A mathematical foundation for the use of cliques in the exploration of data with navigation graphs

by

Pavel Shuldiner

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2023

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	William J. Martin Professor, Dept. of Mathematical Sciences Worcester Polytechnic Institute
Supervisor(s):	Richmond Wayne Oldford Professor, Dept. of Statistics and Actuarial Science University of Waterloo
Internal Member:	Steve Drekic Professor, Dept. of Statistics and Actuarial Science University of Waterloo Greg Rice Associate Professor, Dept. of Statistics and Actuarial Science University of Waterloo
Internal-External Member:	David M. Jackson Professor Emeritus, Dept. of Combinatorics & Optimization University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

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

Statement of Contributions

The work presented in this thesis is focused on the study of a method for exploratory data analysis using the package `loon` (Waddell & Oldford, 2018) based on the Ph.D. work of Waddell (2016) under the supervision of Wayne Oldford. This thesis consists in part of three manuscripts written for publication.

Unless otherwise specified, all of the results presented are novel and I am the principal author of all of the three manuscripts that encompass Chapters 3, 4, 5 and 6 of the thesis. The theorems, their proofs and discussions posed have benefitted from discussions, probing questions and guidance from my supervisor, Wayne Oldford. Exceptions to the sole authorship of material are as follows:

Research presented in Chapter 3:

The research presented in Chapter 3 is based on a submitted paper (Shuldiner & Oldford, 2021), that was written under the supervision and coauthorship of Wayne Oldford. I was the main author of the manuscript with contributions from Wayne Oldford.

Research presented in Chapter 4:

The research presented in Chapter 4 is based on a submitted paper (Shuldiner & Oldford, 2022b), that was written under the supervision and coauthorship of Wayne Oldford. I was the main author of the manuscript with contributions from Wayne Oldford.

Research presented in Chapter 5:

The research presented in Chapter 5 is based on a submitted paper (Shuldiner & Oldford, 2022a), that was written under the supervision and coauthorship of Wayne Oldford. I was the main author of the manuscript with contributions from Wayne Oldford. The proof of Theorem 5.2.4 was greatly improved by an argument illustrated to me by Wayne Oldford.

Research presented in Chapter 6:

The research presented in Chapter 6 is based on a manuscript in progress for submission. I am the main author of the manuscript and I have received input and valuable discussion from Wayne Oldford.

Research presented in Chapter 7:

The research presented in Chapter 7 is based on several projects tangential to the main work. I am the main researcher behind these findings, though I have received input and suggestions for ideas pertaining to the results of Sections 7.1.2 and 7.2.4.

Abstract

Navigation graphs were introduced by Hurley & Oldford (2011a) as a graph-theoretic framework for exploring data sets, particularly those with many variables. They allow the user to visualize one small subset of the variables and then proceed to another subset, which shares a few of the original variables, via a smooth transition. These graphs serve as both a high level overview of the dataset as well as a tool for a first-hand exploration of regions deemed interesting.

This work examines the nature of cliques in navigation graphs, both in terms of type and magnitude, and speculates as to what their significance to the underlying dataset might be. The questions answered by this body of work were motivated by the belief that the presence of cliques in navigation graphs is a potential indicator for the existence of an interesting, possibly unanticipated, relationship among some of the variables.

In this thesis we provide a detailed examination of cliques in navigation graphs, both in terms of type, size and number. The study of types of cliques informs us of the potential significance of highly connected structures to the underlying data and guides our approach for examining the possible clique sizes and counts. On the other hand, the prevalence of large clique sizes and counts is suggestive of an interesting, possibly unexpected, relationship between the variates in the data.

To address the challenges surrounding the nature of cliques in navigation graphs, we develop a framework for the derivation of closed-form expressions for the moments of count random variables in terms of their underlying indecomposable summands is established. We use this framework in conjunction with a connection between intersecting set families to obtain edge counts within a clique cover and thus, obtain closed-form expressions for the moments of clique counts in random graphs.

Acknowledgements

First and foremost, I am grateful to my supervisor, Dr. Wayne R. Oldford for his support and guidance throughout my PhD. His advice, knowledge, and enthusiasm for this project made this work possible. Thank you for taking this journey with me, Wayne.

I would like to thank my committee members Prof. Steve Drekic, Prof. Greg Rice, Prof. David M. Jackson and Prof. William J. Martin for taking the time to read my thesis and provide me with valuable feedback. A special thank you to Prof. David M. Jackson, for his unwavering encouragement and mentorship throughout my academic journey at Waterloo. His uncanny ability to recite the perfect Welsh poem to lift my spirits during difficult times was truly invaluable.

I would also like to thank my master's supervisor, Ian P. Goulden, for helping me grow as a mathematician and playing a pivotal role in initiating my academic journey.

To the friends I made during my time in Waterloo, thank you for your support and making my journey a fondly remembered one, especially Jason, Luis, Nathan, (Chris)Topher, Erik, Melissa, Mark, Nam.

I would also like to express my appreciation to Carolyn for her inspiration and motivation to aim for greater heights, both personally and in the climbing gym.

Last, but certainly not least, I extend my deepest gratitude to my family, Ilya, Irena, Mark, Noch and Sonsa, for their unconditional love and endless patience throughout my journey.

Table of Contents

List of Figures

1

Introduction

> Exploratory data analysis is detective work – numerical detective work – or counting detective work – or graphical detective work.
> A detective investigating a crime needs both tools and understanding. If he has no fingerprint powder, he will fail to find fingerprints on most surfaces. If he does not understand where the criminal is likely to have put his fingers, he will not look in the right places. Equally, the analyst needs both tools and understanding.
>
> – *John W. Tukey (Tukey, 1977)*

John Tukey argued that statisticians neglected exploratory data analysis (EDA), the exploration of data and the search for new directions of research, in favour of confirmatory data analysis. Tukey (1977) emphasized quantitative and visual techniques which reveal possibly unanticipated structure in the data. For example, the simple scatterplot can reveal various patterns in data.

For instance, consider the scatterplot in Figure 1.1 due to Kobak & Shpilkin (2021) and reproduced by *The Economist*. The scatterplot plots the percentage of voter turnout on the $x-$axis, the percentage of support for the United Russia, the current ruling government, for each polling station for every year from 2000 to 2021. It reveals the presence of a monotonic relationship between voter turnout and support for the political party United Russia, several clusters (top right, center and bottom left) and a surprising granularity.

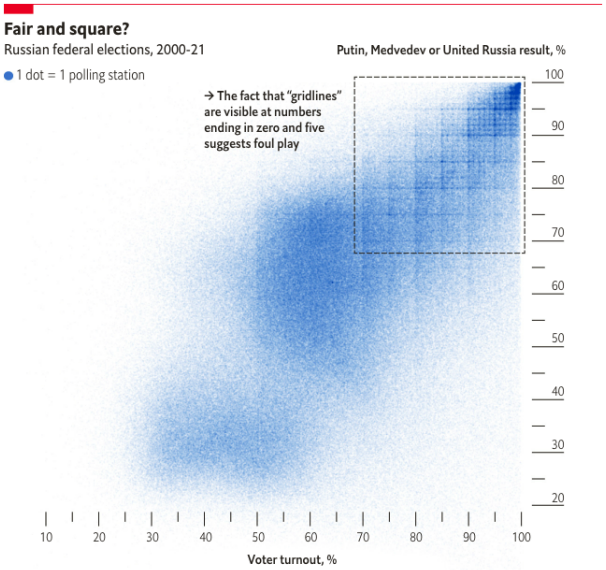


Figure 1.1: Each point stands for a polling station among the 96,325 stations.The raw data was scraped by Kobak & Shpilkin (2021) from the official Russian polling websites maintaining the voter turnout and candidate totals for various regions. Thus, outside of categorical variables, such as region identifiers, their scraped data contains only rows with integer values and hence, decimal rounding does not explain the prevalence of integer peaks in the scatterplot.

These grid lines observed as multiples of five arose suspicions of electoral fraud, hence generating an interesting hypothesis to be tested: how likely is such an unusual high number percentage of voter-turnouts and share to be this tidy? In examining this data, the analysts had suspicions of election fraud based on the granularity of election results (namely, the integer peaks present in polling data) on previous work (Kobak et al., 2016), and thus it was an obvious choice to examine the distribution of percentage voter turnout and percentage of voter turnout in their dataset.

1.0.1 Scagnostics

After cleaning, Kobak & Shpilkin's (2021) data consisted of only three non-categorical variates: number of voters turned out, the number of eligible voters in a region and the percentage of support for United Russia. In addition to their subject matter expertise, the small number of variables made it manageable to explore the relationships present in the data. In practice, even a modest number of variables can significantly tax the analyst's time.

For instance, consider Hofert & Oldford's (2020a) dataset `de_elect`, consisting of n[1] =68 variables from German elections in 2002 and 2005. Outside of `District` and `State`, all other 66 variables are numeric. This results in $\binom{66}{2} = 2145$ scatterplots that can be examined.

[1]Traditionally, the letter p represents the number of variables in a dataset. However, in the interest of being consistent with the established mathematical notation from Johnson graphs, we use the letter n.

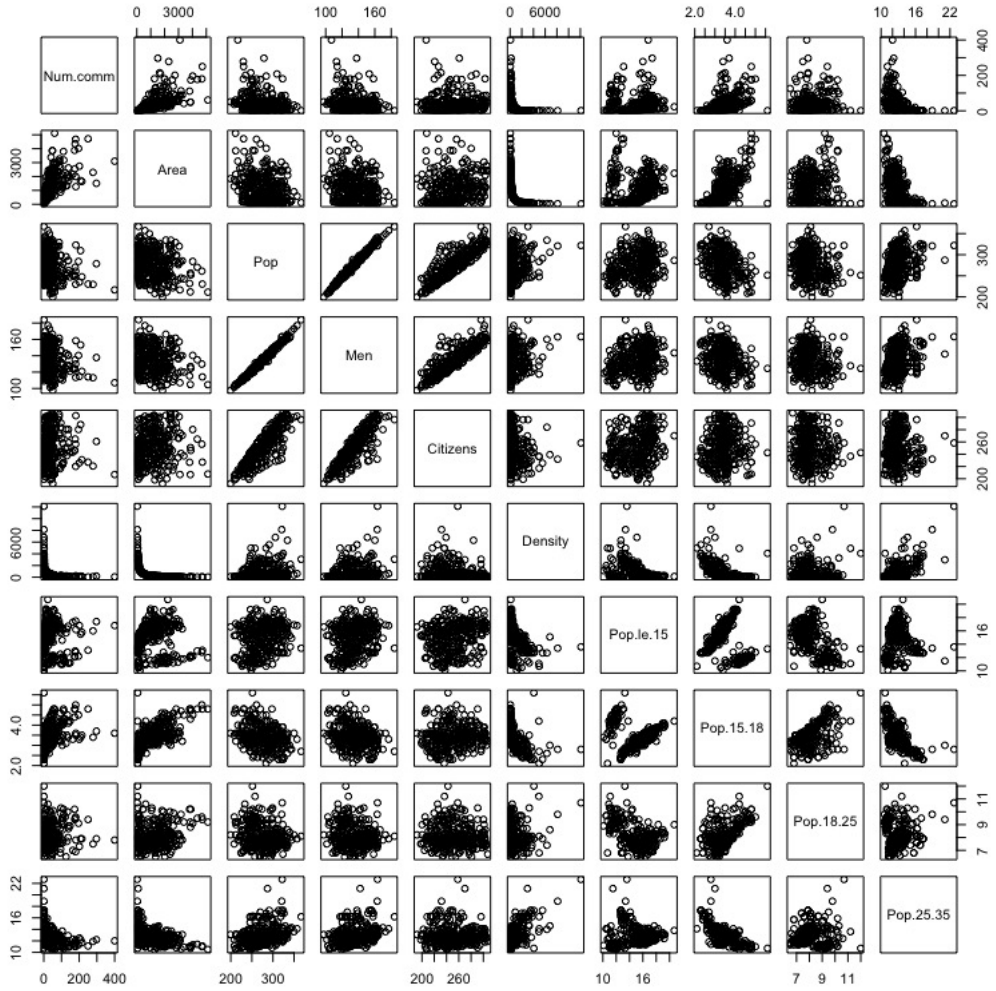


Figure 1.2: A scatterplot matrix of the first 10 numeric variables in the `de_elect` dataset. The $i-$th row and $j-$th column entry plots the scatterplot of the $i-$th variable and versus the j-th variable in the data.

Even when focusing on only a small number of $n = 10$ of the 66 possible numeric variables, there are $\binom{10}{2} = 45$ scatterplots to examine for patterns. As n increases, the task of closely investigating each of the $\binom{n}{2}$ scatterplots becomes quickly intractable. It is therefore advantageous for the analyst to carefully choose the scatterplots to be explored. Thus, there is a need for tools that not only aid the analyst in exploring their data, but also do so in a manner that respects the analyst's time.

Tukey & Tukey (1985) suggested culling the number of scatterplots examined by only focusing on those with the most extreme scores according to some precomputed measures of interest, so-called scatterplot diagnostic measures (scagnostics). Scagnostics assign a quantity between 0 and 1 to a scatterplot based on how strongly it exhibits a particular property of interest.

For instance, consider the stringyness of a scatterplot – the tendency of a scatterplot to resemble a string. This measure can be captured by embedding a dataset into the Euclidean plane, computing its minimum spanning tree T and evaluating the ratio

$$\frac{diameter(T)}{sum(T)},$$

where $diameter(T)$ is the length of the longest shortest path between two nodes and the denominator is the sum of all of the edge weights of T. If this value is approximately 1, then the longest path between two nodes approximately travels through all vertices in the minimum spanning tree and hence there are very few branches – so T has a string-like shape.

The stringy scagnostic is among a collection of graph-theoretic measures developed by Wilkinson et al. (2005) in their extension to Tukey & Tukey's (1985) work. A complete list of scatterplot attributes and measurements examined and developed by Wilkinson et al. (2005) is given as follows. Shape can be assessed via clumpy, skewed, sparse, striated, convex, skinny, and stringy. Additionally, the presence of many outliers can be measured via outlying, and the trend can be measured through monotonicity, the square of the Pearson correlation coefficient of the ranks of the two variables.

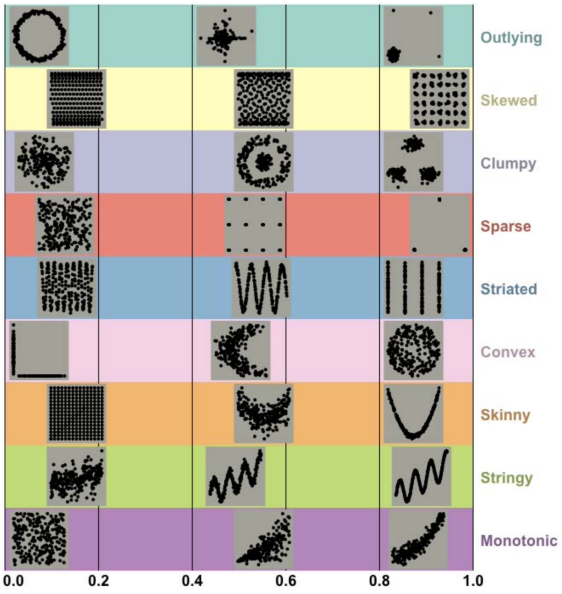
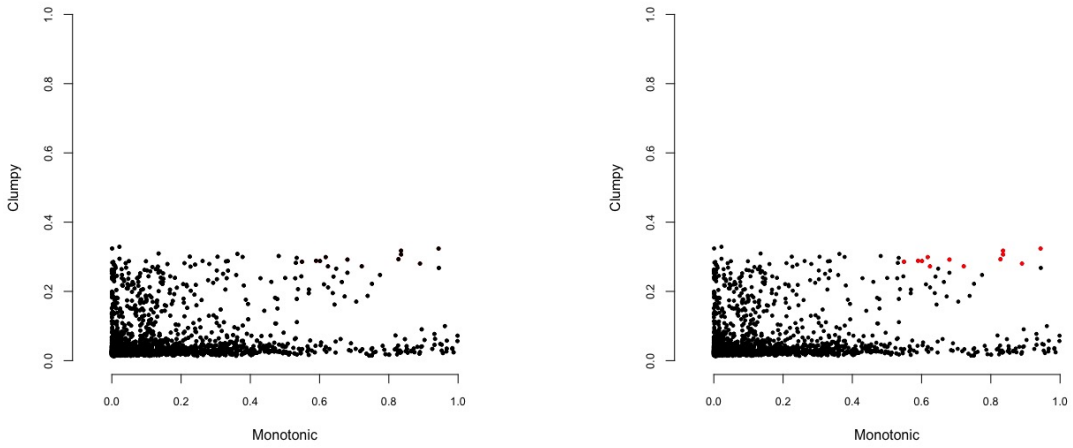


Figure 1.3: Examples of scatterplots and their scoring on the nine scagnostics measures (Dang & Wilkinson, 2014).

Using scagnostics, Tukey & Tukey (1985), sought to detect anomalies in the density, shape and trends of scatterplots. In the Tukeys' design, after evaluating these measures, a scatterplot matrix of the measures would be constructed. According to Wilkinson et al. (Section 2 2005), Paul Tukey suggested viewing the scagnostic scatterplot matrices as a display of pointers (links to scatterplots), which can be assessed to identify irregularities



(a) Monotonic versus clumpyness of all $\binom{66}{2} = 2145$ scatterplots.

(b) The top 10 percentile of monotonic and top 2 percentile of clumpy scatterplots.

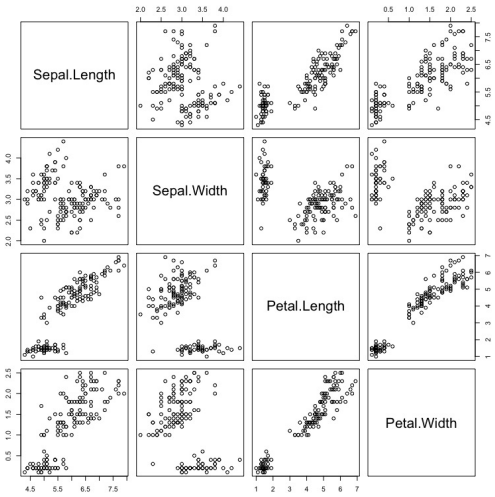
Figure 1.4: Scagnostic measures of monotonic (x-axis) and clumpyness (y-axis) of the 2004 German election dataset `de_elect` from Hofert & Oldford (2020b). The dataset contains 66 numeric variables on 299 observations from the 2004 German election. The data has no scatterplots scoring high on clumpyness.

among the scatterplots and hence the variables. With interactive data visualization software, such as `loon` (Waddell, 2016; Waddell & Oldford, 2018), anomalous scatterplots could be interactively identified from the scagnostics scatterplot matrix. This reduces the problem of examining $\binom{n}{2}$ scatterplots to examining $\binom{k}{2}$ scatterplots, where k is the number of scagnostic measures of interest to the analyst, and then examining only the most interesting plots.

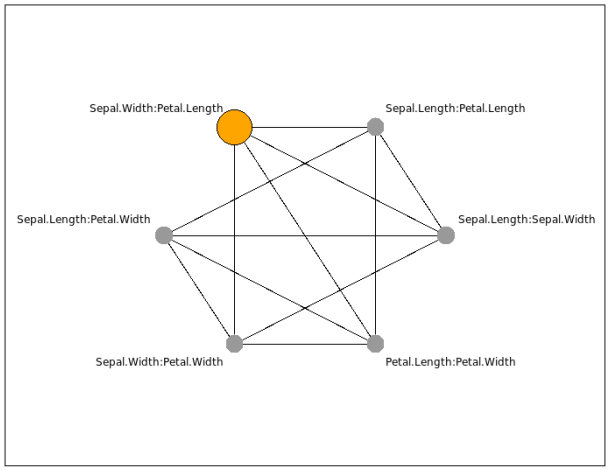
Work on scagnostics has been extended in several ways, such as the development of measures that apply to three-dimensional scatterplots (Fu, 2009) and the implementation of scagnostics-like measures to time series data (Dang et al., 2012). Another interesting direction for extension lies with providing the analyst's with a bird's eye view of the data, indicating which variables are driving the relationships present in the most interesting scatterplots. The navigation graph framework introduced by Hurley & Oldford (2011a) serves as a natural candidate for a graph-theoretic representation of these relationships.

1.1 Navigation graphs

Let $\mathcal{V}$ denote the set of all variables in a dataset, and $n = |\mathcal{V}|$ and $m, k \in \mathbb{N}$ be fixed with $n \geq m \geq k$. The node set of a navigation graph consists of $m-$subsets of the n variables in the dataset along with an additional attribute, a visualization of the corresponding $m-$dimensional space. Two nodes are adjacent in a navigation graph if they share k variables together.



(a) Scatterplot matrix of the `iris` dataset.



(b) `iris` navigation graph.

Figure 1.5: The scatterplots and navigation graph of the `iris` dataset. The navigation graph consists of transitions in 3 dimensional space – there are three variables corresponding to a union of two adjacent nodes and two variables on every node.

For instance, consider `iris` dataset, first popularized by Fisher (1936), consists of 150 observations on $n = 5$ variates (`Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width` and `Species`). Figure 1.5b illustrates the corresponding navigation graph.

Waddell & Oldford's (2018) implementation of navigation graphs aids the analyst by

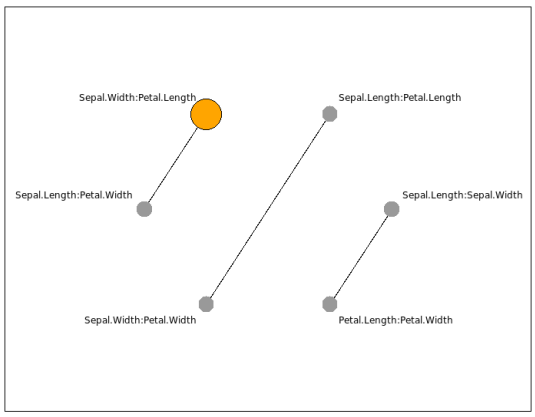
1. facilitating the visualization of a projection onto a subset of the variables and their smooth transition to another subset; and

2. providing the analyst with a high level overview of their dataset.

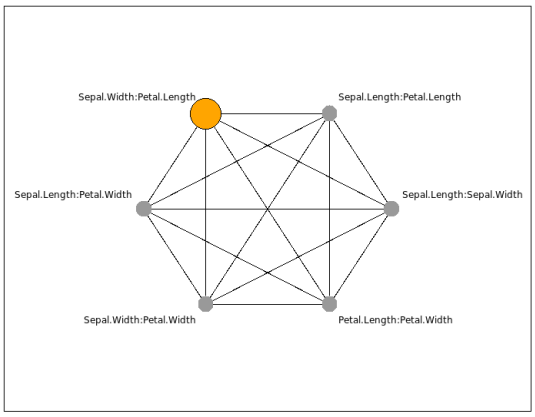
The benefit in visualizations created through smooth transitions is in allowing the analyst to grasp a higher dimensional space than any of the separate projections on their own (Buja & Asimov, 1986).

For instance, transitioning smoothly from one 2-dimensional space to another results in a 3-dimensional movie that may reveal a relationship between the variables as points in the scatterplot shift from one projection to the other. Since the size of intersection of adjacent spaces is controlled by a parameter k, one can also examine higher dimensional projection transformations. Figure 1.6 illustrates two of the other possible navigation graphs on the `iris` dataset: 4d transition navigation and the union of the 3d and 4d transition navigation graphs.

The advantage of a high level overview of the data can be realized through the examination of the data via graph theoretic means. This is the focus point of the majority of the research in this thesis.



(a) `iris` 4d transition navigation graph.



(b) Union of the `iris` 3d and 4d transition navigation graphs.

Figure 1.6: Two spaces are adjacent in the navigation graph if they share no variables in common or at most 2 variables in common, respectively. These two graphs are instances of the so-called generalized Johnson family of graphs.

In spite of the advantages of navigation graphs, there remains a challenge similar to the one addressed by the Tukeys via scagnostics: as n and m increase, the number of spaces to be visited and the resulting navigation graph becomes unmanageably large. Therefore, it is imperative to cull the navigation graph, and the spaces examined, to only the most interesting ones.

1.1.1 Subgraphs of navigation graphs

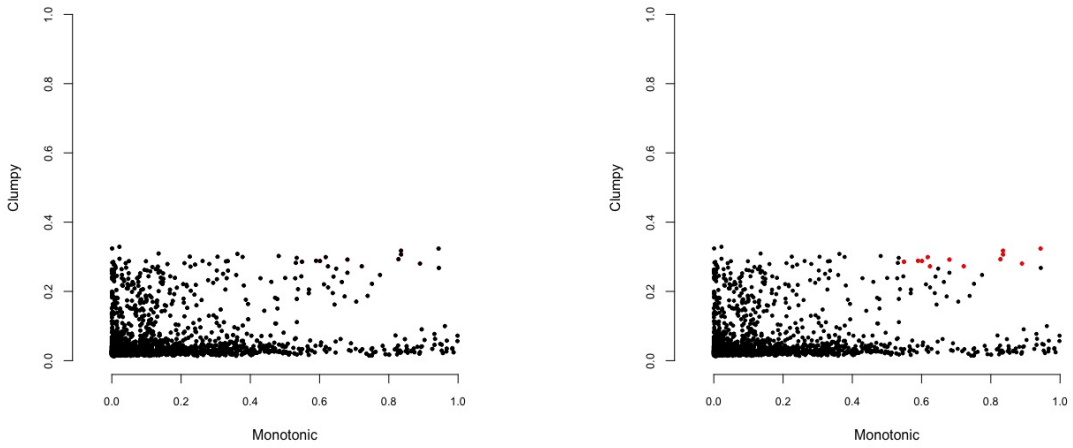
While the analyst may begin by constructing their navigation graph with all $\binom{n}{m}$ projections, and hence a node set of size $\binom{n}{m}$, a large proportion of spaces can be excluded from further investigation if they are deemed uninteresting to the analyst. As a result, we present a framework under which subgraphs of navigation graphs could be generated.

Let $w : \binom{\mathcal{V}}{2} \to \mathbb{R}$ be a function quantifying the 'interestingness' of a relationship between an unordered pair of variables such that more peculiar spaces get mapped to larger values of w, and R is a subset of $\mathbb{R}$ the set of all real numbers. There are two mechanisms we consider for the generation of navigation subgraphs. These are based on a fixed cutoff value and the empirical distribution of w:

M1 Fixing a cutoff value t and choosing the induced subgraph where consisting of all 2-subsets $\{X, Y\}$ for which $w(\{X, Y\}) > t$.

M2 Fixing a proportion q and picking the induced subgraph with node set consisting of all subspaces in the top $q-$th percentile according of the empirical distribution of w.

In contrast with M2, M1 relies on the analyst having an understanding of the underlying distribution of w. Regardless of the mechanism chosen, both paradigms produc e a graph consisting of the most interesting spaces based on the analyst's criterion of choice.



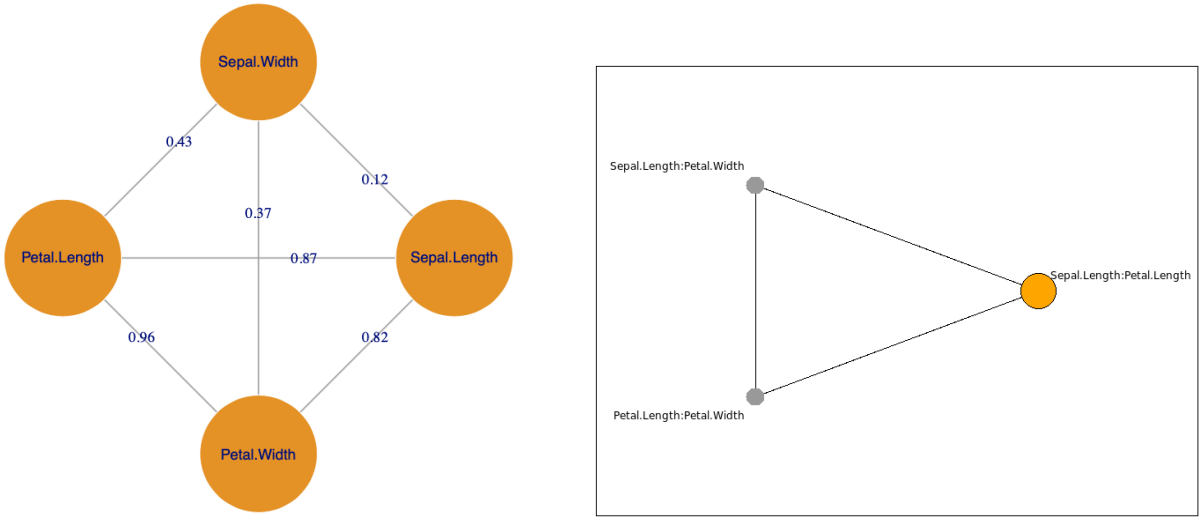
(a) Monotonic versus clumpyness of all $\binom{66}{2} = 2145$ scatterplots.

