 2020.

[VMND16] Michiel Van Meeteren, Zachary P. Neal, and Ben Derudder. Disentangling agglomeration and network externalities: A conceptual typology. *Papers in Regional Science*, 95(1):61–80, 2016.

[Wes95] Julian West. Generating trees and the Catalan and Schröder numbers. *Discrete Math.*, 146(1-3):247–262, 1995.

[Wes96] Julian West. Generating trees and forbidden subsequences. In *Proceedings of the 6th Conference on Formal Power Series and Algebraic Combinatorics (New Brunswick, NJ, 1994)*, volume 157, pages 363–374, 1996.

[XCB20] Wenna Xi, Catherine A Calder, and Christopher R Browning. Beyond activity space: Detecting communities in ecological networks. *Annals of the American Association of Geographers*, pages 1–20, 2020.

[ZH05] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[ZK11] Katharina Anna Zweig and Michael Kaufmann. A systematic approach to the one-mode projection of bipartite graphs. *Social Network Analysis and Mining*, 1(3):187–218, Jul 2011.