(b) The top 10 percentile of monotonic and top 2 percentile of clumpy scatterplots.

Figure 1.7: Scagnostic measures of monotonic (x-axis) and clumpyness (y-axis) of the 2004 German election dataset `de_elect` from Hofert & Oldford (2020b). The dataset contains 66 numeric variables on 299 observations from the 2004 German election. The data has no scatterplots scoring high on clumpyness.

An instance of the navigation graph produced via Mechanism M1 is given by Figure 1.8. Starting with the variable graph, the graph whose node set consists of all variables in the dataset, and filtering for edges scoring above a fixed threshold t according to a scagnostic of interest, we obtain a subgraph of the complete graph. Then, applying a line graph operator (Section 2.1.1), we obtain the subgraph of the navigation graph corresponding to Mechanism M1.

An example of Mechanism M2 is in Figure 1.9, which depicts a navigation graph resulting by filtering for the most interesting spaces according to the empirical distribution of scagnostic measures of interest. Under certain assumptions, Mechanism M1 allows us to examine the resulting navigation graph using a class of random graphs (Section 2.2). Moreover, regardless of the mechanism used to generate subgraphs from a navigation graph, we shall see that all subgraphs are isomorphic to subgraphs of a well-known family of graphs: the generalized Johnson graphs (Godsil & Royle, 2001).



(a) Variable graph with respect to the correlation measure of interest.

(b) Corresponding navigation subgraph with $t = 0.5$ according to Mechanism M1.

Figure 1.8: The variable graph of the `iris` data with correlation measure of interest. The edges were pruned at a cutoff of 0.5 and the corresponding navigation subgraph was generated by the line graph operator.

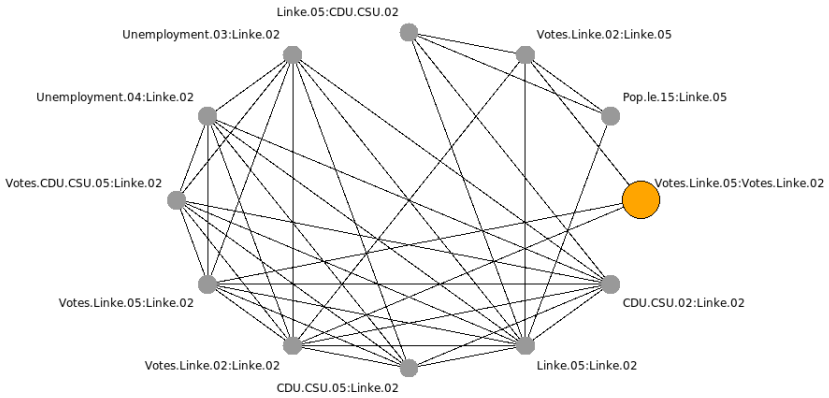


Figure 1.9: The navigation graph of the most monotonic and clumpy variates according to Figure 1.7. Note that `Linke.02` appears in the majority of the scatterplots.

We note that under some assumptions regarding the weight function w, the graph resulting from the first construction is a realization of a line graph of a random graph. Moreover, both approaches result in graphs are the induced subgraphs of a well-known family of graphs known as the Johnson graphs (Godsil & Royle, 2001). Therefore, understanding the graph structure of Johnson graphs is essential for our understanding of the 'most interesting' subgraphs of navigation graphs. While there are many notions of structure that can be interesting to examine in these graphs, the emphasis of the present work is to study subgraphs that exhibit maximal cohesiveness – cliques.

1.2 Cliques

A clique in a graph G is a set of nodes H such that the induced graph $G[H]$ forms a complete graph. Research on complete subgraphs has roots in the graph-theoretic reformulation of Ramsey theory (Erdős & Szekeres, 1935), a branch of mathematics that extends the notion of the pigeon-hole principle: no matter how one partitions a 'large' structure into smaller substructures, one of the substructures will contain a 'large' structure. However, it was not until 1949 that interest in cliques surged when Luce & Perry (1949) popularized the term 'clique' and identified its relevance in sociology: cliques are the quintessential cohesive structure where node connectivity is maximized.

Cliques encapsulate the notion that all of the elements within a group are connected or similar. Luce & Perry (1949) used matrix analysis techniques to examine highly connected group structures in subgraphs of a network. Moreover, they were the first to introduce the clique problem: the computational problem of finding cliques in a graph. In 1957, Harary & Ross (1957) presented the first documented solution to the clique problem. Their work led to several interesting generalizations of the concept of cohesive structure, such as the so-called $n-$clans (as will be discussed in Chapter 7). Furthermore, the study of cliques influenced scientific disciplines outside of graph theory and the social sciences.

Today, researchers examine cliques and their variants in numerous other disciplines, such as neuroscience and computational biology. Community detection algorithms, which are relaxations of clique finding algorithms, are used to study the organization of brain networks (Ashourvan et al., 2019). For instance, segregation in brain networks has been examined (Sporns, 2013), (Stam & Reijneveld, 2007) via the network's modularity: the tendency of a network to organize its nodes into cliques. Additionally, a fundamental measure of segregation on these networks is the clustering coefficient, which measures the connectivity density among nodes and their nearest neighbours. The higher the density, the more likely they will form a cluster or a clique. He et al. (2008) found that Alzheimer's disease patients' brain networks displayed high local clustering and larger shortest path linking individual regions than the control group – participants without Alzheimer's disease. The loss of efficiency in communication between distant brain regions was later examined by studying the average shortest path distance between all pairs of nodes in a network (Lo et al., 2010). This value is smaller when there is a more prominent global clustering coefficient, and hence losses in cognitive function can be explained by the clique topology of the brain network (Yao et al., 2010).

Three of the main problems concerning the cliques are (Bomze et al., 1999):

1. The maximum clique enumeration problem - listing all maximum cliques in a graph (Jain & Seshadhri, 2020; Östergård, 2002);

2. The maximal clique enumeration problem - listing all maximal cliques (Ouyang et al., 1997);

3. The maximum clique optimization problem - identifying the size of the largest clique, i.e. the clique number (Pardalos & Xue, 1994).

These three problems are in general hard – for instance, the maximum clique optimization problem is one of the first problems to be shown to be NP-hard (Gross & Yellen, 2003, Section 5.3). Nonetheless, in this dissertation, we present solutions to these problems in the context of Johnson graphs. These insights aid us in interpreting the possible significance behind the appearance of certain cliques as well as indicate to us which clique configurations are feasible in navigation graphs.

1.3 Cliques in navigation graphs

Since the clique is a prototype of community-like structure in graphs, it is a natural question to ask: what does the presence of cliques suggest about the variables of the underlying navigation graph? In other words, what do communities look like in navigation graphs?

For instance, consider the two cliques from Figure 1.9, one consisting of the many nodes sharing the variable `Linke.02` and a smaller one, consisting of the four spaces sharing the variable `Linke.05`.

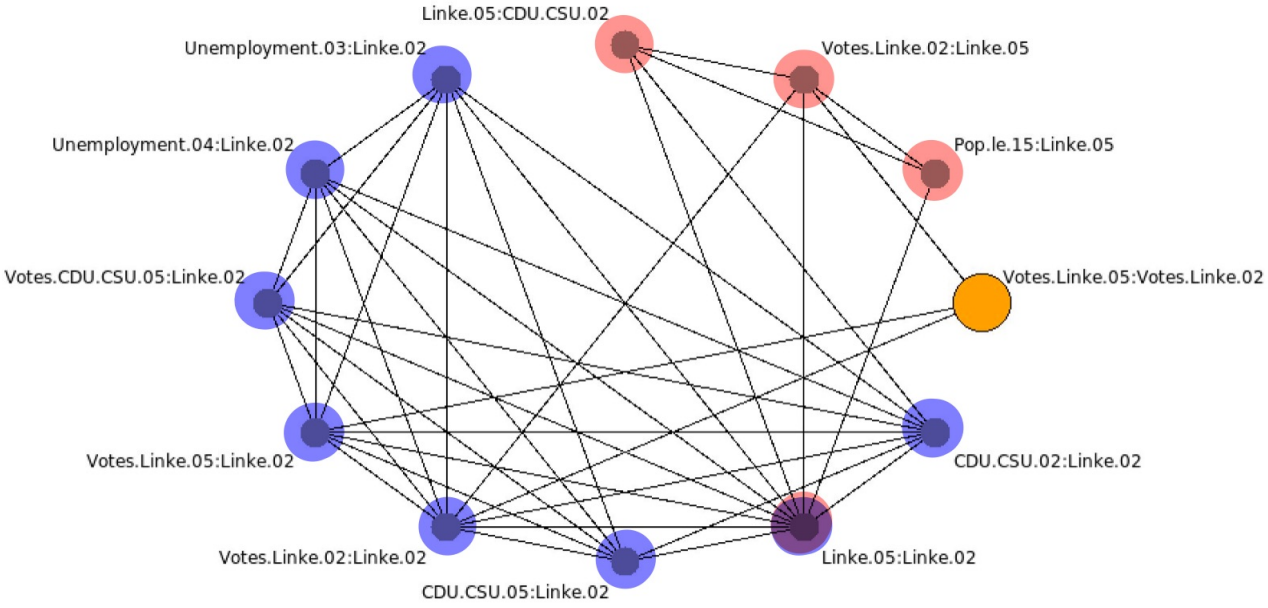


Figure 1.10: Blue nodes correspond to a large clique surrounding `Linke.02`. Red nodes correspond to a small clique surrounding `Linke.05`.

Imagine a clique in a navigation graph: each node has m variables and every pair of nodes share k variables. Of interest would be, which variables appear in the clique (what is the clique's union)? Also of interest, of these variables, which of any appear in every node (what is the clique's intersection)? What is the size of the clique? What configurations of variables can appear together in a cohesive structure? What is the maximum possible size of a clique?

In navigation graphs, cohesive, community-like structures could suggest the presence of interesting relationships between the underlying variables. This thesis focuses on developing the mathematical foundations for a clique-centric study of navigation graphs. Borrowing Tukey's analogy, we believe the presence of cliques, both in terms of magnitude and quantity, will serve as fingerprint powder.

1.4 Overview

The aim of this work is to exploit the connection between navigation graphs, random graphs and Johnson graphs to examine the possible significance behind cliques in navigation graphs. The main two contributions of this work towards the clique-centric study of navigation graphs can be summarized in the following two results:

Theorem 1.4.1. *Let $\mathcal{V}$ denote the set of all n underlying distributions of random variables in a dataset. Let G be the complete variable graph obtained from model **M1** under the assumptions **A1** and **A2**, where the cutoff value is chosen so that $Pr(F > t) = p$. Let H be the corresponding navigation subgraph. Let C_r, X_r, Z_r be the random variables where*

1. *C_r is recording the number of maximal $r-$cliques in H,*
2. *X_r is recording the number of $r-$cliques in $G(n,p)$, and*
3. *Z_r is recording the number of nodes with degree exactly r in G.*

Then the moments of C_r are given by

$$E(C_r^k) = \begin{cases} E(Z_r^k), & r \geq 4 \\ E((Z_3 + X_3)^k), & r = 3 \end{cases},$$

where

$$E(X_3^k) = \sum_{m=1}^{k} \sum_{\{i_1,\cdots,i_m\} \subseteq \mathcal{I}_3} S(k,m) p^{e(i_1,\ldots,i_m)},$$

$i_1, \ldots, i_m$ are distinct triangles in $\mathcal{I}_3$ the set of all 3-subsets of $[n]$, and

$$E(Z_\ell^k) = \sum_{m=1}^{k} \binom{k}{m} S(k,m) \sum_{\mathbf{e} \in \mathcal{E}_m} p^{|\mathbf{e}|} (1-p)^{\binom{m}{2} - |\mathbf{e}|} \times \prod_{i=1}^{m} \binom{n-m}{\ell - \sum_{j \in [n] \setminus \{i\}} e_{ij}} p^{\ell - \sum_{j \in [n] \setminus \{i\}} e_{ij}} (1-p)^{n-m-\ell+\sum_{j \in [n] \setminus i} e_{ij}},$$

where $\mathcal{E}_n$ denotes the set of all vector combinations of subgraphs of the complete graph $\{\mathbf{e} : \mathbf{e} = (e_{ij})_{\{i,j\} \subset [n]}, e_{ij} \in \{0,1\}\}$.

The following theorem states that there are only two types of non-trivial cliques in Johnson graphs.

Theorem 1.4.2. *Let $\mathcal{V}$ denote the set of all n random variables observed in a dataset, let H be its navigation graph where $m = k+1$ and fix an integer $r \geq 3$. If C is an $r-$clique in H, then C has either intersection of cardinality k or union of cardinality m but not both.*

At their core, Theorems 1.4.1 and 1.4.2 exploit the fact that a navigation graph is the line graph of a variable graph. Since cliques arise only as stars and triangles in this configuration, there are only two types of cliques that we may encounter and they are uniquely captured through the distributions of stars and triangles in a random graph.

More broadly, we use and develop algebraic combinatorics tools that specialize to derive closed-form expressions for the moments of clique counts in random graphs (and hence, navigation graphs under certain assumptions), describe the number of cliques induced by a clique cover and demonstrate that Johnson graphs, and hence navigation graphs where $m = k + 1$, have only two types of cliques.

Chapter 2 provides a brief introduction to the necessary mathematical background required for the remaining chapters. Connections between navigation graphs, Johnson graphs and random graphs are discussed. The challenge with the problem of computing the distributions for clique counts and sizes is discussed and an approximate solution is discussed.

Chapter 3 introduces Bernoulli sums, a framework for studying count random variables – random variables with support on the natural numbers. The framework captures the relationship between the moments of a count random variable and the joint distribution of its underlying indecomposable parts.

To apply the Bernoulli sums theory to clique counts in random graphs, one needs to derive the number of edges present in a collection of cliques. Chapter 4 solves the problem of identifying how many cliques are induced by a clique cover. A connection between graphs and intersecting families of sets is established through an orbit partition related to the clique cover. This relationship leads to closed-form expressions for the number of cliques of any size contained within the collection.

While the theory established in Chapters 3 and 4 addresses the questions of how many cliques typically appear in a navigation graph, Chapter 5 investigates the possible meaning behind large clique. In particular, we show that there are only two types of cliques on Johnson graphs, prove that almost all cliques are of a certain type as n grows asymptotically, derive the size of a maximal clique, identify the clique partition number and enumerate the clique counts.

Chapter 6 interprets the results of the previous chapters in the context of several models of navigation graphs. We prove Theorems 1.4.1 and 1.4.2. A discussion of the limitations and implications follows.

Chapter 7 introduces problems related and describes some of our progress in attacking them. The chapter concludes with a reflection of the body of work presented here.

2

Preliminaries

We begin with the necessary mathematical background required for the results in the following chapters. In Section 2.1, graph theoretic terminology is reviewed. We introduce line graph operators, describe Whitney's isomorphism theorem and it's exceptions (Theorem 2.1.3) and describe a construction for Johnson graphs. Section 2.2 introduces random graphs, their univariate degree distributions and a background on the history of the clique counting problem on random graphs. Section 2.3 describes the terms and basic theory of algebraic combinatorics we use in our results. In particular, the multinomial theorem and principle of inclusion and exclusion are revisited.

2.1 Graph theory

Throughout this thesis, we let $[n]$ denote the set of the first n natural numbers $\{1, 2, \ldots, n\}$. The set of all natural numbers $\{1, 2, \ldots\}$ is denoted by $\mathbb{N}$, and the set of all nonnegative integers is denoted by $\mathbb{N}_0$. Given a set A, the set containing all subsets of A will be denoted by $\mathcal{P}(A)$.

A *graph* G is an ordered tuple of sets (V, E) where E is a subset of $V \times V$. In this thesis, any graph G is assumed to be a *simple graph* which means that E can be viewed as a set of $2-$subsets of V. If G is a graph, we will denote by $V(G), E(G)$ the set of nodes/vertices and edges of G, respectively.

For two vertices $u, v \in V$, we call u and v *neighbours* or *adjacent* if $e = \{x, y\} \in E$. In such scenario, we say the vertices u and v are *incident* with e. If u is adjacent to exactly k vertices, then we say that u has *degree* k.

If e and f are edges in E and e and f share exactly one vertex in common, we say that e and f are *incident*.

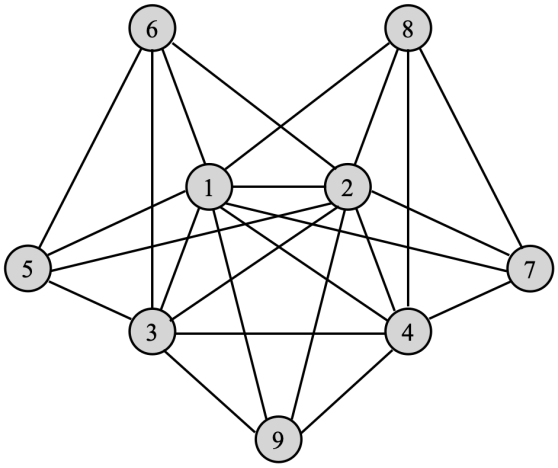


Figure 2.1: A simple graph on 9 nodes.

We call $G' = (V', E')$ a *subgraph* of $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$. Let $U \subseteq V$ be a subset of the set of vertices of V. We call the graph $G[U]$ with the nodeset U the subgraph *induced* by U if every edge between nodes of U in G is present in $G[U]$.

We say that two graphs $G = (V, E)$ and $H = (V', E')$ are *isomorphic* if there is a bijection $f : V \to V'$ for which u, v are adjacent in G if and only if $f(u), f(v)$ are adjacent in H. In such case we write $G \simeq H$.

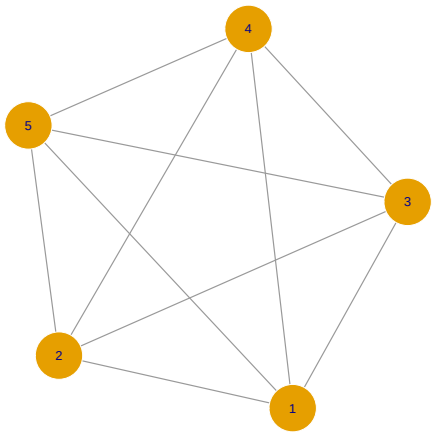


Figure 2.2: The complete graph on 5 vertices, K_5.

Example 2.1.1. The *complete graph* on n vertices is the graph where every two distinct vertices are adjacent. In this thesis, we will denote this graph by K_n, where n is the number of vertices of the graph.

An important graph related to the complete graph we examine in the construction of navigation subgraphs (Section 1.1.1) is the *variable graph*. The variable graph is a weighted complete graph, where edges are weighted according to a measure of interest $w : \binom{\mathcal{V}}{2} \to \mathbb{R}$, such as a scagnostic (Section 1.0.1). Throughout this work, a trimmed or a pruned variable graph is one with edge weights above a certain threshold (as described by Mechanisms M1 and M2, for example).

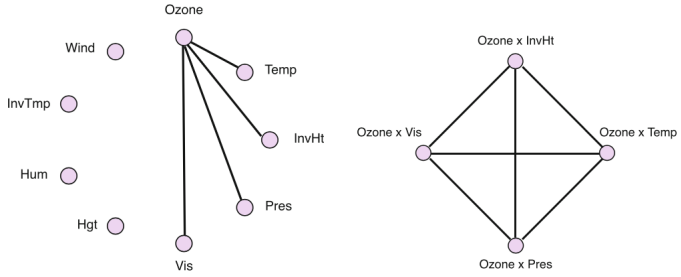


Figure 2.3: A realization of a pruned variable graph of the `Ozone` dataset (Breiman & Friedman, 1985) recreated by Hurley & Oldford (2011a). The variables `Ozone`, `Temp`, `InvHt`, `Pres` and `Vis` score highly according to some scagnostic measure of interest. The figure on the right is the corresponding navigation graph.

If G is a graph and H is a subgraph of G for which $H \simeq K_n$ for some $n \in \mathbb{N}$, then we say that H is a *clique* in G. If H is not a proper subgraph of a larger clique in G, then we say that H is a *maximal clique*. We call H a *maximum clique* if no other clique in G has more vertices than H. We note that any maximum clique is maximal but not every maximal clique is a maximum clique.

If H is a maximum clique in G and H has ℓ vertices, then we say that the *clique number* of G is ℓ and write $\omega(G) = \ell$.

Given a collection $\mathcal{C} = \{C_1, \ldots, C_m\}$ of cliques, we say that $\mathcal{C}$ is a *vertex clique cover* of the graph G if every C_i is a subgraph of G and every node in G belongs to some clique C_i in $\mathcal{C}$. Moreover, if $\mathcal{C}$ is a vertex clique cover for which every edge $e \in E(G)$ appears in some clique C_r in $\mathcal{C}$, then we say that $\mathcal{C}$ is an *edge clique cover* of G.

For example, consider the simple graph G from Figure 2.1. The cliques A, B, C where $A = \{1, 2, 3, 5, 6\}, B = \{1, 2, 4, 7, 8\}, C := \{1, 2, 3, 4, 9\}$ form a vertex clique cover and an edge clique cover of G. Of course, there are other vertex and edge clique covers of G – for instance, consider the edge clique cover formed by using the edges. We will see in Chapter 6 how the problem of evaluating moments of clique count distributions on random graphs is related to edge clique covers.

If every vertex in a graph G has the same degree, say k, we call the graph G a $k-$*regular graph*. For instance, the complete graph K_n is $(n-1)-$regular as every vertex is adjacent to all other vertices.

Example 2.1.2. Graphs are frequently used to model relations between discrete objects in the real world. When discussing a real world phenomena, some researchers use the term *network* to emphasize that a graph's vertices and edges stand for real world objects, such as professors in a department and their collaboration count, respectively.

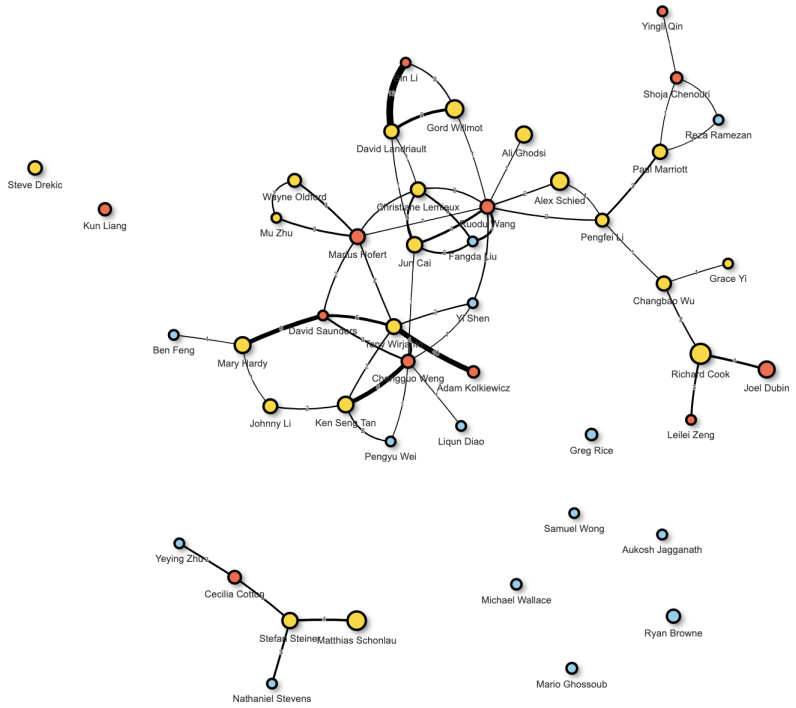


Figure 2.4: A graph where the vertices are professors in the UWaterloo Statistical and Actuarial Science department and edges indicate collaboration within the last 5 years Ogyanova (2020).

A *walk* in a graph G is a sequence of edges $(e_1, e_2, \ldots, e_\ell)$ such that there exist vertices $v_1, v_2, \ldots, v_{\ell+1}$ with the property that edge e_i connects nodes v_i and v_{i+1} in G. If all of the vertices corresponding to the walk are distinct, this walk is known as a *path*. If all of the vertices except the first and the last are distinct, we call the walk a *cycle*. Clearly, up to isomorphism there is only one cycle on n vertices for $n \geq 3$.

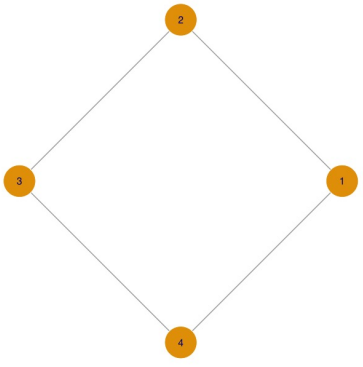


Figure 2.5: A cycle on 4 vertices.

When nodes on a graph have an additional structure (for instance, as sets), we will distinguish between nodes on a graph and their inherent structure through the use of the $\nu(\cdot)$ notation. For instance, if G is the complete graph on $[n]$, then although the intersection of vertices is has no meaning in our context, the intersection of the labels of vertices will be of relevance. In other words, if $a, b \in V$, we will identify a and b with $\nu(a)$ and $\nu(b)$, respectively, and let their intersection be denoted by $\nu(a) \cap \nu(b)$.

2.1.1 Graph operations

There are many operations defined on graphs. In the following chapters, we use graph unions, intersections and line graph operators which we recall are defined as follows.

Let $G_1 = (V_1, E_1), G_2 = (V_2, E_2)$ be two graphs. Then the *graph union* and *graph intersection* of G_1, G_2 are given by $G_1 \cup G_2 = (V_1 \cup V_2, E_1 \cup E_2)$ and $G_1 \cap G_2 = (V_1 \cap V_2, E_1 \cap E_2)$, respectively.

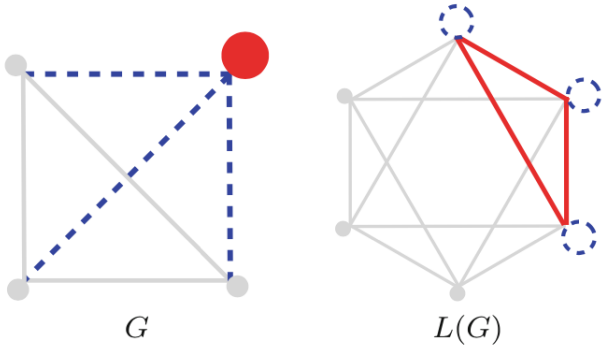


Figure 2.6: The dashed blue edges become nodes in $L(G)$ and they are adjacent because they are incident in G. This figure has been reproduced from Oldford & Waddell (2011) with permission.

In graph theory, the *line graph* of a graph $G = (V, E)$ is the graph $L(G) = (V', E')$ that is constructed in the following way: each $v \in V'$ corresponds to an edge e in E; two vertices $u, v \in V'$ are adjacent in $L(G)$ if they share exactly one vertex in G.

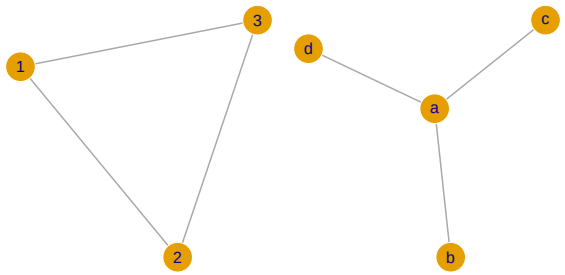


Figure 2.7: The exceptions to Whitney's isomorphism result. Left: The complete graph K_3 on three nodes. Right: The star graph $K_{1,3}$ on degree 3.

As will be discussed in Chapter 6, the line graph operator bridges random graphs and navigation graphs under suitable assumptions. To examine cliques in navigation graphs under this model, we need to understand how cliques arise in random graphs and when the line graph operator translates subgraphs into cliques. Whitney's isomorphism theorem states that the action of the line graph operator on connected graphs is injective except for two graphs: the triangle graph K_3 and the star $K_{1,3}$.

Theorem 2.1.3 (Whitney graph isomorphism theorem)**.** *Let G_1 and G_2 be two connected graphs not equal to the triangle K_3 or the star $K_{1,3}$. Then G_1 and G_2 are isomorphic if and only if $L(G_1)$ and $L(G_2)$ are isomorphic.*

Proof. See Whitney (1992). □

It is easy to check that $L(K_3) = K_3 = L(K_{1,3})$. Moreover, by Theorem 2.1.3, it follows that cliques of size $r \geq 4$ are the images of stars $K_{1,r}$ under the line graph operator.

Example 2.1.4. Consider the complete graph K_5 with node set $V(K_5) = [5] = \{1, 2, 3, 4, 5\}$, and edge set

$$E(K_5) = \{\{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{2,3\}, \{2,4\}, \{2,5\}, \{3,4\}, \{3,5\}, \{4,5\}\}.$$

By construction, the node set of $L(K_5)$ is the edge set of K_5 and hence $V(L(K_5)) = E(K_5)$ and two nodes u_1, u_2 are adjacent if and only if u_1 and u_2 share a node in K_5, which is equivalent to

$$|\nu(u_1) \cap \nu(u_2)| = 1.$$

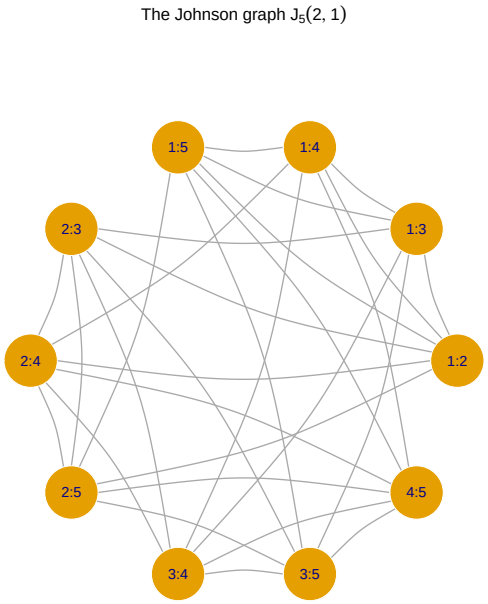


Figure 2.8: The line graph of K_5.

There are many other graph theoretic operations on graphs and the interested reader is encouraged to consult Bondy & Murty (2008). We are interested in the line graphs of complete graphs as they form a special case of a family of graphs known as the *Johnson graphs*.

2.1.2 Johnson graphs

The Johnson graph $J_n(m, m-1)$ has $\binom{n}{m}$ vertices, each labelled by a unique set $\nu(v) \in \binom{[n]}{m}$ where two distinct nodes v_i and v_j are adjacent if, and only if, $|\nu(v_i) \cap \nu(v_j)| = m - 1$. Figure 5.1 shows two examples – $J_4(2, 1)$ in (a) and $J_5(3, 2)$ in (b).

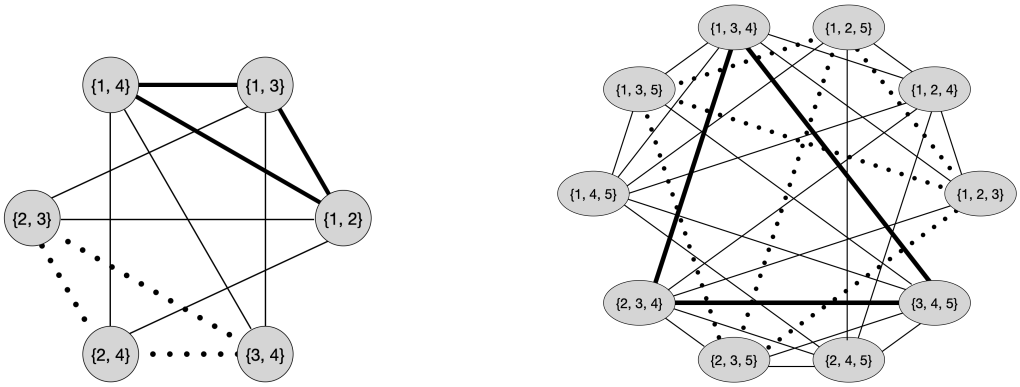
Brouwer et al. (1989) provide a comprehensive examination of Johnson graphs and some of their other properties, such as distance regularity.

Example 2.1.5. The line graph of the complete graph on 5 nodes from example 2.1.4 has two distinct variables from [5] associated to each node. Moreover, two nodes are adjacent precisely when they intersect in one variable and hence the line graph of K_5 is $J_5(2, 1)$.

The example above is an instance of a more general result: the line graph of K_n is the Johnson graph $J_n(2, 1)$. Moreover, if the edges of a complete graph K_n are equipped with

a weight function $w : E(K_n) \to [0, 1]$, the weighted Johnson graph obtained can be turned into a navigation graph akin to the first method discussed in Subsection 1.1.1. That is, if w records the prevalence of a particular facet in scatterplot between two variables (e.g. monotonic or convexity) in a dataset and we omit all edges below a desired threshold, the resulting line graph would have the exact graphic structure of the navigation subgraph described in method 1) of Subsection 1.1.1.

More generally, the Johnson graphs can be viewed as a special case of a larger family of graphs known as the *generalized Johnson graphs*, where the intersection condition is specified via the parameter k. In other words, if $n > m > k \geq 1$ are fixed integers, the *generalized Johnson graph* $J_n(m, k)$ is the graph whose vertices are the $m-$subsets of $[n]$, where two vertices v_1 and v_2 are adjacent if $|\nu(v_1) \cap \nu(v_2)| = k$.



(a) $J_4(2, 1)$ (b) $J_5(3, 2)$

Figure 2.9: Two separate Johnson $J_n(m, m-1)$ graphs with label sets $\nu(v)$ shown on each node v. Nodes are identified as v_1, v_2, ..., beginning from the right most node in each graph and from there in counter-clockwise order. Two maximal cliques are marked on each.

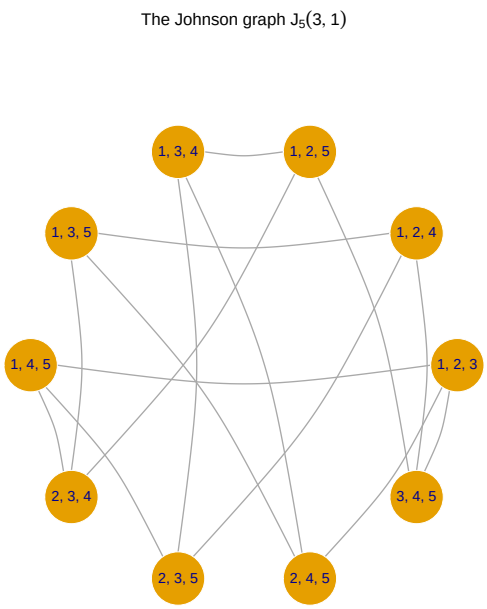


Figure 2.10: The generalized Johnson graph $J_5(3, 1)$.

If $x_1, x_2, \ldots, x_n$ are the variables of a dataset, the Johnson graph $J_n(m, m-1)$ with the addition of visualizations on each of the nodes consisting of the $m-$dimensional subspaces of the data would constitute as a navigation graph on the dataset (as discussed in Section 1.1). Figure 5.1 depicts several different types of cliques that occur in the Johnson graph. For instance, (a) illustrates a two different three-spaces, $\{1, 2\}, \{1, 3\}, \{1, 4\}$ and $\{2, 3\}, \{3, 4\}, \{3, 4\}$. Not only are their labels different, but in this thesis, we argue that they suggest different types of relationships between the underlying variables. The structure of Johnson graph cliques will be examined in Chapter 5.

Because the analyst is typically restricted by time and computational resource constraints, only the most 'interesting' subspaces of the Johnson graph will be explored. As we will describe in Chapter 6, random graphs aid us by serving as components in a model for the navigation graphs we may encounter by when searching for the most interesting spaces.

2.2 Random graphs

The study of random graphs dates to Erdős & Rényi (1959), who used probabilistic ideas to prove the existence of graphs with seemingly conflicting properties. At its core, their idea was that if a collection of objects did not contain an object of a particular type, then the probability of randomly sampling an object of that type must be zero. Therefore, one can prove an object's existence with desired properties by showing that a suitable mechanism for sampling from the collection yielded an object of that type with nonzero

probability. This idea was later applied to several other subdisciplines of mathematics, including number theory, real analysis and linear algebra.

Their idea inspired several more general techniques for non-constructive, existence proofs that rely on the tools of probability. In Chapter 3, we describe moments results that are similar in nature to the main technique of Alon & Spencer (2016, Chapter 2), which serves as a standard reference in the field. Our results are later specialized to special models of networks.

Any model network where the values of some properties are fixed a priori and the rest are random is known as a *random graph*. The simplest example of a random graph is attributed to Erdős and Rényi.

We say that G is a random graph with parameters $G(n, M)$ if G was selected uniformly at random from the collection of all graphs on n nodes and M edges. In this case, we write $G \sim G(n, M)$.

Some of the properties of the $G(n, M)$ are easy to deduce, such as the number of edges, M, and the average degree: $2M/n$. Other properties are not as easy to derive (Newman, 2018). This led to the study of a slightly more flexible model.

We say that G is a *homogeneous Erdős-Rényi random graph* from the $G(n, p)$ model if G has n vertices and an edge between every pair of vertices has an equal probability p of appearing, independently of all other edges. In this case, we write $G \sim G(n, p)$.

Random graphs are often studied because they provide insight into the topology of networks and they provide a foundation on which one can build an understanding of processes taking place on networks, such as the spread of disease (Newman, 2018).

Since there are many potential graphs that arise under various schemes of random graphs (for instance, there are $2^{\binom{n}{2}}$ possible graphs we could encounter under the $G(n, p)$ model for $p \in (0, 1)$), one often studies the properties of the typical random graph by examining the average of a particular property instead. Thus, if $X(G)$ is some theoretical property of networks of interest, such as the count of cliques, one might choose to investigate the average

$$E(X) = \sum_{G} P(G)X(G),$$

where the summation is over all members of the particular class of random graphs.

This idea is useful for several reasons. For one, it is often possible to express $E(X)$ precisely, and in some cases, it is even possible to derive its limiting behaviour as a function of network parameters (i.e. its value when n gets large or p is arbitrarily small). Moreover, if the typical behaviour of a random graph is of interest, then the average is a good proxy.

For instance, the complete graph on n vertices has $\binom{n}{3}$ triangles while the empty graph has none - both graphs are equally probable under $G(n, 0.5)$. Lastly, it has been shown that the distribution of values for many of the commonly used network measures is sharply peaked, becoming concentrated more narrowly around the average as the size of network becomes large, so that all values one is likely to encounter are close to the mean (Bollobás, 2001; Bollobas & Erdős, 1976; Newman, 2018).

Some random graph statistics are easily derived. For instance, since we may view edge inclusions in a homogeneous Erdős-Rényi model as independent identically distributed Bernoulli(p) trials, it is clear that the degree of a vertex in $G(n,p)$ has a binomial distribution:

$$P(\deg(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

Other random graph distributions and statistics are more challenging – despite the simplicity of the degree distribution for a fixed node, there are very few results on the joint degree distribution of the graph itself.

Chapter 6 establishes a connection between random graphs and navigation graphs which served as the initial motivation for this thesis's exploration of cliques in random graphs. Identifying the distribution of cliques in random graphs is a challenging problem which remains without a closed-form solution. To the best of our knowledge, the earliest attempts to resolve this were due to Bollobas & Erdős (1976). Their work presented the following approach for deriving the first and second moments.

Fix $r \geq 3$ an integer and let X_r denote the number of cliques of size r in $G \sim G(n,p)$. For a set A of r nodes in G to span a complete graph, all of the $\binom{r}{2}$ edges between the nodes to be present. Let $Edge(A)$ denote the set of all possible edges between nodes in A. For an edge $e \in Edge(A)$, let Y_e be the indicator random variable recording if e is present in G and let Z_A be the indicator random variable recording if A is an $r-$clique in G. So, the set A is a clique in G if $Y_e = 1$ for all $e \in Edge(A)$. By independence of the Y_e,

$$Pr(Z_A = 1) = Pr(Y_e = 1, \forall e \in E(A)) = p^{\binom{r}{2}}.$$

Since any $r-$subset of $[n]$ could form an $r-$clique in G, the number $X_r(G)$ of $r-$cliques in G is given by

$$X_r(G) = \sum_{I \subseteq [n] : |I| = r} Z_A.$$

By linearity of expectation,

$$E(X_r(G)) = \binom{n}{r} p^{\binom{r}{2}}.$$

By first deriving $E(Z_A Z_B)$ for all pairs of $r-$subsets A, B, Bollobas & Erdős (1976) adapted the argument above to derive the second moment of $X_r(G)$:

$$E(X_r(G)^2) = \sum_{\ell=0}^{r} \binom{n}{r} \binom{r}{\ell} \binom{n-r}{r-\ell} p^{2\binom{r}{2} - \binom{\ell}{2}}.$$

In short, the method described above relies on decomposing a complicated, count random variable into a sum of simple, well-understood indicator random variables. This idea allows one to derive the expected value of a random variable whose distribution is unknown. Chapter 3 expands this idea and presents an expression for all of the moments of any random variable that is a sum of indicator random variables.

2.3 Algebraic combinatorics

This subsection serves as a short review of the basics of generating series. We borrow the notation and terminology found in (Wilf, 2005), (Goulden & Jackson, 1983).

Let $(a_n)_{n\geq 0}$ be a sequence of complex numbers and q an indeterminate. We define the *generating series* $A(q)$ of $(a_n)_{n\geq 0}$ to be the formal sum

$$A(q) := \sum_{n\geq 0} a_n q^n.$$

We call any such summation a *formal power series*. We let $\mathbb{R}[[q]]$ denote the set of all formal power series in q with coefficients in $\mathbb{R}$. In the following chapters, we write $[q^n]A(q)$ to denote the *extraction* of the $n-$th coefficient of $A(q)$. That is,

$$[q^n]A(q) = a_n$$

for all $n \in \mathbb{Z}$ where we use the convention $a_k = 0$ for $k < 0$. We define addition of two formal power series in the usual way:

$$\sum_{n\geq 0} a_n q^n + \sum_{n\geq 0} b_n q^n = \sum_{n\geq 0} (a_n + b_n) q^n.$$

We extend the usual definition of multiplication of polynomials to formal power series as follows:

$$\left(\sum_{n\geq 0} a_n q^n\right)\left(\sum_{n\geq 0} b_n q^n\right) = \sum_{n\geq 0}\left(\sum_{k=0}^{n} a_k b_{n-k}\right) q^n.$$

With respect to these two operations, the set $\mathbb{R}[[q]]$ forms a ring.

Example 2.3.1. Let $(a_n)_{n\geq 0}$ be the sequence defined by $a_n = 1$ for all $n \in \mathbb{N}$. The power series $A(q)$ which corresponds to it is given by

$$A(q) = \sum_{n\geq 0} a_n q^n = \sum_{n\geq 0} q^n = \frac{1}{1-q},$$

and we note that it is easily verified that $\sum_{n\geq 0} q^n$ is the multiplicative inverse of $(1-q)$ in $\mathbb{R}[[q]]$.

In the previous example, $\frac{1}{1-q}$ is an instance of a *closed-form expression* for a generating series. Throughout this thesis, we write *closed-form expression* for a generating series to mean a finite expression in terms of basic arithmetic operations and elementary functions (such as polynomials, trigonometric functions, etc.).

We recall a few standard results regarding expansions of powers of multinomials.

Theorem 2.3.2. *(Binomial theorem) For x, y indeterminates and $n \geq 0$ a positive integer,*

$$(x+y)^n = \sum_{k\geq 0} \binom{n}{k} x^k y^{n-k}.$$

Isaac Newton generalized the binomial theorem by incorporating real exponents. The following generalization also holds for complex numbers.

Theorem 2.3.3 (Newton's generalized binomial theorem)**.** *If x, y are indeterminates and $r \in \mathbb{R}$, then*

$$(x+y)^r = \sum_{k \geq 0} \binom{r}{k} x^k y^{r-k},$$

where

$$\binom{r}{k} := \frac{r(r-1)\cdots(r-k+1)}{k}.$$

If r is a negative integer, then by applying substitutions $x \mapsto 1$, $y \mapsto -y$ into Newton's generalized binomial theorem we obtain

$$(1-y)^r = \sum_{k \geq 0} \binom{k-1-r}{k} y^k.$$

Since binomials are a special case of multinomials, one can also view the binomial theorem as a special case of the multinomial theorem, which is critical to the results presented in Chapter 3.

Theorem 2.3.4 (Multinomial theorem)**.** *Let $y_1, \ldots, y_n$ be a sequence of commutative elements over some ring and fix $k \geq 1$. Then*

$$(y_1 + \cdots + y_n)^k = \sum_{\substack{\ell_1 + \cdots + \ell_n = k \\ \ell_i \geq 0, \forall i}} \binom{k}{\ell_1, \cdots, \ell_n} y_1^{\ell_1} \cdots y_n^{\ell_n}.$$

Proof. The statement follows immediately from the binomial theorem and a simple inductive argument on m. □

We shall use the following well-known Principle of Inclusion and Exclusion. This counting technique enumerates the elements in a finite union of finite sets by counting the number of elements appearing in each possible type of intersection of the underlying sets.

Proposition 2.3.5 (Principle of Inclusion and Exclusion)**.** *If $A_1, \ldots A_n$ is a collection of finite sets, then*

$$\left| \bigcup_{i=1}^{n} A_i \right| = \sum_{i=1}^{n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cap A_j| + \sum_{1 \leq i < j < k \leq n} |A_i \cap A_j \cap A_k| - \cdots + (-1)^{n+1} |A_1 \cap A_2 \cap A_n|.$$

Proof. For an algebraic proof, see (Goulden & Jackson, 1983, pg 47) or (Wilf, 2005, pg 119). For a combinatorial proof, see (Aigner & Axler, 2007, pg 180). □

2.3.1 Directions

The mathematical ideas reviewed above are essential to the clique-centric study of navigation graphs approach in the following chapters. The next three chapters are dedicated to building upon the foundation described above to capture a general theory for the moments of count random variables, obtain expressions for the number of cliques present in a clique cover, describe the clique structure of Johnson graphs and obtain novel expressions for the moments of clique counts in random graphs.

3

Bernoulli sums

As mentioned in Chapter 2, the random graph model for the study of cliques in navigation graphs requires a careful enumeration of the edges associated with a configuration of cliques. In this chapter, we examine a related and more general problem of deriving the moments of a count random variable – a random variable taking on values in $\mathbb{N}_0$.

Consider the random variable

$$X = \sum_{i \in \mathcal{I}} Y_i$$

which sums (not necessarily independent) Bernoulli random variables $Y_i \in \{0, 1\}$ over some countable index set $\mathcal{I}$. When Y_i is an indicator function for some event $\mathcal{A}$, the *Bernoulli sum* X counts the number of occurrences of the event in the set $\mathcal{I}$ and, as such, arises in numerous applications of probability. Of interest here is the determination of the moments, central moments, and factorial moments of any arbitrary Bernoulli sum.

We develop expressions for these moments in terms of the expectation of products of the Bernoulli Y_is. This leads to novel proofs and/or expressions for the moments in many well known problems and to novel approaches to determining such moments for any random variable expressible as a Bernoulli sum.

This chapter is organized as follows. Section 3.1 shows that the power of a sum of idempotents is expressible in terms of the number of surjections from one finite set to another times a sum of their products. This follows as a special case of the multinomial theorem. Section 3.2 builds on this to develop the main general results for the moments of a Bernoulli sum. Both finite and infinite sums are considered and special attention is given to factorial moments and generating functions.

These results are then applied to develop expressions for various classic distributions and problems in Section 3.3. These include the binomial, poisson binomial, hypergeometric, and Conway-Maxwell-Poisson binomial distributions, the Poisson limit of a binomial by moment convergence, and the classic empty urns problem and the matching problem.

Section 3.4 considers the moments for any count random variable, developing expressions based on the upper tail probability of that count. This general theory is then demonstrated on the geometric, Poisson, Ideal Soliton, and Benford distributions.

3.1 Idempotent multinomial theorem

As per Chapter 2, recall that $[k]$ denotes the set $\{1, \ldots, k\}$ for any finite integer k. Let $S(k, m)$ denote the number of surjections from $[k]$ to $[m]$. If $k < m$, no surjective function exists and $S(k, m) = 0$; otherwise, $S(k, m)$ can be written as

$$S(k, m) = \sum_{v=0}^{m-1} (-1)^v \binom{m}{v} (m - v)^k$$

(e.g., see Wilf (2005)). The number $S(k, m)$ figures prominently in the closed form expressions which follow.

In particular, $S(k,m)$ (for all $m \leq k$) will be shown to appear in expressions for the kth moments of a Bernoulli sum. To calculate the smaller moments of importance in statistical inference (say $k \leq 4$), it will be convenient therefore to have $S(k,m)$ evaluated for a few $m \leq k$. Whenever k is at least as large as the second argument, the following values are obtained: $S(k,0) = 0$, $S(k,1) = 1$, $S(k,2) = 2^k - 2$, $S(k,3) = 3^k - 3 \cdot 2^k + 3$, and $S(k,4) = 4^k - 4 \cdot 3^k + 6 \cdot 2^k - 4$. These will appear in calculations up to the 4th moment (e.g. to determine kurtosis). Again, note that $S(k,m) = 0$ whenever $m > k$.

Moments of X are expectations of powers of X which, in the case of $X = \sum_{i=1}^{n} Y_i$, suggests beginning with a multinomial theorem (see Theorem 2.3.4):

$$X^k = (Y_1 + \cdots + Y_n)^k = \sum_{\substack{\ell_1 + \cdots + \ell_n = k \\ \ell_i \geq 0, \forall i}} \binom{k}{\ell_1, \cdots, \ell_n} Y_1^{\ell_1} \cdots Y_n^{\ell_n}.$$

In this section, we only consider the Y_is which are idempotents, as they are in the definition of a Bernoulli sum, only those Y_i with $\ell_i \geq 1$ remain and simplify to $Y_i^{\ell_i} = Y_i$. This leads to the following version of the multinomial theorem where now $S(k,m)$ appears.

Proposition 3.1.1. *Let $y_1, \ldots, y_n$ be a sequence of commutative idempotents over some ring. Then*

$$(y_1 + \cdots + y_n)^k = \sum_{m=1}^{k} S(k,m) \sum_{\{i_1, \ldots, i_m\} \subseteq [n]} y_{i_1} \cdots y_{i_m}$$

where $S(k,m)$ is the number of surjections from $[k]$ onto $[m]$.
(It is understood that the interior sum has $m \leq \min\{k,n\}$.)

Proof. Naively expanding $(y_1 + \cdots + y_n)^k$ gives

$$(y_1 + y_2 + \cdots + y_n)^k = \sum_{(j_1, j_2, \ldots, j_k) \in [n]^k} y_{j_1} y_{j_2} \cdots y_{j_k}. \tag{3.1}$$

Let $\mathcal{F}$ denote the set of all functions $f : [k] \to [n]$. For a product $y_{j_1} y_{j_2} \cdots y_{j_k}$ on the right-hand side of Equation 3.1, let f be the function which maps $\ell \in [k]$ to j_ℓ. Since every $j_\ell \in [n]$, this defines a function $f \in \mathcal{F}$. Conversely, a unique summand of the form $y_{f(1)} y_{f(2)} \cdots y_{f(k)}$ can be assigned to each function $f \in \mathcal{F}$. That is, the naive expansion of $(y_1 + \cdots + y_n)^k$ results in n^k summands of the form $y_{f(1)} y_{f(2)} \cdots y_{f(k)}$ for some function $f : [k] \to [n]$.

Since $y_1, \ldots, y_n$ are commutative idempotents, each product $y_{j_1} \cdots y_{j_k}$ resolves to a unique $y_{i_1} \cdots y_{i_m}$ with indices $i_1 < \ldots < i_m$, for some $m \in [k]$. Equation 3.1 then becomes

$$(y_1 + y_2 + \cdots + y_n)^k = \sum_{(j_1, j_2, \ldots, j_k) \in [n]^k} y_{j_1} y_{j_2} \cdots y_{j_k} = \sum_{m=1}^{k} a(k,m) \sum_{\{i_1, \ldots, i_m\} \subseteq [n]} y_{i_1} \cdots y_{i_m}. \tag{3.2}$$

Here $a(k,m)$ counts the number terms $y_{i_1} \cdots y_{i_m}$ that simplify to $y_{j_1} \cdots y_{j_k}$. It remains only to show that $a(k,m)$ equals $S(k,m)$, the number of surjective maps from $[k]$ onto $[m]$.

To see this, first fix $\{i_1, \ldots, i_m\} \subseteq [n]$ and let $F \subseteq \mathcal{F}$ denote the subset of functions for which $y_{f(1)} \cdots y_{f(k)}$ simplifies to $y_{i_1} \cdots y_{i_m}$. The count $a(k,m)$ is identical to $|F|$. Then consider the set, G, of all surjections $g : [k] \to \{i_1, \ldots, i_m\}$, which must have size $|G| = S(k,m)$. If $F = G$, then $|F| = |G|$ and $a(k,m) = S(k,m)$, as required.

Now $F = G$ iff every $f \in F$ is also in G and every $g \in G$ is also in F. If $f \in F$, then $y_{f(1)} \cdots y_{f(k)} = y_{i_1} \cdots y_{i_m}$, giving $f([k]) = \{i_1, \ldots, i_m\}$, and hence $f \in G$. If $g \in G$, then clearly

$$y_{g(1)} \cdots y_{g(k)} = \prod_{\ell \in g([k])} y_\ell = \prod_{\ell \in \{i_1, \ldots, i_m\}} y_\ell,$$

and so $g \in F$. □

Note that the inner sum $\sum_{\{i_1,\ldots,i_m\} \subseteq [n]} y_{i_1} \cdots y_{i_m}$ vanishes whenever $m > n$ and hence $(y_1 + \cdots + y_n)^k$ is expressible as a sum of at most $min(k,n)$ terms involving the coefficients $S(k,m)$.

A generalization of the result to powers of infinite sums, subject to convergence having been settled for all particular values of the y_is (as is the case, for instance, when the sum of any partial product of terms is absolutely convergent), is relatively straightforward.

Proposition 3.1.2. *Let $(y_i)_{i \geq 1}$ be a sequence of formal, commutative, idempotents over some ring. Then*

$$\left(\sum_{i=1}^{\infty} y_i \right)^k = \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\} \subset \mathbb{N}} y_{i_1} \cdots y_{i_m}$$

where $S(k,m)$ is the number of surjections from $\{1, \ldots, k\}$ onto $\{1, \ldots, m\}$.

Proof. Since the $(y_i)_{i \geq 1}$ are formal, commutative idempotents over some ring, the proof follows that of Proposition 3.1.1. □

Because Propositions 3.1.1 and 3.1.2 express a product as a summation, we can now obtain moment expressions for count random variables via the linearity of expectation.

3.2 Moments of Bernoulli sums

Consider the Bernoulli sum random variable X of Section 3.1 with finite index set $\mathcal{I}$ of size n. The set $\mathcal{I}$ can always be re-indexed to have X appear as

$$X = \sum_{i=1}^{n} Y_i.$$

Expressions for the moments of X can now be derived via Proposition 3.1.1.

Proposition 3.2.1. *When X is expressible as a finite Bernoulli sum $X = \sum_{i=1}^{n} Y_i$, the kth* ***moment*** *of X is expressible as*

$$E(X^k) = \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\}\subseteq[n]} E(Y_{i_1}\cdots Y_{i_m}).$$

Proof. Since $Y_i^2 = Y_i$ for all i, it follows from Proposition 3.1.1 that

$$X^k = (Y_1 + \cdots + Y_n)^k = \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\}\subseteq[n]} Y_{i_1}\cdots Y_{i_m}. \tag{3.3}$$

The result follows by applying expectation $E(\cdot)$ operator to each side of Equation 3.3. □

Note that a similar result holds for any linear operator L applied to both sides of equation 3.3:

$$L(X^k) = \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\}\subseteq[n]} L(Y_{i_1}\cdots Y_{i_m}).$$

Now, since

$$E(Y_{i_1}\cdots Y_{i_m}) = Pr(Y_{i_1} = 1, \cdots, Y_{i_m} = 1)$$

Proposition 3.2.1 shows that the moments of *any* finite Bernoulli sum random variable can be investigated via the joint distribution of those Bernoulli random variables used to construct it – indeed, Proposition 3.2.1 could be rewritten in terms of this probability.

The *central moments* are generally of more statistical interest and a similar result is found for them by applying Proposition 3.2.1. In this case, let $p_i = Pr(Y_i = 1) = E(Y_i)$ denote the ith marginal mean in the sum and $\mu = E(X) = \sum_{i=1}^{n} p_i$ the mean of X. A similar expression for the kth central moment is given in Proposition 3.2.2.

Proposition 3.2.2. *When X is expressible as a finite Bernoulli sum $X = \sum_{i=1}^{n} Y_i$, with $p_i = Pr(Y_i = 1) = E(Y_i)$, then the kth* ***central moment*** *of X is expressible as*

$$E((X-\mu)^k) = (-\mu)^k + \sum_{\ell=1}^{k} \binom{k}{\ell} (-\mu)^{k-\ell} \sum_{m=1}^{\ell} S(\ell,m) \sum_{\{i_1,\ldots,i_m\}\subseteq[n]} E(Y_{i_1}\cdots Y_{i_m})$$

where $\mu = E(X) = \sum_{i=1}^{n} p_i$.

Proof. Applying the binomial expansion, then Proposition 3.2.1, yields

$$\begin{aligned} E((X-\mu)^k) &= \sum_{\ell=0}^{k} \binom{k}{\ell} E(X^\ell)(-\mu)^{k-\ell} \\ &= \binom{k}{0} \cdot 1 \cdot (-\mu)^k + \sum_{\ell=1}^{k} \binom{k}{\ell} (-\mu)^{k-\ell} E(X^\ell) \\ &= (-\mu)^k + \sum_{\ell=1}^{k} \binom{k}{\ell} (-\mu)^{k-\ell} \sum_{m=1}^{\ell} S(\ell, m) \sum_{\{i_1,\ldots,i_m\}\subseteq[n]} E(Y_{i_1} \cdots Y_{i_m}) \end{aligned}$$

□

Of course, whenever the Y_is are also **independently distributed**, the above moment expressions (and those which follow) simplify by replacing $E(Y_{i_1} \cdots Y_{i_m})$ by $p_{i_1} \cdots p_{i_m}$, where each $Y_i \sim \text{Bernoulli}(p_i)$.

3.2.1 Moments of an infinite sequence

Consider now an infinite sequence

$$(Y_i)_{i\geq 1} = Y_1, Y_2, \ldots$$

of Bernoulli random variables and their sum

$$X = \sum_{i=1}^{\infty} Y_i$$

being such that $Pr(X < \infty) = 1$. For example, this condition is satisfied whenever the first moment of X is bounded, that is, whenever $E(X) = \sum_{i=1}^{\infty} E(Y_i) = \sum_{i\geq 1} p_i = \mu < \infty$. From this it follows (e.g., by the Borel-Cantelli lemma) that $Pr(\limsup_{n\to\infty} Y_n = 1) = 0$, and, so, that the probability is zero that infinitely many of the Y_is will be 1.)

In this case, Proposition 3.1.2 gives the kth moment for this sum of countably infinite Bernoulli random variables (whenever all relevant sums converge).

Proposition 3.2.3. *Let $X = \sum_{i=1}^{\infty} Y_i$ be the sum of the sequence of Bernoulli random variables $(Y_i)_{i\geq 1}$, with $p_i = Pr(Y_i = 1) = 1 - Pr(Y_i = 0)$, then we may write*

$$E(X^k) = \sum_{m=1}^{k} S(k, m) \sum_{\{i_1,\ldots,i_m\}\subset\mathbb{N}} E(Y_{i_1} \cdots Y_{i_m}).$$

In the special case where the Y_is are also independent, then Proposition 3.2.3 allows us to draw the interesting conclusion that *a bounded first moment of X implies that all higher order moments are also bounded.* This result is formally given in Proposition 3.2.4:

Proposition 3.2.4. *Let $(Y_i)_{i\geq 1}$ be a sequence of **independent** Bernoulli(p_i) random variables with $\sum_{i\geq 1} p_i = \mu < \infty$. For the Bernoulli (infinite) sum random variable $X = \sum_{i\geq 1} Y_i$, and for any $k \geq 1$,*

$$E(X^k) < \infty.$$

Proof. By Proposition 3.2.3 and independence of the Y_i,

$$\begin{aligned} E(X^k) &= \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\}\subset\mathbb{N}} p_{i_1}\cdots p_{i_m} \\ &\leq \sum_{m=1}^{k} S(k,m) \left(\sum_{i\geq 1} p_i\right)^m \\ &= \sum_{m=1}^{k} S(k,m)\ \mu^m \\ &< \infty. \end{aligned}$$

□

In this special case of an infinite sequence of independent Bernoulli random variables, an expression for the moments involving only the first moments of the Y_is and of X can be easily had as well.

Proposition 3.2.5. *Let $(Y_i)_{i\geq 1}$ be a sequence of **independent** Bernoulli(p_i) random variables with $\sum_{i\geq 1} p_i = \mu < \infty$. For the Bernoulli (infinite) sum random variable $X = \sum_{i\geq 1} Y_i$, and for any $k \geq 2$, the kth **moment** of X is*

$$E(X^k) = \sum_{m=1}^{k} S(k,m) \left[\mu^m - \sum_{s=0}^{k-2} \mu^s (m-1-s) \sum_{\{i_1\ldots,i_{m-1-s}\}\subset\mathbb{N}} p_{i_1}^2 (p_{i_2}\cdots p_{i_{m-1-s}})\right].$$

Proof. Fix an integer $r \geq 2$ and note that

$$\begin{aligned} \sum_{\{i_1,\ldots,i_r\}\subset\mathbb{N}} p_{i_1}\cdots p_{i_r} &= \sum_{\{i_1,\ldots,i_{r-1}\}\subset\mathbb{N}} p_{i_1}\cdots p_{i_{r-1}} \left(\sum_{i\notin\{i_1,\ldots,i_{r-1}\}} p_i\right) \\ &= \sum_{\{i_1,\ldots,i_{r-1}\}\subset\mathbb{N}} p_{i_1}\cdots p_{i_{r-1}} \left(\mu - \sum_{i\in\{i_1,\ldots,i_{r-1}\}} p_i\right) \\ &= \mu \sum_{\{i_1,\ldots,i_{r-1}\}\subset\mathbb{N}} p_{i_1}\cdots p_{i_{r-1}} - (r-2) \sum_{\{i_1,\ldots,i_{r-1}\}\subset\mathbb{N}} p_{i_1}^2\cdots p_{i_{r-1}}, \end{aligned}$$

where the last equality follows from the fact that

$$\sum_{\{i_1,\ldots,i_{r-1}\}\subset\mathbb{N}} p_{i_1}^2 p_{i_2}\cdots p_{i_{r-1}} = \sum_{\{i_1,\ldots,i_{r-1}\}\subset\mathbb{N}} p_{i_1} p_{i_2}^2\cdots p_{i_{r-1}} = \cdots = \sum_{\{i_1,\ldots,i_{r-1}\}\subset\mathbb{N}} p_{i_1} p_{i_2}\cdots p_{i_{r-1}}^2.$$

Recursively rewriting $\sum_{\{i_1,\ldots,i_r\}\subset\mathbb{N}} p_{i_1}\cdots p_{i_r}$ in terms of sums over one fewer index (viz., $r-1$ indices) each time gives the desired result via Proposition 3.2.1. □

3.2.2 Factorial moments

The kth falling factorial of x is the kth degree polynomial in x

$$[x]_k := x(x-1)(x-2)\cdots(x-(k-1)) = \prod_{m=0}^{k-1}(x-m),$$

where $k \in \mathbb{N}$ and $x \in \mathbb{R}$. Replacing x by a random variable X, the corresponding kth **factorial moment** is defined to be $E\left([X]_k\right)$. Like $E(X^k)$ this is the expected value of the product of k terms. Note this is different from $E(X!)$, the **expected factorial** of X, where the number of products in $X!$ is itself be a random variable (viz., X).

The kth power of x can be expressed (Stanley, 2011) in terms of falling factorials as

$$x^k = \sum_{m=1}^{k} S_2(k,m)[x]_m,$$

where $S_2(k,m)$ is the Stirling number of the second kind, typically defined as the number of ways to partition a set of k labelled objects into m nonempty unlabelled subsets. It follows, then, that these are directly related to the number of surjections from a $k-$set onto an $m-$set as

$$S(k,m) = m!\ S_2(k,m)$$

and hence that

$$x^k = \sum_{m=1}^{k} \frac{S(k,m)}{m!}[x]_m = \sum_{m=1}^{k} S(k,m)\binom{x}{m}.$$

Similarly, the falling factorial is written as a sum of powers as

$$[x]_k = \sum_{m=1}^{k} S_1(k,m)x^m$$

where $S_1(k,m)$ is the Stirling number of the first kind. Similar expressions may now be found involving a Bernoulli sum X in place of x.

First, we relate $\binom{X}{m}$ to the Bernoulli random variables that define X.

Proposition 3.2.6. *If X is a Bernoulli sum $X = \sum_{i\in I} Y_i$ for some countable indexing set $\mathcal{I}$, then for $m \geq 1$,*

$$\binom{X}{m} = \sum_{\{i_1,\ldots,i_m\}\subseteq\ \mathcal{I}} Y_{i_1}\cdots Y_{i_m}.$$

Proof. Let $\mathcal{J} = \{i \in \mathcal{I} : Y_i = 1\}$ denote the subset of the indices in $\mathcal{I}$ for which $Y_i = 1$.

Then

$$\begin{aligned}\sum_{\{i_1,\ldots,i_m\}\subseteq\mathcal{I}} Y_{i_1}\cdots Y_{i_m} &= \sum_{\{i_1,\ldots,i_m\}\subseteq\ \mathcal{J}} Y_{i_1}\cdots Y_{i_m}\\ &= \sum_{\{i_1,\ldots,i_m\}\subseteq\ \mathcal{J}} 1\\ &= \binom{|\mathcal{J}|}{m}\\ &= \binom{X}{m}.\end{aligned}$$

□

An earlier, inductive, proof of this result for the case of finite $\mathcal{I}$ is given by Iyer (1958).

A similar approach yields a general result relating $X!$ to its Bernoulli constituents.

Proposition 3.2.7. *Let $X = \sum_{i\in\mathcal{I}} Y_i$ be a Bernoulli sum, where $\mathcal{I}$ is a countable indexing set. Then we may write $X!$ in terms of the (Y_i) as follows*

$$X! = \sum_{\mathcal{H}\subseteq\mathcal{I}} |\mathcal{H}|! \left(\prod_{i\in\mathcal{H}} Y_i \prod_{i\notin\mathcal{H}} (1-Y_i)\right).$$

Proof. Consider the set $\mathcal{J} := \{i \in \mathcal{I} : Y_i = 1\}$. In this case, $|\mathcal{J}| = X$ and $|\mathcal{J}|! = X!$. For any other set $\mathcal{H} \subseteq \mathcal{I}$, either $\mathcal{H} = \mathcal{J}$, or $\mathcal{H} \neq \mathcal{J}$.

If $\mathcal{H} = \mathcal{J}$, then

$$\prod_{i\in H} Y_i \prod_{i\notin\mathcal{H}} (1-Y_i) = \prod_{i\in\mathcal{H}} 1 \prod_{i\notin\mathcal{H}} (1-0) = 1$$

If $\mathcal{H} \neq \mathcal{J}$, then there exists j for which $j \in \mathcal{J}$ but $j \notin \mathcal{H}$. Then,

$$\prod_{i\in\mathcal{H}} Y_i \prod_{i\notin\mathcal{H}} (1-Y_i) = \prod_{i\in\mathcal{H}} Y_i \times 0 = 0.$$

Together these give

$$\sum_{\mathcal{H}\subseteq\mathcal{I}} |\mathcal{H}|! \left(\prod_{i\in\mathcal{H}} Y_i \prod_{i\notin\mathcal{H}} (1-Y_i)\right) = |\mathcal{J}|! \prod_{i\in\mathcal{J}} Y_i \prod_{i\notin\mathcal{J}} (1-Y_i) = |\mathcal{J}|! = X!$$

□

Taking expectations yields the following expressions for a Bernoulli sum $X = \sum_{i\in\mathcal{I}} Y_i$:

- the kth factorial moment in terms of the Bernoulli random variables

$$E([X]_k) = k! \sum_{\{i_1,\ldots,i_k\}\subseteq\ \mathcal{I}} E(Y_{i_1}\cdots Y_{i_k})$$

  or, in terms of the moments of X as

$$E([X]_k) = \sum_{m=1}^{k} S_1(k,m)E(X^m)$$

- the kth moment in terms of the factorial moments of X

$$E(X^k) = \sum_{m=1}^{k} S_2(k,m)\ E([X]_m)$$

- the kth central moment

$$E((X-\mu)^k) = (-\mu)^k + \sum_{j=1}^{k}\left(\sum_{m=j}^{k} S_2(m,j)\binom{k}{m}(-\mu)^{k-m}\right)E([X]_j)$$

- and the expected factorial in terms of the Bernoulli random variables

$$E(X!) = \sum_{\mathcal{H}\subseteq\mathcal{I}} |\mathcal{H}|!\ \times E\left[\prod_{i\in\mathcal{H}} Y_i \prod_{i\notin\mathcal{H}}(1-Y_i)\right].$$

Central moments for small k can always be written in terms of the moments or in terms of the factorial moments. When $k = 2$, a nice symmetry appears in either expression for the variance of X:

$$Var(X) = E(X^2) - (E(X))^2 = E([X]_2) - [E(X)]_2.$$

3.2.3 A statistical interpretation

Central moments are statistically meaningful for any random variable X where available. However, when X is a Bernoulli sum a few more meaningful interpretations are available.

Imagine a collection of individuals $i \in \mathcal{I}$, from which a random number X provides a population $\mathcal{J}$ of size X. Samples of fixed size k are to be drawn from the resulting population $\mathcal{J}$. Here, $X = \sum_{i\in\mathcal{I}} Y_i$ and $\mathcal{J} = \{i \in \mathcal{I} : Y_i = 1\}$ with (possibly dependent) random variables $Y_i \sim Bernoulli(p_i)$ (indicating inclusion in the population $\mathcal{J}$ when $Y_i = 1$ and exclusion when $Y_i = 0$).

In this case, the *expected number of samples of size k*

- is the kth factorial moment $E([X]_k)$ when sampling *without replacement* and
- is the kth moment $E(X^k)$ when sampling *with replacement.*

The expected factorial $E(X!)$ is the *expected number of permutations* one would have in the indices found by forming a population in this way.

3.2.4 Generating functions

Various generating functions for a Bernoulli sum X are now easily had by substitution of

- $E(X^k)$ in the *moment generating function*

$$M_X(s) = E(e^{sX}) = 1 + \sum_{k=1}^{\infty} E(X^k)\frac{s^k}{k!}$$

- $E([X]_k)$ in the *factorial moment generating function* (e.g., see p. 59 of Johnson et al. (2005))

$$H_X(s) = 1 + \sum_{k=1}^{\infty} E([X]_k)\frac{s^k}{k!}$$

- and, from Fréchet (1943),

$$Pr(X = x) = \sum_{j \geq x} (-1)^{x+j} \binom{j}{x} \frac{E([X]_j)}{j!},$$

  or, after substitution for the factorial moments,

$$Pr(X = x) = \sum_{j \geq x} (-1)^{x+j} \binom{j}{x} \sum_{\{i_1,\ldots,i_j\} \subseteq \mathcal{I}} E(Y_{i_1} \cdots Y_{i_j}),$$

  the probability $Pr(X = x)$ into the *probability generating function*

$$G_X(s) = E(s^X) = \sum_{k=0}^{\infty} s^k Pr(X = k).$$

The factorial moment generating function, $H_X(s)$, can be related (again, see Johnson et al. (2005) [p. 59]) to the probability generating function, $G_X(s)$, as

$$H_X(s) = G_X(1 + s) = E((1 + s)^X).$$

It follows that whenever factorial moments are such that $H_X(s)$ has a tidy closed form, the probability generating function of X might be easily obtained through the reverse relation

$$G_X(s) = H_X(s - 1). \tag{3.4}$$

This approach will be illustrated for the binomial distribution in Section 3.3.1, and for the classic matching problem of Section 3.3.5, to determine expressions for the probability generating function of X in each of these classic cases.

3.3 Classic examples

Bernoulli sums naturally arise in many classic problems and lead to well known distributions. In this section, the results of Section 3.2 are applied to several of these where the Bernoulli sum is over a finite index set (of size n), namely

$$X = \sum_{i=1}^{n} Y_i$$

where $Y_i \sim Bernoulli(p_i)$ with $p_i = Pr(Y_i = 1) = 1 - Pr(Y_i = 0)$ for $i = 1, \ldots, n$.

3.3.1 Binomial X

The simplest case where the Y_is are independent and identically distributed (i.i.d.) with $p_i = p \; \forall \; i$, $X \sim binomial(n, p)$. The kth moment of X can be written as

$$\begin{aligned} E(X^k) &= \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\} \subseteq [n]} p^m \\ &= \sum_{m=1}^{k} S(k,m) \binom{n}{m} \; p^m. \end{aligned}$$

An equivalent expression is found by Knoblauch (2008) using a recursive argument. In contrast, the result is easily had here from simple application of the more general Proposition 3.2.1. Central moments follow from Proposition 3.2.2:

$$E((X-\mu)^k) = (-np)^k + \sum_{\ell=1}^{k} (-np)^{k-\ell} \sum_{m=1}^{\ell} S(\ell, m) \binom{n}{m} p^m.$$

The kth factorial moment has a appealingly simple expression $E([X]_k) = [n]_k \; p^k$ derived as

$$E([X]_k) = k! \sum_{\{i_1,\ldots,i_k\} \subseteq [n]} p_{i_1} \cdots p_{i_k} = k! \binom{n}{k} p^k = [n]_k \; p^k,$$

the familiar $E(X) = np$ being the special case when $k = 1$.

Note that whenever $k \geq n$, many terms disappear in the above moment expressions since $S(n, m)$ vanishes whenever $m > n$ and the sum $\sum_{\{i_1,\ldots,i_k\} \subseteq [n]}$ is over the empty set.

The moment generating function of a binomial X also has a new expression following application of Equation 3.5,namely

$$M_X(t) = 1 + \sum_{k \geq 1} \frac{t^k}{k!} \sum_{m=1}^{k} S(k,m) \binom{n}{m} p^m \tag{3.5}$$

compared to $M_X(t) = (1 - p + pe^t)^n$.

Recall that the probability generating function of $X \sim binomial(n, p)$ is

$$G_X(s) = ((1-p) + ps)^n .$$

By Equation 3.4, we find that the factorial moment generating function for X is

$$\begin{aligned} H_X(s) &= ((1-p) + p(1+s))^n \\ &= \sum_{m=0}^{n} s^m \left(\sum_{r=m}^{n} \sum_{\ell=0}^{n-r} \binom{n}{r} \binom{r}{m} (-1)^\ell p^{\ell+r} \right), \end{aligned}$$

by applying the binomial theorem and changing the order of summation. This provides us an additional expression for the $k-$th factorial moment of X:

$$[n]_k p^k = \sum_{r=k}^{n} \sum_{\ell=0}^{n-r} \binom{n}{r} \binom{r}{k} (-1)^\ell p^{\ell+r}.$$

Poisson binomial X

If $Y_i \sim Bernoulli(p_i)$ independently for all i but $p_i \neq p_j$ for (at least one) $i \neq j$, the distribution of X is called a Poisson binomial distribution (e.g., see Shah (1973)). The various moments of X are exactly as given by the relevant results of Section 3.2 with $E(Y_{i_1} \cdots Y_{i_m})$ everywhere replaced by $p_{i_1} \cdots p_{i_m}$. So too for its moment generating function.

3.3.2 Hypergeometric X

Consider a sample of size n randomly drawn without replacement from a population of N individuals where g of them have some trait which is absent from the remaining $N - g$. The indicator random variable, Y_i, records if the ith individual selected has the desired trait ($Y_i = 1$) or not ($Y_i = 0$) and $X = \sum_{i=1}^{n} Y_i$ counts the number in the sample having the trait.

The ith draw is a a Bernoulli random variable Y_i with probability

$$p_i = \frac{g - \ell}{N - (i-1)}$$

where ℓ is the number of previous $(i-1)$ draws having the trait. A sample of m of these Bernoullis satisfies

$$E(Y_{i_1} \cdots Y_{i_m}) = \frac{g(g-1)\cdots(g-m+1)}{N(N-1)\cdots(N-m+1)} = \frac{[g]_m}{[N]_m}$$

provided $m \leq g$ and is zero whenever $m > g$ (since at least one Y_i must be zero).

The kth moment of X, following Proposition 3.2.1, is now

$$E(X^k) = \sum_{m=1}^{\min k,g} S(k,m) \sum_{\{i_1,\ldots,i_m\}\subseteq[n]} \frac{[g]_m}{[N]_m}$$
$$= \sum_{m=1}^{\min k,g} S(k,m)\binom{n}{m}\frac{[g]_m}{[N]_m},$$

where the last equality followed from $m-$symmetry. Similarly, the central moments are

$$E((X-\mu)^k) = \left(-n\frac{g}{N}\right)^k + \sum_{\ell=1}^{M}\binom{k}{\ell}\sum_{m=1}^{\ell} S(\ell,m)\binom{n}{m}\frac{[g]_m}{[N]_m}\left(-n\frac{g}{N}\right)^{k-\ell}$$

where $M = \min k, g$.

The factorial moments again have a pleasingly simple expression when $k \leq g$ (zero whenever $k > g$), namely,

$$E([X]_k) = \; k!\binom{n}{k}\frac{[g]_k}{[N]_k} \;=\; [n]_k\;\frac{[g]_k}{[N]_k}.$$

Where the binomial $E([X]_k) = [n]_k p^k$, the hypergeometric now has $\frac{[g]_k}{[N]_k}$ in place of p^k, as one might expect.

3.3.3 CMP-binomial X

For $n \in \mathbb{N}, p \in [0,1], \nu \in \mathbb{R}$, a random variable X has a *Conway-Maxwell-Poisson (CMP) binomial* distribution with parameters (n,p,ν) if its probability mass at $X = j$ ($j \in [n]$) is given by

$$Pr(X=j) = \frac{1}{C_{n,p,\nu}}\binom{n}{j}^{\nu} p^j(1-p)^{n-j},$$

where $C_{n,p,\nu}$ is the normalizing constant

$$C_{n,p,\nu} = \sum_{j=0}^{n}\binom{n}{j}^{\nu} p^j(1-p)^{n-j}.$$

The distribution is formed from a Conway-Maxwell-Poisson (CMP) random variable conditional on the sum of that variable and another one independently generated from a different CMP-distribution.

Just as the CMP-distribution generalizes a Poisson random variable to model count data having variability larger (over dispersed) or smaller (under dispersed) than that of a Poisson, the *CMP-binomial* generalizes the binomial distribution. A CMP-binomial distribution is binomial when $\nu = 1$ and has larger (smaller) variance than a binomial

when $\nu < 1$ ($\nu > 1$). When $\nu = 0$, the most extreme values of 0 and n are favoured; when $\nu \to \infty$ the count X achieves the middle value of $n/2$ when n is even and $(n \pm 1)/2$ when n is odd. See Shmueli et al. (2005) for details.

As noted by Shmueli et al. (2005), the random variable X can also be viewed as a sum of exchangeable, Bernoulli random variables Y_i with joint probability

$$Pr(Y_1 = y_1, \ldots, Y_n = y_n) = \frac{1}{C_{n,p,\nu}} \binom{n}{\sum_{i=1}^n y_i}^{\nu-1} p^{\sum_{i=1}^n y_i} (1-p)^{n-\sum_{i=1}^n y_i},$$

where $\nu > 1$ in the case of negatively correlated trials and $\nu < 1$ for positively correlated trials. This observation allows an expression to be written for the moments of X from an expression $Pr(Y_{i_1} = 1, \ldots, Y_{i_m} = 1)$ for an arbitrary $m-$set $\{i_1, \ldots, i_m\} \subseteq [n]$.

$$\begin{aligned}
Pr(Y_{i_1}, \ldots, Y_{i_m}) &= \sum_{\substack{y_j \in \{0,1\} \\ \forall j \notin \{i_1,\ldots,i_m\}}} Pr(Y_{i_1} = 1, \ldots, Y_{i_m} = 1, \text{ and } Y_j = y_j, \forall j \notin \{i_1, \ldots, i_m\}) \\
&= \sum_{\substack{y_j \in \{0,1\} \\ \forall j \notin \{i_1,\ldots,i_m\}}} \frac{1}{C_{n,p,\nu}} \binom{n}{m + \sum_{j \notin \{i_1,\ldots,i_m\}} y_j}^{\nu-1} \\
&\qquad \times p^{m + \sum_{j \notin \{i_1,\ldots,i_m\}} y_j} (1-p)^{n - (m + \sum_{j \notin \{i_1,\ldots,i_m\}} y_j)} \\
&= \frac{1}{C_{n,p,\nu}} \sum_{s=0}^{n-m} \sum_{\substack{y_j \in \{0,1\} \\ \forall j \notin \{i_1,\ldots,i_m\} \\ \sum_{j \notin \{i_1,\ldots,i_m\}} y_j = s}} \binom{n}{m+s}^{\nu-1} \times p^{m+s} (1-p)^{n-(m+s)} \\
&= \frac{1}{C_{n,p,\nu}} \sum_{s=0}^{n-m} \binom{n-m}{s} \binom{n}{m+s}^{\nu-1} p^{m+s} (1-p)^{n-(m+s)} \\
&= \frac{1}{C_{n,p,\nu}} \sum_{\ell=m}^{n} \binom{n-m}{\ell-m} \binom{n}{\ell}^{\nu-1} p^{\ell} (1-p)^{n-\ell}.
\end{aligned}$$

The kth moment of $X \sim CMP - binomial(n, p, \nu)$ is then

$$\begin{aligned} E(X^k) &= \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\}\subseteq[n]} E(Y_{i_1}\cdots Y_{i_m}) \\ &= \frac{1}{C_{n,p,\nu}} \sum_{m=1}^{k} S(k,m) \binom{n}{m} \sum_{\ell=m}^{\min n,k} \binom{n-m}{\ell-m} \binom{n}{\ell}^{\nu-1} p^{\ell}(1-p)^{n-\ell} \\ &= \frac{1}{C_{n,p,\nu}} \sum_{m=1}^{k} S(k,m) \sum_{\ell=m}^{\min n,k} \binom{\ell}{m} \binom{n}{\ell}^{\nu} p^{\ell}(1-p)^{n-\ell}. \end{aligned}$$

Similarly, the kth central moment is

$$(-np)^k + \frac{1}{C_{n,p,\nu}} \sum_{\ell=1}^{k} \binom{k}{\ell} (-np)^{k-\ell} \sum_{m=1}^{\ell} S(\ell,m) \sum_{\ell=m}^{\min n,k} \binom{\ell}{m} \binom{n}{\ell}^{\nu} p^{\ell}(1-p)^{n-\ell}$$

and the kth factorial moment

$$E([X]_k) = \frac{k!}{C_{n,p,\nu}} \sum_{\ell=k}^{n} \binom{\ell}{k} \binom{n}{\ell}^{\nu} p^{\ell}(1-p)^{n-\ell}.$$

3.3.4 The empty urns problem

Consider the problem of assigning ℓ indistinguishable balls uniformly at random into n distinguishable urns. Let Y_i be 1 if urn i is empty and 0 otherwise, and $X = \sum_{i=1}^{n} Y_i$ be the Bernoulli sum counting the total number of empty urns.

Through a straightforward counting argument, it can be shown that there are $\binom{\ell+n-1}{n}$ ways to distribute ℓ indistinguishable balls into n distinguishable urns and therefore

$$\begin{aligned} Pr(Y_i = 1) &= \frac{\#\text{ ways to distribute } m \text{ balls into } n-1 \text{ urns}}{\#\text{ ways to distribute } m \text{ balls into } n \text{ urns}} \\ &= \frac{\binom{n+\ell-2}{\ell}}{\binom{n+\ell-1}{\ell}} \\ &= \frac{n-1}{\ell+n-1}. \end{aligned}$$

By the same argument, for a subset $\{i_1, \ldots, i_m\}$ of $[n]$,

$$Pr(Y_{i_1} = 1, Y_{i_2} = 1, \ldots, Y_{i_m} = 1) = \frac{(n-1)(n-2)\cdots(n-m)}{(\ell+n-1)(\ell+n-2)\cdots(\ell+n-m)} = \frac{[n-1]_m}{[\ell+n-1]_m}.$$

The kth moment of X is

$$E(X^k) = \sum_{m=1}^{k} \sum_{\{i_1,\ldots,i_m\}\subseteq[n]} S(k,m)E(Y_{i_1}\cdots Y_{i_m})$$
$$= \sum_{m=1}^{k} S(k,m)\binom{n}{m}\frac{[n-1]_m}{[\ell+n-1]_m},$$

the kth central moment (from Proposition 3.2.2)

$$E((X-\mu)^k) = (-\mu)^k + \sum_{\ell=1}^{k}\binom{k}{\ell}(-\mu)^{k-\ell}\sum_{m=1}^{\ell} S(\ell,m)\binom{n}{m}\frac{[n-1]_m}{[\ell+n-1]_m},$$

where $\mu = n\frac{n-1}{\ell+n-1}$. The kth factorial moment has a particularly simple representation as

$$E([X]_k) = \frac{[n]_k[n-1]_k}{[\ell+n-1]_k}.$$

Matching moments shows X of the urn problem to have a Hypergeometric distribution with parameters $(n, n-1, \ell+n-1)$.

3.3.5 The matching problem

The matching problem dates back to de Montmort (1713) and is the problem of taking n paired elements, randomly permuting the first elements over all pairs, then letting X be the number of correctly matched pairs after the random permutation. Examples are n letters matched correctly to n envelopes, couples separated at a dance and dance partners formed by randomly assigning one of each sex to the pair, and so on. The random variable X can be expressed as a sum of Bernoulli random variables taking value 1 when a correct match occurs and zero otherwise.

More abstractly, let $f : [n] \to [n]$ be a permutation on $[n]$ and let X denote the number of fixed points of f (i.e., the number of $i \in [n]$ for which $f(i) = i$). If f is picked uniformly at random from the set of all permutations on $[n]$, denoted $Sym(n)$, then the distribution of X can be shown to tend to $Poisson(1)$ as $n \to \infty$. We do that by expressing X as a sum of Bernoullis and then examining and comparing moments.

Let Y_i be the Bernoulli random variable recording if $f(i) = i$. As f is chosen uniformly at random from $Sym(n)$,

$$Pr(Y_i = 1) = \frac{1}{n}$$

for all $i = 1, \ldots, n$.

Fix $m \leq n$ and consider an $m-$subset $\{i_1, \ldots, i_m\} \subseteq [n]$. The probability that all $\{i_1, \ldots, i_m\}$ are fixed points of f is $(n-m)!/n!$. This is because if $f(i_j) = i_j$ for all

$j \in \{1, \ldots, m\}$, one must only consider how to assign the other $(n-m)$ points so that f is a bijection. There are $(n-m)!$ ways to do this as $|Sym(n-m)| = (n-m)!$. Therefore,

$$Pr(Y_{i_1} = 1, \ldots, Y_{i_m} = 1) = \frac{(n-m)!}{n!}.$$

By Proposition 3.2.1, for $k \leq n$

$$\begin{aligned}
E(X^k) &= \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\} \subseteq [n]} Pr(Y_{i_1} = 1, \ldots, Y_{i_m} = 1) \\
&= \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\} \subseteq [n]} \frac{(n-m)!}{n!} \\
&= \sum_{m=1}^{k} S(k,m) \binom{n}{m} \frac{(n-m)!}{n!} \\
&= \sum_{m=1}^{k} \frac{S(k,m)}{m!} \\
&= \sum_{m=1}^{k} S_2(k,m) \\
&= B_k,
\end{aligned}$$

where B_k is the $k-$th *Bell number*, the number of ways to partition a set of size k into a family of nonempty, unlabelled, pairwise disjoint subsets. On the other hand, for $k > n$, the inner sum vanishes whenever $m > n$ and hence

$$\begin{aligned}
E(X^k) &= \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\} \subseteq [n]} Pr(Y_{i_1} = 1, \ldots, Y_{i_m} = 1) \\
&= \sum_{m=1}^{n} S(k,m) \sum_{\{i_1,\ldots,i_m\} \subseteq [n]} Pr(Y_{i_1} = 1, \ldots, Y_{i_m} = 1) \\
&= \sum_{m=1}^{n} \frac{S(k,m)}{m!} \\
&= \sum_{m=1}^{n} S_2(k,m) \\
&= B_k - \sum_{m=n+1}^{k} S_2(k,m),
\end{aligned}$$

which can be interpreted as the number of ways to partition a set of size k into at most n classes.

For $k \leq n$, the $k-$th factorial moments of X is given by

$$\begin{aligned} E([X]_k) &= k! \sum_{\{i_1,\ldots,i_k\}\subseteq[n]} Pr(Y_{i_1} = 1, \ldots, Y_{i_k} = 1) \\ &= k! \frac{(n-k)!}{n!} \binom{n}{k} \\ &= 1. \end{aligned}$$

Therefore, the factorial moment generating function of X is given by

$$H_X(s) = \sum_{k=0}^{n} \frac{s^k}{k!},$$

which by Equation 3.4 gives us that the probability generating function of X is

$$\begin{aligned} G_X(s) &= H_X(s-1) \\ &= \sum_{k=0}^{n} \frac{(s-1)^k}{k!} \\ &= \sum_{\ell=0}^{n} s^\ell \sum_{k=\ell}^{n} \binom{k}{\ell} \frac{(-1)^{k-\ell}}{k!}, \end{aligned}$$

which provides us with another method for deriving the probability distribution of X *without* using the principle of inclusion and exclusion or counting derangements in permutations.

Now, the exponential generating function for B_k given by (e.g., see Stanley (2011) p. 74)

$$\sum_{k\geq 0} B_k \frac{t^k}{k!} = e^{e^t - 1}$$

is a special case (viz., $\lambda = 1$) of the moment generating function for a Poisson(λ) random variable W

$$M_W(t) = \sum_{k\geq 0} E(W^k) \frac{t^k}{k!} = e^{\lambda(e^t - 1)}.$$

For $k \leq n$, the moments of X match those of $W \sim$ Poisson(1); as $n \to \infty$, the moment generating functions agree and X converges to a Poisson(1) random variable.

3.3.6 The Poisson limit of a binomial

Proposition 3.2.1 can also be used to provide a novel proof that a binomial(n, p) random variable approaches a Poisson(λ) with $np \to \lambda$ as $n \to \infty$.

For $X = \sum_{i=1}^{n} Y_i$, each Y_i is independent, identically distributed Bernoulli(p) random variable and, as seen earlier, the kth moment

$$\begin{aligned} E(X^k) &= \sum_{m=1}^{k} S(k,m) \binom{n}{m} p^m \\ &= \sum_{m=1}^{k} S(k,m) \binom{n}{m} \left(\frac{np}{n}\right)^m \\ &= \sum_{m=1}^{k} \frac{S(k,m)}{m!} \left(\frac{[n]_m}{n^m}\right) (np)^m . \end{aligned}$$

As $n \to \infty$, the ratio $\left(\frac{[n]_m}{n^m}\right) \to 1$, $np \to \lambda$, and

$$E(X^k) \to \sum_{m=1}^{k} \frac{S(k,m)}{m!} \lambda^m = \sum_{m=1}^{k} S_2(k,m) \lambda^m$$

which is the kth moment of a Poisson(λ) random variable expressed as a Touchard polynomial in λ (e.g., see Riordan (1937)). It follows that as $n \to \infty$, $X \sim$ *binomial*(n,p) converges to a Poisson(λ) with $\lambda = \lim_{n\to\infty} np$.

When $p = \frac{1}{n}$, then binomial X converges to Poisson(1) and its kth moment is the Bell number B_k, as in the matching problem of Section 3.3.5.

3.4 Counts more generally

Focus has been on Bernoulli sums that arise naturally as counts of events. In this section, we consider any "count random variable" N to be that having support on *any* subset of the extended natural numbers $\mathbb{N}_0$. Previous results are extended to N by matching it to a Bernoulli sum X constructed from the upper tail probabilities of N.

Proposition 3.4.1. *Let N denote any discrete random variable on $\mathbb{N}_0$. Consider the Bernoulli random variable Y_i indicating whether $N \geq i$ or not; that is*

$$Y_i = \begin{cases} 1 & \text{if } N \geq i, \\ 0 & \text{otherwise.} \end{cases}$$

Then the Bernoulli sum $X = \sum_{i=1}^{\infty} Y_i$ has the same distribution as N.

Proof.

$$\begin{aligned} Pr(X = x) &= Pr\left(\sum_{i=1}^{\infty} Y_i = x\right) \\ &= Pr([Y_i = 1, \forall i : i \leq x] \cap [Y_i = 0, \forall i : i > x]) \\ &= Pr(N = x). \end{aligned}$$

□

The moments of N are identified with those of X and the Bernoullis Y_i defined above. The results for the moments and factorial moments now follow.

Proposition 3.4.2. *For any discrete random variable N on $\mathbb{N}_0$, the kth moment of N is*

$$E(N^k) = \sum_{m=1}^{k} S(k,m) \sum_{M \geq m} \binom{M-1}{m-1} Pr(N \geq M),$$

and the kth factorial moment is

$$E([N]_k) = k! \sum_{M \geq k} \binom{M-1}{k-1} Pr(N \geq M).$$

Proof. From Proposition 3.2.1,

$$E(N^k) = \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\} \subset \mathbb{N}} Pr(Y_{i_1} = 1, \ldots, Y_{i_m} = 1)$$

Now,

$$\begin{aligned}
\sum_{\{i_1,\ldots,i_m\} \subset \mathbb{N}} Pr(Y_{i_1} = 1, \ldots, Y_{i_m} = 1) &= \sum_{\{i_1,\ldots,i_m\} \subset \mathbb{N}} Pr(\max(i_1, \ldots, i_m) \leq N) \\
&= \sum_{M=1}^{\infty} \sum_{\{i_1,\ldots,i_{m-1}\} \subseteq M-1} Pr(N \geq M) \\
&= \sum_{M=1}^{\infty} \binom{M-1}{m-1} Pr(N \geq M).
\end{aligned}$$

Therefore,

$$E(N^k) = \sum_{m=1}^{k} S(k,m) \sum_{M=1}^{\infty} \binom{M-1}{m-1} Pr(N \geq M).$$

As for the factorial moment,

$$\begin{aligned}
m! \sum_{M=m}^{\infty} \binom{M-1}{m-1} Pr(N \geq M) &= m! \sum_{M \geq m} \sum_{\ell \geq M} Pr(N = \ell) \binom{M-1}{m-1} \\
&= m! \sum_{\ell \geq m} Pr(N = \ell) \sum_{M=1}^{\ell} \binom{M-1}{m-1} \\
&= m! \sum_{\ell \geq m} Pr(N = \ell) \binom{\ell}{m} \\
&= m! \sum_{\ell \geq m} \frac{[\ell]_m}{m!} Pr(N = \ell) \\
&= \sum_{\ell \geq m} [\ell]_m Pr(N = \ell) \\
&= \sum_{\ell \geq 0} [\ell]_m Pr(N = \ell) \\
&= E([N]_m) \quad \text{by definition.}
\end{aligned}$$

□

Note that an expression for $E(N^k)$ has also recently been derived by (Chakraborti et al., 2019, eq. 10), namely

$$E(N^k) = \sum_{i=0}^{\infty} ((i+1)^k - i^k) Pr(N > i).$$

Chakraborti et al. (2019) claim their formulation to be the first for $E(N^k)$ expressed in terms of the upper tail probability of N; if so, then Proposition 3.4.2 may provide the second for $E(N^k)$ and the first for $E([N]_k)$. These results are best appreciated whenever the cumulative distribution function of N has form allowing simplification, especially when multiplied by binomial coefficients.

The remainder of this section explores application of Proposition 3.4.2 to several familiar cases.

3.4.1 Geometric distribution

Suppose interest lay in the number X of tosses of a coin at which the first head occurs; X is a geometric(p) distribution with p being the probability of heads ($p = 0.5$ for a fair coin). Surprisingly, the random variable X can be written as a Bernoulli sum $X = \sum_{i=1}^{\infty} Y_i$ for suitably defined Bernoulli Y_is, allowing the previous results to be applied.

The representation is as follows. Take $(Z_i)_{i \geq 1}$ to be the sequence of independent Bernoulli(p) random variables representing the sequence of potential coin tosses ($Z_i = 1$

for heads and zero otherwise). Let N be the index of the first $Z_i = 1$ in the sequence, that is

$$N = \min\{i : Z_i = 1\}.$$

Consider now the Bernoulli random variables formed from the upper tail of the distribution of the index N:

$$Y_i = \begin{cases} 1 & \text{if } N \geq i \\ 0 & \text{otherwise} \end{cases}$$

The Bernoulli sum $X = \sum_{i=1}^{\infty} Y_i$ is the number of coin tosses (Z_is) that have occurred when the first head ($Z_i = 1$) appears in the sequence. While $X = N$, each provides a different way of looking at the same random variable.

Given the Bernoulli sequence $(Y_i)_{i\geq 1}$ and any $m-$set of indices $\{i_1, \ldots, i_m\} \subseteq [n]$, the joint probability of m Y_is can be written as

$$\begin{aligned} Pr(Y_{i_1} = 1, \ldots, Y_{i_m} = 1) &= Pr(N \geq i_1, \ldots, N \geq i_m) \\ &= Pr(N \geq \max i_1, \ldots, i_m). \end{aligned}$$

Then by Proposition 3.2.3

$$\begin{aligned} E(X^k) &= \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\}\subset\mathbb{N}} E(Y_{i_1} \cdots Y_{i_m}) \\ &= \sum_{m=1}^{k} S(k,m) \sum_{\{i_1,\ldots,i_m\}\subset\mathbb{N}} Pr(N \geq \max i_1, \ldots, i_m) \\ &= \sum_{m=1}^{k} S(k,m) \sum_{M\geq 1} \sum_{\{i_1,\ldots,i_{m-1}\}\subseteq[M-1]} Pr(N \geq M) \quad (M \text{ being the max index}) \\ &= \sum_{m=1}^{k} S(k,m) \sum_{M\geq 1} \binom{M-1}{m-1} [Pr(N > M) + Pr(N = M)] \\ &= \sum_{m=1}^{k} S(k,m) \sum_{M\geq 1} \binom{M-1}{m-1} [(1-p)^M + (1-p)^{M-1}p] \\ &= \sum_{m=1}^{k} S(k,m) \sum_{M\geq 1} \binom{M-1}{m-1} (1-p)^{M-1}. \end{aligned}$$

For an indeterminate y and $k \in \mathbb{N}$,

$$\sum_{n\geq 0} \binom{n}{k} y^n = \frac{y^k}{(1-y)^{k+1}}$$

from which it follows that

$$\sum_{M\geq 1} \binom{M-1}{m-1} (1-p)^{M-1} = \frac{(1-p)^{m-1}}{p^m}$$

giving the expression for the kth moment to be

$$E(X^k) = \sum_{m=1}^{k} S(k,m) \frac{(1-p)^{m-1}}{p^m}.$$

Similarly, the kth factorial moment has the simpler expression

$$E([X]_k) = \frac{(1-p)^{k-1}}{p^k}.$$

We note that since the probability generating function of X has the closed form expression

$$G_X(s) = \frac{p}{1-s(1-p)},$$

by Equation 3.4, the factorial moment generating function is given by

$$\begin{aligned} H_X(s) &= \frac{p}{1-(1+s)(1-p)} \\ &= p \sum_{\ell \geq 0} [(1+s)(1-p)]^{\ell}. \end{aligned}$$

3.4.2 Poisson distribution

In Section 3.3.6 the tidy expression

$$E(N^k) = \sum_{m=1}^{k} S_2(k,m) \lambda^m$$

for the kth moment of $N \sim Poisson(\lambda)$ appeared. The proof of this result given by Riordan (1937) is recursive. It can also be proved by direct application of Proposition 3.4.2 as follows:

$$\begin{aligned} E(N^k) &= \sum_{m=1}^{k} S(k,m) \sum_{M=m}^{\infty} \binom{M-1}{m-1} Pr(N \geq M) \\ &= \sum_{m=1}^{k} S(k,m) \sum_{M=m}^{\infty} \binom{M-1}{m-1} \sum_{\ell=M}^{\infty} e^{-\lambda} \frac{\lambda^{\ell}}{\ell!} \\ &= e^{-\lambda} \sum_{m=1}^{k} S(k,m) \sum_{\ell=m}^{\infty} \frac{\lambda^{\ell}}{\ell!} \sum_{M=1}^{\ell} \binom{M-1}{m-1} \quad \text{by reorganizing the sums} \\ &= e^{-\lambda} \sum_{m=1}^{k} S_2(k,m) m! \sum_{\ell=m}^{\infty} \frac{\lambda^{\ell}}{\ell!} \binom{\ell}{m} \end{aligned}$$

as $\sum_{M=1}^{\ell} \binom{M-1}{m-1} = \binom{\ell}{m}$,

$$
\begin{aligned}
&= e^{-\lambda} \sum_{m=1}^{k} S_2(k,m)\lambda^m \sum_{\ell=m}^{\infty} \frac{\lambda^{(\ell-m)}}{(\ell-m)!} \\
&= e^{-\lambda} \sum_{m=1}^{k} S_2(k,m)\lambda^m e^{\lambda} \\
&= \sum_{m=1}^{k} S_2(k,m)\lambda^m.
\end{aligned}
$$

Following the same route as for $E(N^k)$, the kth factorial moment of $N \sim Poisson(\lambda)$ has the even simpler expression:

$$E([N]_k) = \lambda^k.$$

3.4.3 Ideal soliton distribution

The ideal soliton distribution appears in erasure correcting codes, a subject of coding theory concerned with using information redundancy to accommodate for missing data. Luby (2002) introduced the ideal and robust soliton distributions as initial models for the transmission of messages over a noisy medium (e.g., see MacKay et al., 2003, Chapter 50). In this section, we apply our theory to derive the moments of the ideal soliton distribution. To the best of our knowledge, this has not been described before in the literature.

For an integer r with $r \geq 2$, we say that N follows the ideal soliton distribution, $soliton(r)$, when

$$
\begin{aligned}
Pr(N = 1) &= \frac{1}{r}, \\
Pr(N = i) &= \frac{1}{i(i-1)},
\end{aligned}
$$

for $i \in \{2, \ldots, r\}$ and is zero otherwise. It can be shown by induction that

$$\sum_{i=2}^{\ell} \frac{1}{i(i-1)} = \frac{\ell - 1}{\ell}.$$

From this it follows that

$$Pr(N \leq \ell) = \frac{1}{r} + \frac{\ell - 1}{\ell} \quad \text{for } \ell = 1, 2, \ldots r$$

and

$$Pr(N \geq \ell) = \begin{cases} 1 & \text{when } \ell = 1 \\ \frac{r-1}{r} - \frac{\ell-2}{\ell-1} & \ell = 2, 3, \ldots, r. \end{cases}$$

To derive the moments and factorial moments, we need the following lemma.

Lemma 3.4.3. *For any $m, r \in \mathbb{N}$,*

$$\sum_{M=0}^{r} M[M]_m = (m+1)!\binom{r+1}{r-m-1} + m \cdot m!\binom{r+1}{r-m}.$$

Proof. First, an expression for the ordinary generating function $f(x) = \sum_{i\geq 0} i[i]_{m-2}\ x^j$ corresponding to the sequence $(i[i]_{m-2})_{i\geq 0}$ is determined. Multiplying by $\frac{1}{1-x}$ then gives the generating series for $(\sum_{i=0}^{r} i[i]_{m-2})_{r\geq 0}$.

Since $\sum_{i\geq 0} x^i = \frac{1}{1-x}$, differentiating with respect to x and then multiplying by x gives

$$\sum_{i\geq 0} ix^i = \frac{x}{(1-x)^2}.$$

Differentiating with respect to x, $(m-2)$ times, the left hand side becomes

$$\sum_{i\geq 0} i \times i(i-1)\cdots(i-m-1)\ x^{i-m-2} = \sum_{i\geq 0} i[i]_{m-2}\ x^{i-m-2}.$$

Applying the general Leibniz rule, the same derivative of right hand side is

$$\frac{x(m-1)!}{(1-x)^m} + (m-2)\frac{(m-2)!}{(1-x)^{m-1}}.$$

Multiplying both sides by $\dfrac{x^{m+2}}{(1-x)}$ gives

$$\sum_{r\geq 0}\sum_{i=0}^{r} i[i]_{m-2}x^r = \frac{x^{m+3}(m-1)!}{(1-x)^{m+1}} + (m-2)\frac{x^{m+2}(m-2)!}{(1-x)^m}.$$

Since

$$\frac{1}{(1-x)^k} = \sum_{n\geq 0}\binom{n+k-1}{n}x^n,$$

the claim immediately follows. □

Proposition 3.4.4. *Let N follow the soliton(r) distribution with $r \geq 2$. For $k \geq 1$, the moments of N are given by*

$$E(N^k) = H_r + \sum_{m=2}^{k} S(k,m)\left[\frac{r-1}{r}\binom{r}{m} - \binom{r-3}{m} - (m-2)^2\binom{r-3}{m-1}\right]$$

For $k \geq 2$, the $k-$th factorial moment of N is

$$E([N]_k) = k!\left[\frac{r-1}{r}\binom{r}{k} - \binom{r-3}{k} - (k-2)^2\binom{r-3}{k-1}\right].$$

Proof. Once more, by Proposition 3.2.1 and Proposition 3.2.6, we need to evaluate sums of the form

$$\sum_{M=m}^{r} \binom{M-1}{m-1} Pr(N \geq M).$$

If $m = 1$, this becomes

$$\begin{aligned}
\sum_{M=m}^{r} \binom{M-1}{m-1} Pr(N \geq M) &= \sum_{M=1}^{r} Pr(N \geq 1) \\
&= \sum_{M=1}^{r} \sum_{\ell=M}^{r} Pr(N = \ell) \\
&= \sum_{\ell=1}^{r} \sum_{M=1}^{\ell} Pr(N = \ell) \\
&= \sum_{\ell=1}^{r} \ell \; Pr(N = \ell) \\
&= \frac{1}{r} + \sum_{j=2}^{r} \frac{j}{j(j-1)} = H_r.
\end{aligned}$$

Otherwise, for $m \geq 2$,

$$\begin{aligned}
\sum_{M=m}^{r} \binom{M-1}{m-1} Pr(X \geq M) &= \sum_{M=m}^{r} \binom{M-1}{m-1} \left[\frac{r-1}{r} - \frac{M-2}{M-1}\right] \\
&= \frac{r-1}{r} \sum_{M=m}^{r} \binom{M-1}{m-1} - \sum_{M=m}^{r} \binom{M-1}{m-1} \frac{M-2}{M-1}.
\end{aligned}$$

By straightforward algebraic manipulation,

$$\sum_{M=m}^{r} \binom{M-1}{m-1} \frac{M-2}{M-1} = \frac{1}{(m-1)!} \sum_{M=m}^{r} (M-2)[M-2]_{m-2}.$$

By replacing m and r by $m-2$ and $r-2$, respectively, in Lemma 3.4.3 and dividing by $(m-1)!$ it follows that

$$\begin{aligned}
\frac{1}{(m-1)!} \sum_{M=m}^{r} (M-2)[M-2]_{m-2} &= \frac{1}{(m-1)!} \sum_{M=m}^{r-2} (M)[M]_{m-2} \\
&= \binom{r-1}{r-m-1} + (m-2)^2 \binom{r-1}{r-m}.
\end{aligned}$$

The result now follows from a straightforward application of Propositions 3.2.6 and 3.2.1. □

3.4.4 Benford distribution

Benford's distribution encapsulates the notion that in many real world settings, the leading digits in a numerical data set are more likely to be small. In particular, we say that D follows Benford's distribution if for a digit $d \in \{1, \ldots, 9\}$,

$$Pr(D = d) = \log_{10}(d+1) - \log_{10}(d).$$

Due to its telescoping nature, the complementary CDF of D is given by

$$Pr(D \geq d) = 1 - \log_{10}(d).$$

Therefore, by Proposition 3.4.2, the $k-$th moment of D is

$$\begin{aligned} E(D^k) &= \sum_{m=1}^{k} S(k,m) \sum_{M=m}^{9} \binom{M-1}{m-1} (1 - \log_{10}(M)) \\ &= \sum_{m=1}^{k} S(k,m) \left[\binom{9}{m} - \sum_{M=m}^{9} \binom{M-1}{m-1} \log_{10}(M) \right], \end{aligned}$$

and factorial moments with the form

$$E([D]_k) = k! \left[\binom{9}{k} - \sum_{M=k}^{9} \binom{M-1}{k-1} \log_{10}(M) \right],$$

for $k \leq 9$. Of course, the results above extend to any general base b by noting that $D_b \sim$ Benford(b) satisfies

$$\sum_{M=m}^{b-1} \binom{M-1}{m-1} Pr(D_b \geq M) = \left[\binom{b-1}{m} - \sum_{M=k}^{b-1} \binom{M-1}{k-1} \log_b(M) \right].$$

3.5 Discussion

A multinomial theorem for commutative idempotents (Proposition 3.1.1) led to new general expressions for the moments (including central and factorial) of a Bernoulli sum (e.g., Propositions 3.2.1 to 3.2.3) as well as corresponding generating functions. The general expressions depend on the determination of the expected product of subsets of the Bernoulli random variables. By evaluating these in particular cases a number of new expressions for moments and generating functions of many common distributions and classic problems.

The success of the approach in these examples mark it as potentially fruitful in more novel distributions and problems where this expectation might be more readily available. To that end, the representation of $\binom{X}{m}$ for random count X and fixed m (Proposition 3.2.6), and of $X!$ (Proposition 3.2.7), as the product of Bernoullis may also be more generally useful.

In other instances, the general representation of the various moments for a count variable N expressed in terms of the upper probability of that N (Proposition 3.4.2) may be valuable in yet other problems, as shown in the examples of Section 3.4.

The Bernoulli sum approach provides another tool for problems involving count data, particularly those where expectations of products of the individual Bernoulli random variables are easily accessed. The results presented here apply in Chapter 6 to derive novel expressions for the moments of clique counts when combined with the edge counts derived in Chapter 4. Additional directions for application of the theory are discussed in Chapter 6.

4

Clique covers

Before deriving the expressions for the moments of clique counts in a random graph using the Bernoulli sums framework of Chapter 3, we need to evaluate probabilities of the form

$$Pr(c_i \text{ is a clique} : i = 1, \ldots, m)$$

for a collection of subsets of $[n]$. Due to the independence and homogeneity assumptions of random graphs, it follows that

$$Pr(c_i \text{ is a clique} : i = 1, \ldots, m) = p^{e(c_1, \ldots, c_m)},$$

where p is the probability of an edge inclusion, and $e(\mathcal{C})$ is the number of edges induced by a set of cliques $\mathcal{C}$. The overarching goal of this chapter is to derive expressions for $e_r(\mathcal{C})$ the number of $r-$cliques contained in a collection of cliques $\mathcal{C}$ by presenting a connection between clique covers and intersecting families of sets through a special kind of partition, which translates some commonly studied objects from extremal set theory into the language of graph theory.

Section 4.3 introduces and illustrates these concepts using the three clique collection example of Figure 4.1. Section 4.4 then provides a more general treatment with formal definitions and proved results. The general Γ-partition is derived in Section 4.4.1 for any collection of subsets of $[m]$ and applied to clique collections. It is shown to be an orbit partition in Section 4.4.2 and its quotient graph defined. Support and signatures are formally defined in Section 4.4.3 and used to define different types of isomorphic graphs. Section 4.4.3 establishes some counting results on signatures as does Section 4.4.3 as they relate to subgraph connectedness. Section 4.4.3 ends with a generating function for the number of induced connected subgraphs of size k. Section 4.4.3 shows H induces a clique if, and only if, its support is an intersecting family; Theorem 4.4.13 provides the conditions for the clique to be maximal. Section 4.4.3 shows how the quotient graph, G/Γ, can be used to directly determine cliques and maximal cliques in the original graph union G and ends with some minor results on the number of maximal cliques and the clique number of G.

Section 4.5 uses the framework of Section 4.4 to finally get down to counting cliques. Results include expressions for the number of cliques containing any particular clique H, the number of cliques of size r, and, in Theorem 4.5.3, a generating function for clique counts in the graph union of m cliques. Theorem 4.5.3 is then applied to give a new expression for the number of r-cliques and for the number of edges induced by a collection of m cliques of size r. We apply the results with an application of the "hand-shaking lemma" to yield an expression for the number of edges induced by any collection of cliques.

4.1 Clique covers

Recall that for any graph $G = (V, E)$ and node subset $H \subseteq V$, the induced subgraph $G[H]$ has nodes H and those edges in E whose endpoints lie in H. A clique of size r is induced whenever $G[H]$ is a complete graph on r nodes. Allowing trivial cliques (i.e., $r = 1$ or $r = 2$), a collection of cliques $\mathcal{C} = \{c_1, \ldots, c_m\}$ can always be found (for some m) which

covers the graph G – in the sense that the graph union, $G[c_1] \cup G[c_2] \cup \cdots \cup G[c_m]$, of the induced subgraphs has the same vertex set as G.

Such a collection is called a *vertex clique cover* of G. When its cliques are also non-intersecting (i.e., $c_i \cap c_j = \emptyset \ \forall i \neq j$), then the collection will be called a *clique cover partition*, so as to clearly identify this special case.

A collection of cliques whose graph union contains all edges in G is called an *edge clique cover* (e.g., see Roberts, 1985). In what follows, interest lies in counting the number of cliques, of any specified size r, formed by the graph union over *any* of these clique covers, indeed over any *collection of cliques* of G.

Suppose the graph G has n nodes numbered 1 to n, so that the power set, $\mathcal{P}([n])$, of $[n] = \{1, \ldots, n\}$ identifies, by node indices, all possible induced subgraphs of G. For index set $i \subseteq [n]$, provided G is understood, the induced subgraph $G[i]$ may be more simply denoted by its index set i. A collection of cliques, then, is denoted by a *family of sets* $\mathcal{C} = \{c_1, \ldots, c_m\} \subseteq \mathcal{P}([n])$, provided each $c_j \in \mathcal{C}$ identifies a clique induced in G.

For example, suppose $n \geq 9$ and G contains three cliques $A := \{1, 2, 3, 5, 6\}, B := \{1, 2, 4, 7, 8\}$ and $C := \{1, 2, 3, 4, 9\}$, each of size 5. Then $\mathcal{C} = \{A, B, C\}$ is a collection of three size 5 cliques, being a *vertex clique cover* only if $n = 9$ (and not if $n > 9$). Its graph union is shown in Figure 4.1. It may, or may not, also be an *edge clique cover*, depending

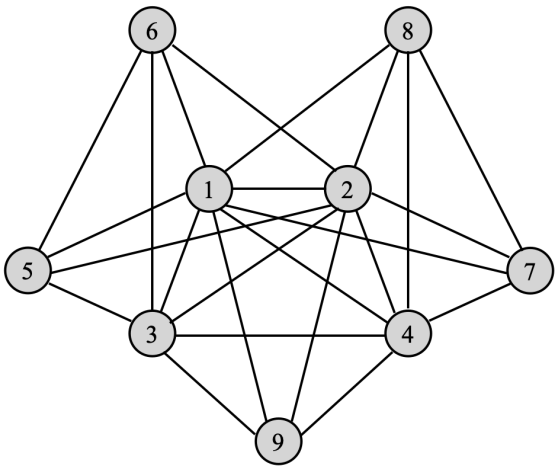


Figure 4.1: The graph union of $\mathcal{C} = \{A, B, C\}$ with $A := \{1, 2, 3, 5, 6\}, B := \{1, 2, 4, 7, 8\}$ and $C := \{1, 2, 3, 4, 9\}$. How many cliques are there of size $r = 1, 2, 3, \ldots$?

on whether, or not, the union contains *all* edges of G. It is *not* a *clique cover partition* because the intersection of at least one pair of A, B, and C is non-null (here all pairs intersect).

Our interest lies in determining the number of cliques of any size r in the union. When $r = 5$, there are exactly three 5-cliques, namely A, B and C. Consulting Figure 4.1, there are no cliques in the union of size $r \geq 6$, though this need not be true in general – e.g., were the 3-clique $D := \{4, 5, 6\}$ added to the collection $\mathcal{C}$, the 6-clique $\{1, 2, 3, 4, 5, 6\}$ would arise. For $r < 5$, the intersections of the cliques in $\mathcal{C}$ must also be considered. If all

intersections are null, then $\mathcal{C}$ would be *clique cover partition* of its union, and the number of cliques of size $r \leq 5$ would simply be the sum of the number of r-cliques within each clique of $\mathcal{C}$. But that is not the case here – e.g., the 3-clique $\{1, 2, 3\}$ appears in both A and C – so care is needed to avoid overcounting. Careful examination of Figure 4.1 will yield 15 cliques of size 4, 28 of size 3, and 24 of size 2.

Given a clique cover $\mathcal{C} = \{c_1, \ldots, c_m\}$ of a graph G, an expression for the number of cliques of size r in $G = \cup_{i=1}^{m} c_i$ can be had by applying the principle of inclusion and exclusion. This is done in Section 4.2.

A richer approach is to first form a partition of $G = \cup_{i=1}^{m} c_i$ based on index sets J, now consisting of the indices from $\{1, \ldots, m\}$ which identify the cliques in the collection (or cover) $\mathcal{C}$. That is, each partition cell is identified with one set $J \in \mathcal{P}([m])$; the set of graph node indices in cell J will be denoted $\Gamma_J \in \mathcal{P}([n])$ and the partition called a Γ-partition. The cardinality of Γ_J will be denoted γ_J. This is the primary approach introduced and explored in this chapter.

Subgraphs H of $G = \cup_{i=1}^{m} c_i$ will have nodes appearing in some cells Γ_J (for some $J \in \mathcal{P}([m])$) and not in others. The clique index cells J whose Γ_J contain nodes in the subgraph H will be called the *support* of H and the tuple containing the count of nodes of H in each Γ_J its *signature*. Whether H is connected, or is a clique of size r, or forms a maximal clique, can be determined using characteristics of its support and/or signature. This is shown by connecting these concepts to intersecting sets and *intersecting families* of sets (e.g., see Meyerowitz, 1995).

The Γ-partition is itself an *orbit* partition and hence an *equitable* partition. The quotient graph, G/Γ, which results compresses and contains all information needed to determining connected subgraphs, cliques, and maximal cliques on G.

We begin by considering the approach which does not rely on the partition.

4.2 Counting by Principle of Inclusion and Exclusion

As the example of Section 4.1 suggests, the key to clique counting over the graph union of a collection of cliques will be identifying the intersection of the various index sets. Unsurprisingly, then, our first approach to enumerating cliques makes use of the *Principle of Inclusion and Exclusion*.

This yields the following result for the count of the number of r-cliques in the union of an arbitrary collection of cliques.

Proposition 4.2.1. *Let $\mathcal{C} = \{c_1, \ldots, c_m\}$ be a collection of cliques. The total number of $r-$cliques that are induced by $\{c_1, \ldots, c_m\}$ is*

$$\sum_{J:\emptyset \neq J \subseteq \{1,\ldots,m\}} (-1)^{|J|+1} \binom{I_J}{r},$$

where $I_J := |\bigcap_{j \in J} c_j|$.

Proof. We count the number of $r-$cliques that are induced by at least one of the cliques in $\mathcal{C}$. Let $\binom{c_j}{r} := \{\{v_1, \ldots, v_r\} \subseteq c_j : v_1 \neq \cdots \neq v_r\}$ denote the set of $r-$cliques induced by the clique c_j. We will prove that for any nonempty $J \subseteq \{1, \ldots, m\}$,

$$\left|\bigcap_{j\in J} \binom{c_j}{r}\right| = \binom{I_J}{r},$$

by showing that

$$\bigcap_{j\in J} \binom{c_j}{r} = \binom{\bigcap_{j\in J} c_j}{r}.$$

If $\{v_1, \ldots, v_r\} \in \bigcap_{j\in J} \binom{c_j}{r}$, then $\{v_1, \ldots, v_r\} \subset c_j$ for all $j \in J$ and so

$$\{v_1, \ldots, v_r\} \in \binom{\bigcap_{j\in J} c_j}{r}.$$

Conversely, if $\{v_1, \ldots, v_r\} \in \binom{\bigcap_{j\in J} c_j}{r}$ then $\{v_1, \ldots, v_r\} \subset c_j$ for all $j \in J$. Therefore,

$$\{v_1, \ldots, v_r\} \in c_j,$$

for all $j \in J$ and the claim follows.

Therefore, the total number of $r-$cliques within A is $\left|\bigcup_{j\in\{1,\ldots,m\}} \binom{c_j}{r}\right|$. By Principle of Inclusion and Exclusion (Proposition 2.3.5),

$$\left|\bigcup_{j\in J} \binom{c_j}{r}\right| = \sum_{\emptyset\neq J\subseteq\{1,\ldots,m\}} (-1)^{|J|+1} \left|\bigcap_{j\in J} \binom{c_j}{r}\right| = \sum_{\emptyset\neq J\subseteq\{1,\ldots,m\}} (-1)^{|J|+1} \binom{I_J}{r},$$

as needed to be shown. □

This leads to an expression for the *total* number of typically interesting cliques (i.e., $r \geq 3$; non-trivial: no single edge, no single vertex, cliques):

Corollary 4.2.2. *Let $\{c_1, \ldots, c_m\}$ be a collection of cliques. The total number of non-trivial cliques that are contained within $\{c_1, \ldots, c_m\}$ is*

$$\sum_{J:\emptyset\neq J\subseteq\{1,\ldots,m\}} (-1)^{|J|+1} \left(2^{I_J} - \binom{I_J}{2}\right) - \left|\bigcup_{j=1}^{m} c_j\right| - 1.$$

Proof. By Proposition 4.2.1, the total number of cliques is given by

$$\begin{aligned}\sum_{r=0}^{\infty} \sum_{J:\emptyset\neq J\subseteq\{1,\ldots,m\}} (-1)^{|J|+1} \binom{I_J}{r} &= \sum_{J:\emptyset\neq J\subseteq\{1,\ldots,m\}} (-1)^{|J|+1} \sum_{r=0}^{\infty} \binom{I_J}{r} \\ &= \sum_{J:\emptyset\neq J\subseteq\{1,\ldots,m\}} (-1)^{|J|+1} \sum_{r=0}^{\infty} 2^{I_J},\end{aligned}$$

by the Binomial Theorem. Now, since there is only one 0−clique on a set of nodes, the 1-cliques correspond to the $\left|\bigcup_{j=1}^{m} c_j\right|$ vertices and 2-cliques is the number of edges,

$$\sum_{J:\emptyset\neq J\subseteq\{1,\ldots,m\}} (-1)^{|J|+1} \sum_{r=0}^{\infty} 2^{I_J} = \sum_{J:\emptyset\neq J\subseteq\{1,\ldots,m\}} (-1)^{|J|+1} \left(2^{I_J} - \binom{I_J}{2}\right) - \left|\bigcup_{j=1}^{m} c_j\right| - 1.$$

□

For example, let $r = 3$ and let $\mathcal{C} = \{c_1, c_2\}$ be a collection of the two triangles $c_1 = \{1, 2, 3\}$ and $c_2 = \{2, 3, 4\}$. If e_j is the number of edges induced by triangle j, then the total number of edges in the collection is given by

$$e_1 + e_2 - \binom{|c_1 \bigcap c_2|}{2} = \binom{|c_1|}{2} + \binom{|c_2|}{2} - \binom{2}{2} = 3 + 3 - 1 = 5$$

since $\binom{|c_1 \bigcap c_2|}{2}$ is the number of edges common to both c_1 and c_2 (one edge for every 2 vertices).

4.3 A partition framework

Consider again the example of Section 4.1, where the collection $\mathcal{C} = \{A, B, C\}$ consisting of the three 5-cliques $A = \{1, 2, 3, 5, 6\}, B = \{1, 2, 4, 7, 8\}$, and $C = \{1, 2, 3, 4, 9\}$ in some graph. Figure 4.3 shows the graph union over the cliques of $\mathcal{C}$.

Because various intersections of the cliques in $\mathcal{C}$ are important to identify, we introduce a separate notation to distinguish those subgraphs, of the graph union over $\mathcal{C}$, that *uniquely* appear in an intersection of specified cliques in $\mathcal{C}$ *but not* in any of the unspecified cliques. We call this partition Γ and its cells $\Gamma - sets$.

The set of indices is denoted by Γ with the specified cliques identified by subscript are shown in Figure 4.2 for a collection of cliques $\mathcal{C} = \{A, B, C\}$. The Γ sets partition its graph union while the indexing on Γ partitions the power set of $\mathcal{C}$, which we denote by $\{\emptyset, A, B, C, AB, AC, BC, ABC\}$.

For each cell in Γ, its cardinality is denoted by $\gamma = |\Gamma|$ – e.g., $\gamma_{AB} = |\Gamma_{AB}|$. Figure 4.3 shows the partition of the graph union of $\mathcal{C}$ from Figure 4.1 according to its Γ sets, as in Figure 4.1. The contents of each Γ set are easily read off from the graph, as shown. The γs are simply the cardinalities of the sets. For example, the cell Γ_{AB} contains no nodes from the collection because every element common to both A and B is also common to C.

The Γ-sets turn out to have useful properties related to cliques. From Figures 4.2 and 4.3, note that each original clique $A = \Gamma_A \cup \Gamma_{AB} \cup \Gamma_{AC} \cup \Gamma_{ABC}$, $B = \Gamma_B \cup \Gamma_{AB} \cup \Gamma_{BC} \cup \Gamma_{ABC}$, and $C = \Gamma_C \cup \Gamma_{AC} \cup \Gamma_{BC} \cup \Gamma_{ABC}$, is the union of Γ-sets whose subscript sets have a common intersection, namely A, B, or C. Moreover, the size of each clique is simply the sum of the corresponding γs. Similar results hold for the union of any two Γ-sets Γ_{J_1} and Γ_{J_2}. If the index sets are such that $J_1 \cap J_2 \neq \emptyset$, then the union $\Gamma_{J_1} \cup \Gamma_{J_2}$ forms a clique of size $\gamma_{J_1} + \gamma_{J_2}$; if $J_1 \cap J_2 = \emptyset$, then there is no clique spanning Γ_{J_1} and Γ_{J_2}.

$$
\begin{aligned}
\Gamma_A &= A \cap \overline{B} \cap \overline{C} \\
\Gamma_B &= \overline{A} \cap B \cap \overline{C} \\
\Gamma_C &= \overline{A} \cap \overline{B} \cap C \\
\Gamma_{AB} &= A \cap B \cap \overline{C} \\
\Gamma_{AC} &= A \cap \overline{B} \cap C \\
\Gamma_{BC} &= \overline{A} \cap B \cap C \\
\Gamma_{ABC} &= A \cap B \cap C \\
\Gamma_\phi &= \overline{A} \cap \overline{B} \cap \overline{C}
\end{aligned}
$$

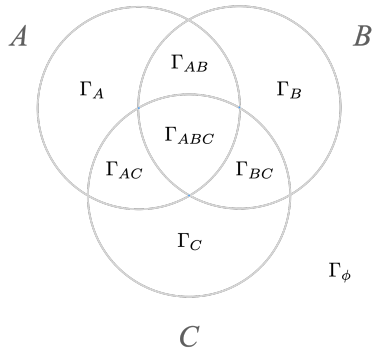
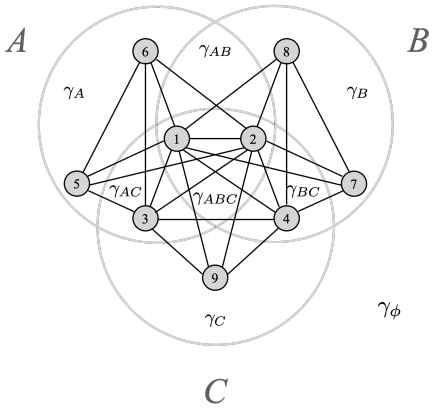


Figure 4.2: The Γ sets for $\mathcal{C} = \{A, B, C\}$.



$$
\begin{aligned}
\Gamma_A = \{5,6\} &\implies \gamma_A = 2 \\
\Gamma_B = \{7,8\} &\implies \gamma_B = 2 \\
\Gamma_C = \{9\} &\implies \gamma_C = 1 \\
\Gamma_{AB} = \emptyset &\implies \gamma_{AB} = 0 \\
\Gamma_{AC} = \{3\} &\implies \gamma_{AC} = 1 \\
\Gamma_{BC} = \{4\} &\implies \gamma_{BC} = 1 \\
\Gamma_{ABC} = \{1,2\} &\implies \gamma_{ABC} = 2 \\
\Gamma_\phi = \emptyset &\implies \gamma_\phi = 0
\end{aligned}
$$

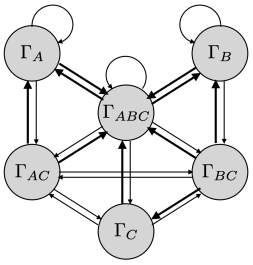
Figure 4.3: A partition of the graph union of $\mathcal{C} = \{A, B, C\}$ with $A = \{1, 2, 3, 5, 6\}, B = \{1, 2, 4, 7, 8\}$, and $C = \{1, 2, 3, 4, 9\}$ according to its Γ sets, together with their sizes γ. This is also a decomposition of [9] following Proposition 4.4.1.

4.3.1 An orbit partition

Consider any Γ-set in Figure 4.2 and the node numbers it contains in Figure 4.3. The node numbers within *any* Γ-set could be permuted without any change in the structure of the graph in Figure 4.3. These cells are called *orbits* and the partition an *orbit partition* (e.g., see Lerner, 2005, Definition 9.3.4 and Proposition 9.3.5). That the Γ-sets, as defined above, form an orbit partition in general will be proved in Proposition 4.4.5.

For any equitable partition (e.g., an orbit partition), $\Gamma = \{\Gamma_1, \ldots, \Gamma_m\}$, of the vertex set of a graph G, a directed multi- (or weighted) *quotient* graph can be defined having nodes Γ_i and b_{ij} edges (or edge weights) from Γ_i to Γ_j where b_{ij} is the number of neighbours in Γ_j of every vertex in Γ_i – called the *quotient* of G modulo Γ and denoted G/Γ (e.g., Lerner, 2005, Definition 9.3.2).

For the graph union of Figure 4.3, the partition $\Gamma = \{\Gamma_A, \Gamma_B, \Gamma_C, \Gamma_{AC}, \Gamma_{BC}, \Gamma_{ABC}\}$ produces the quotient graph and matrix $\mathbf{B} = [b_{ij}]$ shown in Figure 4.4. This graph can



orbit	Γ_A	Γ_B	Γ_C	Γ_{AC}	Γ_{BC}	Γ_{ABC}
Γ_A	1	0	0	1	0	2
Γ_B	0	1	0	0	1	2
Γ_C	0	0	0	1	1	2
Γ_{AC}	2	0	1	0	1	2
Γ_{BC}	0	2	1	1	0	2
Γ_{ABC}	2	2	1	1	1	1

$= \mathbf{B}$

Figure 4.4: The quotient graph of the graph union of $\mathcal{C}$ modulo Γ and its edge weight matrix $\mathbf{B} = [b_{ij}]$. Edges are shown with width proportional to their weight in $\mathbf{B}$.

be thought of as a compression of the original graph union. As such, some information will be lost, but much remains. Its (weighted) adjacency matrix and graph are enough to determine several properties of the graph union (e.g., see Godsil & Royle, 2001)

including the path distances between nodes, the graph diameter, and a partial spectral decomposition – the characteristic roots of $\mathbf{B}$ are a subset of those of the adjacency matrix $\mathbf{A}$ of the graph union.

4.3.2 Equivalent graphs

The orbit partition, Γ, has particular features. For example, if *any* node u in Γ_{J_1} connects to k nodes in Γ_{J_2}, then *every* node in Γ_{J_1} connects to the *same* k nodes in Γ_{J_2}, and vice versa. And, since permuting node numbers in any Γ-set does not change the graph, if $u \sim v$ for any $u \in \Gamma_{J_1}$ and any $v \in \Gamma_{J_2}$, then *all* nodes in Γ_{J_1} connect to *all* nodes in Γ_{J_2}. It follows, then, that the nodes of $\Gamma_{J_1} \cup \Gamma_{J_2}$ form a clique of size $\gamma_{J_1} + \gamma_{J_2}$ (since each of Γ_{J_1} and Γ_{J_2} also form cliques).

This suggests that the orbit partition, given by the Γ-sets, provides a structure to identify sets of equivalent subgraphs which may, or may not, form a clique. If we choose an ordering of the orbits, say $(\Gamma_A, \Gamma_B, \Gamma_C, \Gamma_{AB}, \Gamma_{AC}, \Gamma_{BC}, \Gamma_{ABC})$, then a unique tuple of the counts of nodes from each orbit identifies a set of subgraphs which are isomorphic to one another (under node permutation within each orbit). For example, both $\{1, 3, 6\}$ and $\{2, 3, 6\}$ share the tuple (1, 0, 0, 0, 1, 0, 1), but $\{1, 2, 3\}$ with tuple (0, 0, 0, 0, 1, 0, 2) is a unique subgraph (under permutation within orbits). Each of these forms a 3-clique. The size of the subgraph is the sum of the tuple elements and the number of subgraphs the tuple represents is the product of the size of the orbit choose that element of the tuple (e.g., here $\binom{2}{1} \times \binom{1}{1} \times \binom{2}{1} = 4$ in total, the remaining two being $\{1, 3, 5\}$ and $\{2, 3, 5\}$).

The index sets associated with each non-zero tuple element determine whether subgraphs produced by the tuple are also a clique. For example, the tuple (1, 0, 0, 0, 1, 0, 1) takes nodes from Γ_A, Γ_{AC}, and Γ_{ABC} whose index sets are $\{A\}$, $\{A, C\}$, and $\{A, B, C\}$. The intersection of these index sets is non-null, and every subgraph induced by this tuple is a clique of size equal the sum of its elements. In contrast, the index sets $\{B\}$ and $\{C\}$ corresponding to the tuple (0, 2, 0, 1, 0, 0, 0) have null intersection and this tuple's induced graph does not form a clique. A tuple induces a clique if, and only if, the intersection of any two of its index sets is non-null – this is formally established by Proposition 4.4.2. The tuple associated with a clique we call its *signature*, and cliques having the same signature are of the same *type*.

4.3.3 Maximal cliques

Consider the problem of finding all maximal cliques which contain some specific clique. For example, from Figure 4.3, find all maximal cliques which contain the 2-clique $\{1, 2\}$. These are

(i) $M_1 = \{1, 2, 3, 5, 6\} = \Gamma_A \cup \Gamma_{AC} \cup \Gamma_{ABC}$,

(ii) $M_2 = \{1, 2, 4, 7, 8\} = \Gamma_B \cup \Gamma_{BC} \cup \Gamma_{ABC}$, and

(iii) $M_3 = \{1, 2, 3, 4, 9\} = \Gamma_C \cup \Gamma_{AC} \cup \Gamma_{BC} \cup \Gamma_{ABC}$.

The maximal cliques help enumerate the total number of cliques which contain a specified clique by identifying the nodes which can be added to expand that clique. In the case of $\{1, 2\}$, M_1 provides three additional nodes (viz., 3, 5, and 6) and so $2^3 - 1 = 7$ larger cliques containing $\{1, 2\}$. The same holds for M_2 and M_3, but care must be taken for double counting. The total number of cliques containing $\{1, 2\}$ (including itself) is expressed in terms of its maximal cliques as

$$\sum_{i=1}^{3} \left(2^{|M_i| - |\{1,2\}|} - 1\right) - \sum_{\{i,j\} \subset \{1,2,3\}} \left(2^{|M_i \cap M_j| - |\{1,2\}|} - 1\right) + \left(2^{|M_1 \cap M_2 \cap M_3| - |\{1,2\}|} - 1\right) + 1$$

$$= (7 + 7 + 7) - \left((2^0 - 1) + (2^1 - 1) + (2^1 - 1)\right) + (1 - 1) + 1 = 20,$$

where the last summand 1 corresponds to the edge $\{1,2\}$ on its own. A general expression for this count is given in Proposition 4.5.1.

We might call the union $M_1 \cup M_2 \cup M_3$, of its maximal cliques, the *clique extension* of $\{1,2\}$ within the cover, or simply the *clique extent* of $\{1,2\}$. In this case, the extent of $\{1,2\}$ is the entire cover but this is not generally the case (e.g., the extent of $\{5,6\}$ is simply the set $A = \{1,2,3,5,6\}$). More generally, if the intersection of all cliques in the collection is non-null, then the clique extension of any node (or clique) in that intersection will generate the entire cover. In a social network context, for example, such individuals (or cliques) might be deemed to be highly influential in the entire cover – wherever they are located in the cover, those having larger clique extents might be regarded as more influential than those having smaller ones.

4.3.4 Intersecting families

Reading off the set of subscripts from the Γ-sets defining each of the maximal cliques, M_1, M_2, M_3, respectively, gives the sets:

(i) $\mathcal{F}_1 = \{\{A\}, \{A,C\}, \{A,B,C\}\}$,

(ii) $\mathcal{F}_2 = \{\{B\}, \{B,C\}, \{A,B,C\}\}$,

(iii) $\mathcal{F}_3 = \{\{C\}, \{A,C\}, \{B,C\}, \{A,B,C\}\}$.

Each of these sets, $\mathcal{F}_i$, is called an *intersecting family* (e.g., see Meyerowitz, 1995), meaning that each set $\mathcal{F}_i$ is a subset of the power set of $\{A,B,C\}$ and that its elements have non-null pairwise intersection. When the context is clear, the notation for an intersecting family will be simplified, from a set of sets, to a set of the subscripts identifying the corresponding Γ-sets – so, the contents of $\mathcal{F}_1$ can be simplified to $\{A, AC, ABC\}$. Being cliques, each of the above families share the additional property that they have non-null intersection over all of their elements, not just pairwise.

When an intersecting family $\mathcal{F}$ is not a proper subset of any other intersecting family, it is called a *maximally intersecting family* (Meyerowitz, 1995). The family $\mathcal{F}_3 = \{C, AC, BC, ABC\}$ is a maximally intersecting family and corresponds to the maximal clique c_3. The families $\mathcal{F}_1$ and $\mathcal{F}_2$ are not; though, since $\Gamma_{AB} = \emptyset$, adding AB to each will make them maximally intersecting as well (i.e., equivalently, $M_1 = \Gamma_A \cup \Gamma_{AB} \cup \Gamma_{AC} \cup \Gamma_{ABC}$ and $M_2 = \Gamma_B \cup \Gamma_{AB} \cup \Gamma_{BC} \cup \Gamma_{ABC}$). Note that maximal intersecting families do not necessarily produce maximal cliques. For example, the only other maximal intersecting family here is $\mathcal{F}_4 = \{AB, AC, BC, ABC\}$ corresponds to the 4-clique $\{1,2,3,4\}$ which is not maximal. Necessary and sufficient conditions for an intersecting family to determine a maximal clique are given in Theorem 4.4.13 provides the necessary and sufficient conditions for a clique to be maximal. Intersecting families have interesting structure (e.g., see Meyerowitz, 1995).

We call a collection of sets $\mathcal{F}$ *path intersecting* if, for any $A, B \in \mathcal{F}$ there exists a sequence sets $J_1, J_2, \ldots, J_\ell$ in $\mathcal{F}$ from $A = J_1$ to $B = J_\ell$ having that $J_j \cap J_{j+1} \neq \emptyset$ for all

$j = 1, 2, \ldots, \ell - 1$. The set collection $\{\{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}\}$ is path intersecting, but is not an intersecting family.

In Section 4.4.2, the sequence of index sets of nodes along any path in the quotient graph is shown to be path intersecting and the set of index sets from any clique to be an intersecting family.

4.4 The general approach

By example, a number of results were illustrated in Section 4.3 relating the properties of a particular partition of the vertex set of the graph union of a collection of cliques to the cliques within the union. In this section, we show more generally that this kind of partition is a link between cliques in the union of a clique collection and certain intersecting families of sets. This link allows for a nuanced enumeration of several clique counting problems on these graphs, including total number of cliques, maximal cliques, maximum cliques and cliques containing any specific subset of interest. Finally, as with the example of Section 4.3, we establish that this kind partition is an orbit partition, hence capturing salient features of the original graph.

The example clique collection of Section 4.3 had three maximal 5-cliques as its elements – while possibly desirable, this is not necessary. In this section, general results for cliques in the graph union of *any* collection of cliques are derived (i.e., each of any size, including possibly as a single edge). We begin with the general construction of a vertex partition of the graph union which permits a deeper examination of all cliques through intersecting families derived from that partition. Maximal intersecting families will be shown to correspond to the largest cliques obtainable from particular sub-collections of cliques.

4.4.1 The partition

The general construction of the partition, and its cells, are defined in Proposition 4.4.1. Here, for any set $J \subseteq [m]$, its set complement is with respect to $[m]$ and is denoted as $\overline{J} := [m] \setminus J$.

Proposition 4.4.1. *For any $m \geq 1$, given a sequence $(A_i)_{i=1}^m$ of subsets of $[n] = \cup_{i=1}^m A_i$, the family of sets given by*

$$\Gamma := \left\{ \bigcap_{i \in J} A_i \setminus \left(\bigcup_{i \in \overline{J}} A_i \right) : J \subseteq [m] \right\} := \{\Gamma_J : J \subseteq [m]\}$$

is a partition of $[n]$.

Moreover, for any $i \in [m]$,

$$A_i = \bigcup_{J \subseteq [m]:\ i \in J} \Gamma_J.$$

Proof. First, we show that

$$\bigcup_{J\subseteq[m]} \Gamma_J = \bigcup_{J\subseteq[m]} \left[\bigcap_{i\in J} A_i \setminus \left(\bigcup_{i\in\overline{J}} A_i\right)\right] = [n].$$

For every $J \subseteq [m]$,

$$\Gamma_J = \left[\bigcap_{i\in J} A_i \setminus \left(\bigcup_{i\in\overline{J}} A_i\right)\right] \subseteq [n],$$

as each $A_i \subseteq [n]$. To see the reverse inclusion, fix *any* choice $x \in \cup_{i=1}^m A_i = [n]$ and let $J_x := \{i : x \in A_i\} \subseteq [m]$ denote the set of all indices i with $x \in A_i$, and its complement in $[m]$ as $\overline{J_x} = ([m] \setminus J_x)$. Now $x \in [n]$ appears in at least one A_i, since $\cup_{i=1}^m A_i = [n]$, so it follows that $x \in \cap_{i\in J_x} A_i$ and $x \notin \cup_{i\in\overline{J_x}} A_i$. Thus,

$$x \in \left[\bigcap_{i\in J_x} A_i \setminus \left(\bigcup_{i\in\overline{J_x}} A_i\right)\right] = \Gamma_{J_x}$$

for *any* $x \in [n]$, and hence

$$[n] = \bigcup_{x\in[n]} \Gamma_{J_x} = \bigcup_{J\subseteq[m]} \Gamma_J.$$

It remains only to show that the intersection of any two distinct non-null members of Γ is empty – the proof is by contradiction. Let $J, H \subseteq [m]$ be distinct, respectively producing

$$\Gamma_J = \left[\bigcap_{i\in J} A_i \setminus \left(\bigcup_{i\notin J} A_i\right)\right] \quad \text{and} \quad \Gamma_H = \left[\bigcap_{i\in H} A_i \setminus \left(\bigcup_{i\notin H} A_i\right)\right]$$

as members in Γ. Suppose $x \in \Gamma_J \cap \Gamma_H \neq \emptyset$, then $x \in \Gamma_J \implies x \in A_i \ \forall i \in J$ and $x \in \Gamma_H \implies x \in A_i \ \forall i \in H$. Since J and H are distinct, there exists some $k \in J \setminus H$ for which $x \in \Gamma_J$ appears in A_k. Now $k \notin H$ means $k \in \overline{H}$ and hence A_k appears in the union $\cup_{i\notin H} A_i$ being removed from $\cap_{i\in H} A_i$ in the definition of Γ_H. Therefore $x \notin \Gamma_H$ and, so, $x \notin \Gamma_J \cap \Gamma_H$, a contradiction. It follows that Γ_J and Γ_H are disjoint, whenever $J \neq H$ and hence that the sets of Γ form a partition of their union, $[n]$.

Finally, for any $i \in [m]$, it remains only to show that the original sets A_i are the union of those Γ-sets, Γ_J, whose index set J contains i. That is,

$$A_i = \bigcup_{J\subseteq[m]:\ i\in J} \Gamma_J.$$

If $i \in J$, then $\Gamma_J = \left[\bigcap_{j\in J} A_j \setminus \bigcup_{j\in\overline{J}} A_j\right]$ intersects A_i, and hence $\Gamma_J \subseteq A_i$ whenever $i \in J$. It follows, then, that

$$\bigcup_{J\subseteq[m]:\ i\in J} \Gamma_J \subseteq A_i.$$

Conversely, for every $x \in A_i$, then $i \in J_x$ and

$$x \in \left[\bigcap_{j \in J_x} A_j \setminus \bigcup_{j \in \overline{J_x}} A_j \right] = \Gamma_{J_x} \subseteq \bigcup_{J \subseteq [m]:\ i \in J} \Gamma_J.$$

So $A_i \subseteq \bigcup_{J \subseteq [m]:\ i \in J} \Gamma_J \subseteq A_i$, and it follows that $A_i = \bigcup_{J \subseteq [m]:\ i \in J} \Gamma_J$. □

We will call a partition produced as in Proposition 4.4.1, a Γ-*partition* and note that it will be peculiar to the sets A_i from which it is constructed.

Applied to a clique collection

For a collection of cliques $\mathcal{C} = \{c_1, \ldots, c_m\}$, defined by index sets $c_j \subset [n]$, with graph union $\bigcup_{j=1}^m c_j = [n]$, Proposition 4.4.1 provides a general means to find Γ-sets, namely as $(\Gamma_J)_{J \subseteq [m]}$ with

$$\Gamma_J = \left(\bigcap_{j \in J} c_j \right) \cap \left(\bigcap_{j \notin J} \overline{c_j} \right)$$

where complement is with respect to $[n]$. That is, each cell Γ_J is the set of vertices common to all c_j for all $j \in J$ and absent from every c_j for which $j \notin J$. Again, the cardinality of Γ_J is denoted as $\gamma_J = |\Gamma_J|$.

The Γ-partition provides an equivalence relation on nodes $u, v \in [n]$ via the indices of those cliques which contain u or v – namely, $J_u = \{j \in [m] : u \in c_j\}$ and $J_v = \{j \in [m] : v \in c_j\}$. The nodes u and v are equivalent, $u \equiv v$, if, and only if, $J_u = J_v$; that is, u and v are in the same Γ-set.

The Γ-partition can also be used directly to infer some properties of the graph union. For example, as in Section 4.3.2, the adjacency of nodes in the graph union is related to the intersection of those Γ-sets which contain them:

Proposition 4.4.2. *Let u and v be two nodes in the graph union, $\bigcup_{j=1}^m c_j$, of the clique collection $\mathcal{C} = \{c_1, c_2, \ldots, c_m\}$. If $u \in \Gamma_{J_u}$ and $v \in \Gamma_{J_v}$, then $u \sim v$ if, and only if, $J_u \cap J_v \neq \emptyset$.*

Proof. We note that $u \sim v$ if, and only if, for some $j \in [m]$, $u \in c_j$ and $v \in c_j$, which is equivalent to $J_u \cap J_v \neq \emptyset$. □

It follows, for example, that $u \sim v$ for every pair of nodes $u, v \in \Gamma_J$ (for any $J \subseteq [m]$). Moreover, the cardinalities, γ_J, determine the degree of every vertex in Γ_J. Proposition 4.4.3 establishes that all nodes in a cell of Γ have the same degree.

Proposition 4.4.3. *For a non-null set $J \subset [m]$, every vertex in Γ_J has degree d_J where*

$$d_J = \sum_{I \subseteq [m]\ :\ I \cap J \neq \emptyset} \gamma_I \qquad - 1.$$

Proof. If $u \in \Gamma_J$, then $u \sim v$ if, and only if,

$$v \in \bigcup_{I \subseteq [m] \,:\, I \cap J \neq \emptyset} \Gamma_I = \bigcup_{j \in J} c_j,$$

with $v \neq u$. Therefore, the degree of u is

$$\begin{aligned} deg(u) &= \left| \bigcup_{j \in J} c_j \right| - 1 \\ &= \left| \bigcup_{j \in J} \left(\bigcup_{I : j \in I} \Gamma_I \right) \right| - 1 \\ &= \sum_{I : j \in I, \text{ for some } j \in J} \gamma_I - 1. \end{aligned}$$

□

Note that different clique collections having the same graph-union produce different Γ-partitions, these being peculiar to the particular cliques in the collection. The cliques of the collection in Section 4.3, for example, were all of size 5; had they all been of size 3 the same graph union of (now many more) cliques in the collection would be the same but the resulting Γ-sets would be different.

The special case that the collection consists of exactly m cliques of size r, as in Section 4.3, can also be determined from the cardinalities, γ_J. For Γ to have been formed from a collection of m distinct $r-$cliques, the following must hold:

$$\sum_{J \subseteq [m]} \gamma_J = n, \qquad \text{...for the graph union to have } n \text{ nodes}$$

$$\sum_{J \subseteq [m] \,:\, j \in J} \gamma_J = r \qquad \text{...for each } c_j \text{ to have } r \text{ nodes}$$

$$\sum_{J \subseteq [m] \,:\, \{j,k\} \subseteq J} \gamma_J < r \qquad \text{...to ensure distinct cliques: when } j \neq k,\ c_j \neq c_k.$$

Of these, only the last may not be self-evident; it follows from:

Proposition 4.4.4. *Let $\mathcal{C} = \{c_1, \ldots, c_m\}$ be a collection of $r-$cliques and fix $I \subseteq [m]$. Then $\{c_i : i \in I\}$ consists of a single clique if, and only if,*

$$\sum_{J : I \subseteq J} \gamma_J = r.$$

Proof. By Proposition 4.4.1

$$\bigcap_{i \in I} c_i = \bigcap_{i \in I} \bigcup_{J \subseteq [m] : i \in J} \Gamma_J = \bigcup_{J : I \subseteq J} \Gamma_J$$

So, $|\cap_{i \in I} c_i| = |\bigcup_{J : I \subseteq J} \Gamma_J| = \sum_{J : I \subseteq J} \gamma_J$. Since all c_i are $r-$sets, their intersection is an $r-$set if, and only if, they are all equal. □

4.4.2 The general Γ-quotient graph

In light of Proposition 4.4.2, Γ is an *equitable partition* (e.g., see Godsil & Royle, 2001; Lerner, 2005) – the number of neighbours in Γ_H of vertex $u \in \Gamma_J$ depends only on the choice of H and J. In fact, Γ is an orbit partition induced by a group of automorphisms of H.

Proposition 4.4.5. *The partition* $(\Gamma)_{\emptyset \neq J \subseteq [m]}$ *is an orbit partition.*

Proof. For a nonempty $J \subseteq [m]$, let π_J be any permutation of the elements of Γ_J. Let $\pi : V \to V$ be the extension of the π_J to V, where $\pi(i) = i$ for all $i \notin J$. It immediately follows that the orbits of π are the cells of Γ and, by Proposition 4.4.2, that π is an automorphism of V. □

Each Γ_J cell has an $n \times 1$ characteristic vector $\mathbf{c}_J$ having value 1 in row i if vertex i is in Γ_J, and 0 otherwise, so that $\mathbf{c}_J{}^{\mathsf{T}}\mathbf{c}_J = \gamma_J$. The characteristic matrix $\mathbf{C}$ is formed with columns $\mathbf{c}_J$ placed in order of the Γ_Js of the partition Γ. If $\mathbf{A}$ is the adjacency matrix of the graph union, G, the matrix $\mathbf{B} = (\mathbf{C}^{\mathsf{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathsf{T}}\mathbf{A}\mathbf{C}$ determines the structure of the quotient graph of G modulo Γ (e.g., see Godsil & Royle, 2001, Lemma 9.3.1, p. 196).

4.4.3 Type equivalent graphs

For the example of Figure 4.3, Section 4.3.2, introduced the *type* of a subgraph H of the graph union of $\mathcal{C} = \{c_1, \ldots, c_m\}$ associated with its Γ-partition and identified by a *signature*, namely, the tuple of the counts of nodes from H appearing in each cell of the partition. In this section, these ideas are formalized to provide a more nuanced sense of equivalent graphs in the context of a Γ-partition of the graph union $G = \cup_{j=1}^{m} c_j$.

For subgraph H of G, the *signature* of H defined by the Γ-partition of $\mathcal{C}$ is the *function* $f_H : \mathcal{P}([m]) \to \mathbb{N}_0$ defined as $f_H(J) = |H \cap \Gamma_J|$ for all $J \subseteq [m]$. Note that this is defined for any subgraph H, not necessarily only cliques H. Two subgraphs H_1 and H_2 are said to be of the same *type*, or to be *type-isomorphic*, if, and only if, they have identical signatures (i.e., $f_{H_1} = f_{H_2}$). Finally, the *support* of H (or of f_H) is the set of all subsets J of $[m]$ for which $f_H(J) > 0$; we write the support as $Supp(H) = \{J : J \subseteq [m] \text{ and } f_H(J) > 0\}$, or as $Supp(f_H)$ when emphasizing the signature. Note also that all of these are predicated on the particular clique collection $\mathcal{C}$ and its associated Γ-partition.

For example, consider the clique collection of Figure 4.3 and the subgraphs $H_1 = \{1, 2, 3, 4\}$, $H_2 = \{1, 2, 3, 5\}$, $H_3 = \{1, 2, 3, 6\}$, and $H_4 = \{1, 2, 3, 5, 6\}$. The first three are graph isomorphic to each other and the complete graph, K_4 while H_4 is isomorphic to K_5. In contrast only H_2 and H_3 are type isomorphic; H_1 has a different signature (and support), while H_4 shares the same support as H_2 and H_3 but is of a different type.

Because it differs from the usual graph equivalence, the notion of type could be of interest whenever the node labels, or the cliques defining the collection, carry additional meaning.

Γ-signatures

This section develops a number of counting results pertaining types of subgraphs (as defined by signature) from any specific clique collection.

The number of different types of induced subgraphs is easily captured by the cell sizes of the partition:

Proposition 4.4.6. *The number of distinct signatures for the Γ-partition of a collection of m cliques is*

$$\prod_{J\in\mathcal{P}([m])} (\gamma_J + 1).$$

Proof. A function $f : \mathcal{P}([m]) \to \mathbb{N}_0$ is a signature if, and only if, $|f(J)| \leq \gamma_J$. Thus, there are $\gamma_J + 1$ choices for every $J \in \mathcal{P}([m])$. □

Proposition 4.4.7. *For any signature f_H, the number of signatures having the same support, $Supp(H)$, is*

$$\prod_{J\in Supp(H)} \gamma_J.$$

Proof. For signatures f_{H_1} and f_{H_2} to have the same support, they must have the same Γ-cells, Γ_J for $J \in Supp(H_1) = Supp(H_2)$, and each signature can have values $1, \ldots, \gamma_J$ for the Jth cell. The total possible is therefore $\prod_{J\in Supp(H)} \gamma_J$. □

Proposition 4.4.8. *Let $f : \mathcal{P}([m]) \to \mathbb{N}_0$. The number of induced subgraphs having signature f in the graph union of the clique collection $\{c_1, \ldots c_m\}$ is*

$$\prod_{J\in Supp(f)} \binom{\gamma_J}{f(J)}.$$

Proof. The signature is invariant to the choice of nodes within each Γ-cell – provided the same number of nodes from each cell is chosen, the signature is the same. Each cell has γ_j nodes giving

$$\prod_{J\in Supp(f)} \binom{\gamma_J}{f(J)}$$

choices for type-isomorphic induced subgraphs. □

Connected subgraphs

The Γ-signature of an induced graph also tells whether it is *connected*. This is captured by the notion of a *path-intersecting* collection of sets defined in Section 4.3.4.

Proposition 4.4.9. *A subgraph H of the graph union over a clique collection is connected if, and only if, its support is path-intersecting.*

Proof. Since, f_H is defined by the Γ-partition of $\mathcal{C}$, every node must appear in exactly one set J of $Supp(H)$. Moreover, any pair of nodes $u, v \in H$ appearing in the same set $J \in Supp(H)$ are connected by construction of the partition. So, we need only consider nodes u and v which lie in different sets of the support.

Suppose $Supp(H)$ is path-intersecting. Then for any pair of nodes $u, v \in H$, which appear in different subsets $J_u, J_v \in Supp(H)$, a sequence of sets $J_{w_1}, J_{w_2}, \ldots, J_{w_\ell}$ can be found in $Supp(H)$ such that $J_u = J_{w_1}$, $J_v = J_{w_\ell}$, and $J_{w_i} \cap J_{w_{i+1}} \neq \emptyset$ for all $i = 1, \ldots, (\ell-1)$. From Proposition 4.4.2 $w_i \sim w_{i+1}$ for all $i = 1, \ldots, (\ell - 1)$, $u = w_1 \to w_2 \to \cdots \to w_\ell = v$, is a path from u to v in H, and so the subgraph H is connected.

Conversely, suppose H is connected. Every pair of nodes u, v appearing in separate sets J_u and J_v of $Supp(H)$ have a path connecting them in H. By the construction of Γ, this path can be chosen to be $u = w_1 \to w_2 \to \cdots \to w_\ell = v$ such that each w_i comes from a different J_i in $Supp(H)$. Again, by Proposition 4.4.2, $w_i \sim w_{i+1}$ implies $J_i \cap J_{i+1} \neq \emptyset$, and hence that $\{J_1, \ldots, J_\ell\}$ is path-intersecting. This holds for any $u, v \in H$ and hence any $J_u, J_v \in Supp(H)$, implying that it holds for the whole of $Supp(H)$. It follows that $Supp(H)$ is path-intersecting. □

Proposition 4.4.10. *Let $\mathcal{I}_P$ be the set of all path-intersecting collections of non-empty cells from the Γ-partition of a clique collection $\mathcal{C}$. The number of distinct signatures that induce a connected subgraph in the graph union over $\mathcal{C}$ is*

$$\sum_{\mathcal{F} \in \mathcal{I}_P} \prod_{J \in \mathcal{F}} \gamma_J.$$

Proof. Proposition 4.4.9 states that for a subgraph H to be connected, its support must be path-intersecting; Proposition 4.4.9 determines the number of distinct signatures having the same support. Together they give the result. □

It follows that the number of induced *disconnected* subgraphs is

$$\prod_{J \in \mathcal{P}([m])} (\gamma_J + 1) - \sum_{\mathcal{F} \in \mathcal{I}_P} \prod_{J \in \mathcal{F}} \gamma_J$$

where $\mathcal{I}_P$ denotes the set of all path-intersecting collections of non-empty cells from Γ.

Proposition 4.4.11. *Let $\mathcal{I}_P$ be the set of all path-intersecting collections of non-empty cells from the Γ-partition of a clique collection $\mathcal{C}$. The number of induced connected subgraphs of size k in the graph union over $\mathcal{C}$ is the k-th coefficient of the generating series*

$$\sum_{\mathcal{F} \in \mathcal{I}_P} \prod_{J \in \mathcal{F}} \left[(1 + x)^{\gamma_J} - 1 \right].$$

Proof. By Proposition 4.4.9, every induced connected subgraph H is contained in some path-intersecting family. In fact, there exists a unique smallest path-intersecting family $\mathcal{F}_H := Supp(H)$ containing it. Clearly, the contribution of H to the generating function

$$\sum_{H'} x^{|V(H')|}$$

is x^k, where $|V(H)| = k$, and the sum is over all H' induced connected subgraphs whose support is $\mathcal{F}_H$.

Conversely, given a path-intersecting family $\mathcal{F}$, the induced connected subgraphs whose support is $\mathcal{F}$ are constructed uniquely by choosing $\alpha_J \geq 1$ nodes from Γ_J for every $J \in \mathcal{F}$. The generating series corresponding to this is

$$\prod_{J \in \mathcal{F}} \left[(1+x)^{\gamma_J} - 1\right].$$

□

Γ-support and cliques

The support of a subgraph H provides information on whether H is a clique and whether it is maximal.

Proposition 4.4.12. *For any clique collection $\mathcal{C} = \{c_1, \ldots, c_m\}$, the subgraph induced by H on the graph union $\bigcup_{j=1}^m c_m$, is a clique, if, and only if, is support, $Supp(H) = \{J : J \subseteq [m]$ and $\Gamma_J \cap H \neq \emptyset\}$ is an intersecting family.*

Proof. Suppose the induced graph on H is a clique. Fix two distinct sets $J_1, J_2 \in Supp(H)$. Let $u_1 \in \Gamma_{J_1} \cap H$ and $u_2 \in \Gamma_{J_2} \cap H$. Since $u_1 \sim u_2$, it must be that $u_1, u_2 \in c_j$ for some $j \in [m]$. Therefore, it follows that $j \in J_1$ and $j \in J_2$, by the definition of the partition $(\Gamma_J)_{J \subseteq [m]}$. Thus, $|J_1 \cap J_2| \geq 1$ and $Supp(H)$ is an intersecting family.

On the other hand, suppose that $Supp(H)$ is an intersecting family. Fix $u, v \in H$ and suppose that $u \in \Gamma_{J_u}$ and $v \in \Gamma_{J_v}$. Since $Supp(H)$ is an intersecting family, $|J_u \cap J_v| \geq 1$ and there exists some $j \in [m]$ with $j \in J_u \cap J_v$. Thus, we have that $u, v \in c_j$ and since c_j is a clique, $u \sim v$. □

So a subgraph H is connected if, and only if, its support is path-intersecting (Prop. 4.4.9) and is a clique if, and only if, its support is an intersecting family (Prop. 4.4.12). Theorem 4.4.13 gives necessary and sufficient conditions for H to be a *maximal* clique.

Theorem 4.4.13. *For any clique collection $\mathcal{C} = \{c_1, \ldots, c_m\}$, a clique induced by H on the graph union $\bigcup_{j=1}^m c_m$, is maximal, if, and only if, for any $J \subseteq [m]$,*

1. *$J \in Supp(H) \implies |\Gamma_J \cap H| = \gamma_J$, and*
2. *$J \notin Supp(H) \implies$ either $\Gamma_J = \emptyset$ or $\Gamma_J \neq \emptyset$ and $\{J\} \cup Supp(H)$ is not an intersecting family.*

Proof. First, to prove necessity, assume H is a maximal clique. For any $J \in Supp(H)$, at least one node in Γ_J is in H, and, so, connected to all other nodes in H. It follows from Proposition 4.4.2 that every node of Γ_J is also in H and hence $|\Gamma_J \cap H| = \gamma_J$ for

all $J \in Supp(H)$. To show statement 2 holds, suppose now that $J \notin Supp(H)$. Further, suppose that $\{J\} \cup Supp(H)$ is an intersecting family and so, by Proposition 4.4.12, that $\Gamma_J \cup J$ is a clique. Since $J \notin Supp(H)$, $\Gamma_J \cap H = \emptyset$ and, since H is maximal, it follows that $\Gamma_J = \emptyset$.

To prove sufficiency, assume H is a clique and that both statements 1 and 2 hold. By statement 1, all nodes in Γ_J for $J \in Supp(H)$ are in H and no nodes remain in Γ_J to increase H. Statement 2 ensures that no nodes exist in any Γ_J with $J \notin Supp(H)$ that could enlarge H and still be a clique. Hence, H is maximal. □

Statement 2 of Theorem 4.4.13 shows that, not only does a maximal clique have an intersecting family as its support (like all cliques), but that its intersecting family can only be expanded by sets $J \notin Supp(H)$ having no nodes in Γ_J.

The Γ-quotient graph and maximal cliques

Theorem 4.4.13 suggests that instead of considering intersecting families that are subsets of the entire power set, $\mathcal{P}([m])$, we need only those that are subsets of the support of the graph union $G = \cup_{j=1}^{m} c_j$, namely, $Supp(G) = \{J : J \subseteq [m] \text{ and } \Gamma_J \neq \emptyset\} \subseteq \mathcal{P}([m])$.

This effectively ignores empty cells of the Γ partition to focus on intersecting families formed from the index sets that define the nodes of the quotient graph G/Γ. The relevant families are intrinsic to the quotient graph. For example,

- any path on G/Γ corresponds to a path-intersecting set (Prop. 4.4.9),
- any clique on G/Γ determines an intersecting family and hence a clique on G, and
- any maximal clique on G/Γ gives a maximal intersecting family and, so, a maximal clique on G.

The last two points are proved below in Proposition 4.4.14.

Proposition 4.4.14. *If $\mathcal{F}$ is a nonempty intersecting family on $Supp(G)$, then the graph $H_{\mathcal{F}}$ induced by $\{\Gamma_J : J \in \mathcal{F}\}$ is a clique. Furthermore, $\mathcal{F}$ is a maximal intersecting family on $Supp(G)$ if, and only if, $H_{\mathcal{F}}$ is a maximal clique.*

Proof. The fact that $H_{\mathcal{F}}$ is a clique follows immediately from Proposition 4.4.2.

Suppose $\mathcal{F}$ is a maximal intersecting family on $Supp(G)$ and $H_{\mathcal{F}}$ is not a maximal clique. Then there exists some $u \in V(G)$ with u adjacent to all nodes in $H_{\mathcal{F}}$. Suppose $u \in \Gamma_{J_u}$, then Γ_{J_u} is nonempty and by Proposition 4.4.2, $\Gamma_{J_u} \cap J \neq \emptyset$ for all $J \in \mathcal{F}$. Therefore, either $\mathcal{F}$ is not a maximal intersecting family or $H_{\mathcal{F}}$ was not the subgraph induced by $\mathcal{F}$ – a contradiction.

The proof of the converse is almost identical.

□

Corollary 4.4.15. *If $\Gamma_J \neq \emptyset$ for all $\emptyset \neq J \subseteq [m]$, then every maximal intersecting family on $\mathcal{P}([m])$ induces a unique maximal clique in G.*

Proof. Suppose $\Gamma_J \neq \emptyset$ for all $\emptyset \neq J \subseteq [m]$. Then $Supp(G)$ is the set of all nonempty subsets of $\mathcal{P}([m])$. Therefore, by Proposition 4.4.14, each maximal intersecting family gives to a unique maximal clique. □

This means that the number of maximal cliques, $M(\mathcal{C})$, in G is equal to the number of maximal intersecting families on $Supp(G)$ which in turn is bounded above by the number of maximal intersecting families on $[m]$.

Corollary 4.4.16. *The number, $M(\mathcal{C})$, of maximal cliques in the graph union of $\mathcal{C} = \{c_1, \ldots, c_m\}$ is bounded above by $\lambda(m)$, the number of maximal intersecting families on $[m]$.*

Proof. By Theorem 4.4.13, each maximal intersecting family would correspond to at most one maximal clique in the graph union of the collection $\{c_1, \ldots, c_m\}$. Thus, $\lambda(m)$ is an upperbound for $M(\mathcal{C})$. □

Corollary 4.4.17. *The clique number of the graph union of the collection of cliques $\{c_1, \ldots, c_m\}$ is*

$$\max_{\mathcal{F} \in \mathcal{M}} \sum_{J \in \mathcal{F}} \gamma_J,$$

where $\mathcal{M}$ is the set of all maximal intersecting families on $[m]$.

Proof. By Theorem 4.4.13, a clique H is maximal if, and only if, its corresponding intersecting family $\mathcal{F}_H$ is only extendible by trivial elements and H uses all of the vertices in the cells Γ_J that contain members from H. Therefore, for every maximal intersecting family $\mathcal{F}$, there is a corresponding unique maximal clique H contained within the union of the cells $\{\Gamma_J : J \in F\}$.

Since the clique number is the maximum of the size of all maximal cliques in a graph, and each maximal clique has the form $\sum_{J \in \mathcal{F}} \gamma_J$ for some maximal intersecting family $\mathcal{F}$, the proof follows. □

To summarize, an intersecting family on $Supp(G)$ identifies a clique (Prop 4.4.12) and that clique is maximal if, and only if, its corresponding intersecting family is also maximal (Prop. 4.4.14). Whether an intersecting family, $\mathcal{F}$, is maximally intersecting can be determined from its cardinality, namely an intersecting $\mathcal{F} \subset [m]$ is a maximal intersecting family if, and only if, $|\mathcal{F}| = 2^{m-1}$ (e.g., Meyerowitz, 1995, Lemma 2.1); note that the intersecting family corresponding to an identified clique might have to be extended by adding subsets $J \in [m]$ having $\Gamma_J = \emptyset$ to achieve this cardinality (Thm. 4.4.13). Every such maximal intersecting family produces a unique maximal clique (Cor. 5.3.3). The number of such maximal cliques is bounded above by $\lambda(m)$, the number of maximally intersecting families on $[m]$ (Cor. 4.4.16). Unfortunately, $\lambda(m)$ is typically computationally intractable (e.g.,

see Brouwer et al., 2013) though is presently feasible on today's laptops for $m \leq 10$, for example. In the special case where $\gamma_J > 0$ for all $J \subseteq [m]$, every maximal intersecting family induces precisely one maximal clique so that the upper bound (Cor. 4.4.16) is achieved and $M(\mathcal{C}) = \lambda(m)$.

4.5 Counting cliques

For a family of sets $\mathcal{F}$, let $N(\mathcal{F}) := \sum_{J \in \mathcal{F}} \gamma_J$ denote the number of nodes in the sets contained in the family.

Given the collection of all maximal intersecting families on the support of G, we can apply the principle of inclusion and exclusion in the following manner.

Proposition 4.5.1. *Let H be a clique in the graph union of $\{c_1, \ldots, c_m\}$ and let $\mathcal{F}_H$ denote its support. Let $\mathcal{M}_H$ be the set of all maximal intersecting families $\mathcal{F}$ on Supp(G) that extend $\mathcal{F}_H$. The number of cliques that contain H in the graph union of $\{c_1, \ldots, c_m\}$ is*

$$1 + \sum_{\mathcal{J} \subseteq \mathcal{M}_H} (-1)^{|\mathcal{J}|+1} \left(2^{N\left(\bigcap_{\mathcal{F} \in \mathcal{J}} \mathcal{F}\right) - |H|} - 1\right).$$

Proof. Any clique that contains H would be a subclique of one of the maximal cliques that contain H. Therefore, by Theorem 4.4.13, it suffices to examine the collection $\mathcal{M}_H$ of maximal intersecting families that generate a unique maximal clique in the graph union of $\{c_1, \ldots, c_m\}$. If $\mathcal{F} \in \mathcal{M}_H$ corresponds to a maximal clique with $N(\mathcal{F})$ total nodes, then the selection of a nonempty subset from $(\cup_{J \in \mathcal{F}} \Gamma_J) \setminus H$ corresponds to a clique that properly contains H. This can be done in

$$\left(2^{N(\mathcal{F}) - |H|} - 1\right)$$

ways.

Since some cliques are subgraphs of several different maximal cliques, we use the principle of inclusion and exclusion and obtain

$$\sum_{\mathcal{J} \subseteq \mathcal{M}_H} (-1)^{|\mathcal{J}|+1} \left(2^{N\left(\bigcap_{\mathcal{F} \in \mathcal{J}} \mathcal{F}\right) - |H|} - 1\right)$$

cliques. However, this count does not include the clique H on its own and hence we add a 1. □

The proof of Proposition 4.5.1 relies on the fact that every clique is contained in some maximal clique. This observation can also be used to enumerate the total number of $r-$cliques in the graph union, by considering the collection of maximal cliques.

For instance, suppose we are interested in the number of triangles in the clique union from Figure 4.3. There are only three maximal cliques in this graph union, each of size 5. To enumerate the triangles in the graph union, then, simply count the triangles in each maximal clique, subtract the number of triangles common to the each of the intersections

of maximal cliques, and finally, add the number of triangles common to all three maximal cliques. This yields

$$\binom{5}{3}+\binom{5}{3}+\binom{5}{3}-\binom{2}{3}-\binom{3}{3}-\binom{3}{3}+\binom{2}{3}=28$$

triangles.

The following proposition follows the same logic to generalize to counting the number of cliques of any size r for any graph union of cliques. Reminiscent of Proposition 4.2.1, an advantage here is that the number of maximal cliques in the graph induced by the collection can be smaller than the number of cliques in the initial collection.

Proposition 4.5.2. *The number of $r-$cliques induced by the graph union of the cliques $\{c_1,\dots,c_m\}$ is*

$$\sum_{\mathcal{J}\subseteq\mathcal{M}}(-1)^{|\mathcal{J}|+1}\binom{N(\cap_{\mathcal{F}\in J}\mathcal{F})}{r},$$

where $\mathcal{M}$ is the collection of all maximal intersecting families $\mathcal{F}$ with $\gamma_J>0$ for all $J\in\mathcal{F}$.

Proof. As every clique is a subclique of a maximal clique, the induced graph by the maximal collection of cliques is the same as the induced graph by the collection $\{c_1,\dots,c_m\}$. Thus, the proof is exactly as in Proposition 4.2.1. □

A more subtle expression for clique counts is had by considering their signatures.

Theorem 4.5.3. *The generating function for clique counts induced by a collection $\{c_1,\dots,c_m\}$ is*

$$\Phi(\boldsymbol{x})=\sum_{\mathcal{F}\in\mathcal{I}_m}\prod_{J\in\mathcal{F}}\left[(1+x_J)^{\gamma_J}-1\right],$$

where $\mathcal{I}$ is the set of all intersecting families on $\mathcal{P}([m])$, and $\boldsymbol{x}$ is the vector $(x_J:J\in\mathcal{P}([m]))$.

Proof. A clique H is determined uniquely by its signature and the node labels. By Proposition 4.4.12, the support must be an intersecting family on on $Supp(G)$, and hence it is also an intersecting family on $\mathcal{P}([m])$.

For a cell J to contribute $\alpha_J\geq 1$ nodes to H is accomplished in $\binom{\gamma_J}{\alpha_J}$ ways, which corresponds to the coefficient of $x_J^{\alpha_J}$ in the generating series

$$\left[(1+x_J)^{\gamma_J}-1\right],$$

and the result follows. □

Extracting the coefficient of x^r in the generating function $\Phi(x_J\to x)$ in Theorem 4.5.3 yields the number of $r-$cliques as given in Corollary 4.5.4:

Corollary 4.5.4. *The number of $r-$cliques in the graph union of the clique collection $\{c_1, \ldots c_m\}$ is*

$$\sum_{\ell=1}^{r} \sum_{(\alpha_1,\ldots,\alpha_\ell)} \sum_{(J_1,\ldots,J_\ell)} \prod_{i=1}^{\ell} \binom{\gamma_{J_i}}{\alpha_i}$$

where $(J_1, \ldots, J_\ell)$ is an intersecting family on $Supp(G)$ of size ℓ with signature $(\alpha_1, \ldots, \alpha_\ell)$ being an integer composition of r having $1 \leq \alpha_i \leq \gamma_i$.

A third expression for the total number of cliques of any size, induced by the collection, can also be had by substituting $x_J = 1$ in the generating series in Theorem 4.5.3. The expression is given as Corollary 4.5.5:

Corollary 4.5.5. *The total number of cliques of size at least 1 induced by a collection $\{c_1, \ldots, c_m\}$ is*

$$\sum_{\mathcal{F} \in \mathcal{I}_m} \prod_{J \in \mathcal{F}} \left[2^{\gamma_J} - 1\right],$$

where $\mathcal{I}_m$ is the set of all intersecting families on $\mathcal{P}([m])$.

When $r = 2$, the interesting special case of the edge count is obtained (e.g., essential to edge count distributions for many random graph models, such as the Erdős-Rényi model):

Corollary 4.5.6. *The number of edges induced by the collection of r-cliques $\{c_1, \ldots, c_m\}$ is*

$$\sum_{J \subseteq [m]} \binom{\gamma_J}{2} + \frac{1}{2} \sum_{J \subseteq [m]} \gamma_J \sum_{I \neq J:\ |I \cap J| \geq 1} \gamma_I.$$

Alternatively, edges can also be enumerated via the degree sequences of the vertices in the various cells Γ_J. For every $J \subseteq [m]$, any two nodes within Γ_J have the same degree. For instance, if $u \in \Gamma_{\{k\}}$ for some $k \in [m]$, then it must be that $deg(u) = r - 1$ because $u \in c_k$ and $u \notin c_j$ for all $j \neq k$ by the definition of $\Gamma_{\{k\}}$. On the other extreme, if $u \in \Gamma_{[m]}$, then $u \in c_j$ for all $j \in [m]$ and hence u must be adjacent to all other nodes in G which are in at least one of the $\{c_1, \ldots, c_m\}$. Therefore,

$$deg(u) = n - \gamma_\emptyset - 1 = n - 1,$$

The "handshaking lemma" immediately gives the number of edges induced by the collection as below:

Proposition 4.5.7. *The number of edges induced by the collection of cliques $\{c_1, \ldots, c_m\}$ is*

$$\frac{1}{2} \sum_{J:\emptyset \neq J \subseteq [m]} \gamma_J \left(\sum_{I:\ |I \cap J| \geq 1} \gamma_I - 1 \right).$$

Proof. Follows immediately from Proposition 4.4.3 and the fact that number of edges in the graph is half the sum of the degrees in the graph. □

4.6 Discussion

In this chapter, connections were established and exploited between several graph-theoretic properties of clique covers, and notions of intersecting families on a special partition, the Γ-partition, of a graph $G = \cup_{i=1}^{m} c_i$ formed from a collection of $\mathcal{C} = \{c_1, \ldots, c_m\}$ of m cliques c_i.

The partition was formed using elements J of the power set $\mathcal{P}([m])$ from the m clique indices. The support of G is a subset of the power set, $Supp(G) \subseteq \mathcal{P}([m])$, and induces the partition Γ of $[n]$, which partitions the set of n distinct nodes in G. This Γ-partition frames the unique contributions to G from the various cliques of $\{c_1, \ldots, c_m\}$ via sets from the power set of $[m]$.

The quotient graph, G/Γ, induced by the Γ-partition succinctly captures the information provided by the collection of cliques. This description serves as a dictionary between graph-theoretic traits, such as cliques, maximal cliques, and connected induced subgraphs, and their extremal set theory counterparts (viz., intersecting families, maximal intersecting families and path-intersecting families, respectively). The natural connection between these objects facilitates determination of expressions for several classes of counting problems arising from clique covers.

Of course, the Γ-partition and quotient graph are determined by the particular cliques given as elements of the collection. Coarser partitions (those which produce fewer Γ_J cells) are preferred – ideally, the collection will consist of a minimal number of unique maximal cliques.

The techniques enabled by this partition approach may be adapted to enumerating graph components other than cliques (e.g., spanning trees or cycles). The methods also show promise in stochastic settings (e.g., since probability of particular graph configurations in Erdős-Rényi models is a function of edge counts, one can obtain the moments of clique counts on homogeneous Erdős-Rényi graphs using the techniques above). In fact, Chapter 6 revisits the expressions here to obtain the moments of clique counts in Erdős-Rényi random graphs by considering the edge counts induced by a collection of cliques.

Finally, from a topological standpoint, we note that, since the number of $(r+1)-$cliques in a graph is corresponds to the number of r-faces in the clique complex of the graph, these results can be extended to study the bounds on the number of generators in the $r-$th homology class of the clique complex (e.g., Kahle, 2009).

5

Johnson graphs

Our motivation for studying Johnson graphs originates with the problem of visual exploration of high dimensional data. If there are N observations on n variables in a statistical data set, then the data can be thought of as a set of N points in n dimensional real space. Such data are most naturally viewed in 2 and 3 dimensional scatterplots.

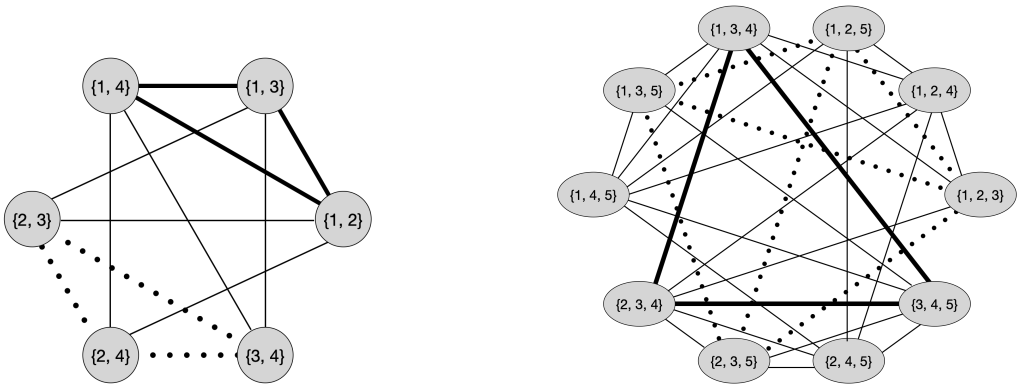
Hurley & Oldford (2011b) suggest interpreting a $J_n(2,1)$ Johnson graph as having 2-dimensional spaces as its nodes (defined by the indices of the variables) and, as its edges, 3-dimensional spaces defined by the union of the variables defining the adjacent nodes (e.g., see Figure 5.1(a) for $n = 4$). The Johnson graph provides a "navigation graph" for exploring high dimensional point clouds along lower dimensional trajectories. Following a path along the graph traverses from one $2d$-space to another along $3d$-transitions. Dynamic $3d$-scatterplot rotations from one $2d$-space to another have been used effectively in data analysis (e.g., see Oldford & Waddell, 2011; Waddell & Oldford, 2022), as have static displays of large numbers of $2d$-scatterplots laid out by following paths in the $J_n(2,1)$ with neighbouring scatterplots sharing a common axis (e.g., see Hofert & Oldford, 2017). Hurley & Oldford (2011b) also consider using paths along $J_n(m, m-1)$ Johnson graphs in conjunction with more complex visualizations to perceive structure in point clouds of dimension $m \geq 3$. Understanding the clique structure of the $J_n(m, m-1)$ for arbitrary m will help data analysts better understand, and make use of, lower dimensional regions of the full n-dimensional space of the data.

In what follows, we present novel, elementary proofs of the characterization of the structure of cliques, particularly maximal cliques in the $J_n(m, m-1)$ Johnson graph, which were previously alluded to without proof by Brouwer et al. (1989, p. 256) and Godsil & Meagher (2016, p. 113). Section 5.1 begins with some preliminary results for the $J_n(2,1)$ Johnson graph of interest in our motivating example. This section illustrates the logic, and provides the base case, for many of the more general results developed in Section 5.2 for the $J_n(m, m-1)$ Johnson graph. Results in both sections are derived without reference to the motivating example. These include the characterization of maximal cliques (there are only two types) of a $J_n(m, m-1)$, from which follows the clique number. Section 5.3 characterizes the nature of any r-clique, from which the clique partition number follows in Section 5.3.1.

Johnson "graphs are important because they enable us to translate many combinatorial problems about sets into graph theory" (Godsil & Royle, 2001, p. 9). Section 5.4 discusses the results of earlier sections in the context of the intersecting families of sets from extremal set theory (Gerbner & Patkós, 2018). Chapter 4 illustrated how intersecting families of sets can also be related to cliques in a clique cover. The last section ends with some discussion on the implications of the results in the context of the motivating example of statistical data analysis.

5.1 On the clique structure of $J_n(2,1)$

Figure 5.1 shows two examples of $J_n(m, m-1)$ graphs for (a) $n = 4$, $m = 2$, and (b) $n = 5$, $m = 3$. Identifying the nodes of $J_4, (2,1)$ as $v_1, \ldots, v_6$ in counter-clockwise order



(a) $J_4(2,1)$ (b) $J_5(3,2)$

Figure 5.1: Two separate Johnson $J_n(m, m-1)$ graphs with label sets $\nu(v)$ shown on each node v. Nodes are identified as v_1, v_2, ..., beginning from the right most node in each graph and from there in counter-clockwise order. Two maximal cliques are marked on each.

beginning from the rightmost node of Figure 5.1(a) gives $\nu(v_1) = \{1, 2\}$, $\nu(v_2) = \{1, 3\}$, ..., $\nu(v_6) = \{3, 4\}$. Similarly, beginning from the rightmost node of Figure 5.1(b), and moving counter-clockwise, yields label sets $\nu(v_1) = \{1, 2, 3\}$, $\nu(v_2) = \{1, 2, 4\}$, ..., $\nu(v_{10}) = \{3, 4, 5\}$ for $J_5(3, 2)$.

For any subgraph $H \subseteq G$, the *intersection of H* will refer to the set

$$S = \cap_{v \in V(H)} \nu(v) \text{ or, simply, } S = \cap_{v \in H} \nu(v),$$

the intersection of the label sets for the nodes $V(H)$ of H. For example, in the $J_4(2, 1)$ Johnson graph of Figure 5.1(a), consider the subgraphs H_1, H_2, and H_3 induced by vertex sets $V(H_1) = \{v_1, v_2, v_3\}$ (shown with thick edges in Fig. 5.1(a)), $V(H_2) = \{v_4, v_5, v_6\}$ (shown with thick dotted edges in Fig. 5.1(a)), and $V(H_3) = \{v_2, v_3, v_5, v_6\}$ (not shown), respectively. The intersection

- of H_1 is $S_1 = \cap_{v \in H_1} \nu(v) = \{1\}$,
- of H_2 is $S_2 = \cap_{v \in H_2} \nu(v) = \emptyset$, and
- of H_3 is $S_3 = \cap_{v \in H_3} \nu(v) = \emptyset$.

Of special interest is the relationship between this intersection set and cliques, H of G (e.g., H_1 and H_2 above; not H_3). For example, for a clique H to be *maximal* (i.e., no larger clique contains H) in a $J_n(2, 1)$ graph, it is easy to see that the size of its intersection set is either 1 (e.g., $|S_1| = 1$) or 0 (e.g., $|S_2| = 0$), as shown below in Lemma 5.1.1.

Lemma 5.1.1. *For $G = J_n(2, 1)$, the size of the intersection of node label sets on any maximal clique in G is at most 1.*

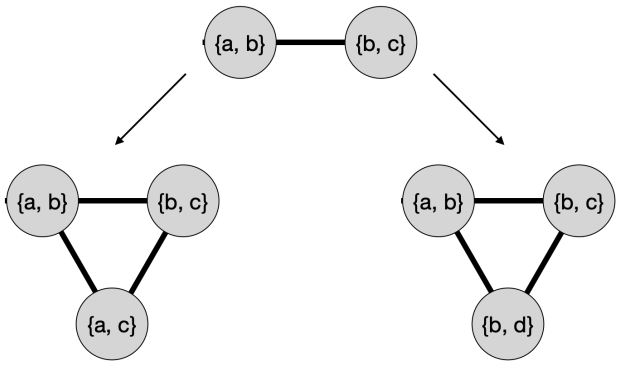
Proof. Let H be a maximal clique in $G = J_n(2,1)$ and $S = \cap_{v \in H} \nu(v)$. Since every node label set has 2 elements, and any two nodes intersect in exactly one element, it follows that $|S| \leq 1$. □

Both 0 and 1 are possible sizes for the intersection set S of a maximal clique in $G = J_n(2,1)$. Moreover, maximal cliques in $J_n(2,1)$ are only of two possible sizes according to the size of their intersection set H. This is shown in Lemma 5.1.2 below.

Lemma 5.1.2. *For any maximal clique H of $G = J_n(2,1)$, with intersection set S, for $n \geq 3$*

$$|H| = \begin{cases} 3 & \iff |S| = 0 \\ (n-1) & \iff |S| = 1 \end{cases}$$

Proof. Begin with the smallest non-trivial clique $H \subset G = J_n(2,1)$ of size two with vertex label sets $\nu(v_1) = \{a, b\}$ and $\nu(v_2) = \{b, c\}$, for distinct numbers $a, b, c \in [n]$. H is not maximal for it can be extended by a single node v_3 in one of only two possible ways, as shown below



for a fourth distinct number $d \in [n]$.

Consider first the leftmost triangle. Its intersection set $S = \{a, b\} \cap \{b, c\} \cap \{a, c\} = \emptyset$ and $|S| = 0$. No fourth node, v_4, can be added and the clique maintained, so this triangle is also a maximal clique. To see this, recall that any node v adjacent to both v_1 and v_2 must have label set of either $\{a, c\}$ or $\{b, d\}$, as shown above. But the first is already in the triangle and the second has no intersection with $\{a, c\} = \nu(v_3)$, meaning v_4 cannot be adjacent to v_3. So, this triangle, with $|S| = 0$, cannot be extended into a larger clique and, hence, must be maximal of size 3.

Now consider the rightmost triangle. This clique has intersection $S = \{b\}$ giving $|S| = 1$, but the clique is maximal only when $n = 3$. For $n > 4$, any node v with label $\nu(v) = \{b, e\}$, where $e \in [n]$ is distinct from a, b, c, and d, will be adjacent to *all* nodes in the triangle. There are exactly $(n-4)$ such choices remaining in $[n]$ to be paired with b in a label set. The clique can therefore grow maximally to size $(n-4)+3 = (n-1)$ with intersection set $S = \{b\}$ of size $|S| = 1$. □

Together, Lemmas 5.1.1 and 5.1.2 yield the clique number, the size of the maximum clique in $G = J_n(2,1)$, denoted $\omega(J_n(2,1))$, as follows:

Corollary 5.1.3. *The clique number of $G = J_n(2,1)$, for $n \geq 3$, is*

$$\omega(J_n(2,1)) = max\{n-1, 3\}.$$

According to Lemma 5.1.2, there are two different types of maximal cliques in $J_n(2,1)$. One set, say $\mathcal{M}_{min}$, contains maximal cliques $H \subset J_n(2,1)$ having *minimal intersection* set of size $|S| = 0$ (with each $|H| = 3$ for $H \in \mathcal{M}_{min}$); the other, say $\mathcal{M}_{max}$, contains those maximal cliques having *maximal intersection set* of size $|S| = 1$ (with each $|H| = n-1$ when $n \geq 4$ for every $H \in \mathcal{M}_{max}$). The number of maximal cliques of each type is $|\mathcal{M}_{min}| = \binom{n}{3}$ and $|\mathcal{M}_{max}| = \binom{n}{1}$.

Figure 5.1(b) suggests that similar results could exist for the Johnson graph $J_n(m, m-1)$ more generally. There, two different types of maximal cliques are shown for $G = J_5(3,2)$. One, shown with dotted line edges, is a 4-clique with intersection set $S = \{1,2,3\} \cap \{1,2,5\} \cap \{1,3,5\} \cap \{2,3,5\} = \emptyset$ of size $|S| = 0$ is in keeping with Lemma 5.1.2 identifying a maximal clique seemingly in $\mathcal{M}_{min}$ for a $J_5(3,2)$. Another, shown by thick solid line edges, is of size three and has intersection set $S = \{1,3,4\} \cap \{2,3,4\} \cap \{3,4,5\} = \{3,4\}$ of size $|S| = 2$ which is like $\mathcal{M}_{max}$ in that its intersection set also appears to be of maximum size, though this time 2 instead of 1. More generally, it turns out that maximal cliques in any Johnson graph $G = J_n(m, m-1)$ either have an intersection set of size $|S| = 0$ or $|S| = m-1$. This is proved as Theorem 5.2.4 in the next section.

5.2 General results

As with Lemma 5.1.1, the size of an intersection set for any maximal clique of a Johnson graph $G = J_n(m, m-1)$ cannot be larger than $m-1$, given that is the size of the intersection of node label sets for a single edge. The main result of this section, analogous to Lemma 5.1.2, proves that this *maximum* size, $m-1$, and the *minimum* size, 0, are the *only* values possible for the size of the intersection set of a maximal clique in $J_n(m, m-1)$.

In Lemma 5.1.2, the types of maximal cliques for the simplest case of $J_n(2,1)$ were found by beginning with a clique of size two and seeing how it might be expanded by adding nodes. There were only two possible ways to do this, each leading to a different type of maximal clique. Here, we follow the same reasoning, but for vertices and edges from a $J_n(m, m-1)$ graph. The figure and proof of Lemma 5.1.2 guide the intuition in this more general case.

We begin with the case corresponding to the left most diagram of Lemma 5.1.2. There, a third node, v_3, was added by selecting the elements for its label set, $\nu(v_3) = \{a, c\}$ from the *union* of the label sets of the first two vertices v_1 and v_2, that is $\nu(v_3) \subset B = \nu(v_1) \cup \nu(v_2)$. This choice had repercussions in Lemma 5.1.2 in that the intersection set was null for $J_n(2,1)$ and the clique could not be enlarged past size 3 as in the dotted clique of Figure 5.1 (a). For the dotted clique of Figure 5.1 (b), however, the set B for two vertices from

a $J_n(3,2)$ is larger so that the clique can be enlarged to include a fourth node. In both dotted clique examples of Figure 5.1 the intersection of the label sets is null.

The next proposition characterizes the intersection sets for maximal cliques, H, formed in this way from a Johnson graph $G = J_n(m.m-1)$.

Proposition 5.2.1. *Let $G = J_n(m, m-1)$ be a Johnson graph with $n \geq m+1$ and let H be a maximal clique in G. If B denotes the union*

$$B := \bigcup_{v \in H} \nu(v),$$

then $\bigcap_{v \in V(H)} \nu(v) = \emptyset$ if, and only if, $\nu(v_1) \cup \nu(v_2) = B$ for any v_1, v_2 distinct in $V(H)$.

Proof. Suppose that for any distinct $v_1, v_2 \in V(H)$, $\nu(v_1) \cup \nu(v_2) = B$. It follows that for any node $v \in V(H)$, $\nu(v)$ is an $m-$subset of the $(m+1)-$set B. Moreover, since every $m-$subset of B corresponds to a node adjacent to every node in H, it follows that

$$\nu(V(H)) = \{A : A \subset B, |A| = m\}.$$

For any $i \in B$, the node with label $\{j \in B : j \neq i\}$ eliminates i from the intersection $\bigcap_{v \in V(H)} \nu(v)$. Thus, the intersection $\bigcap_{v \in V(H)} \nu(v)$ must be empty.

Conversely, suppose that $\bigcap_{v \in V(H)} \nu(v) = \emptyset$. If $\nu(v_1) \cup \nu(v_2) \neq B$ for some $v_1, v_2 \in V(H)$, then there exists some $v_3 \in V(H)$ such that $x_3 \in \nu(v_3)$ and $x_3 \notin \nu(v_1) \cup \nu(v_2)$. It follows that v_3 satisfies

$$\nu(v_3) = (\nu(v_1) \cap \nu(v_2)) \cup \{x_3\}.$$

Thus, v_1, v_2 and v_3 satisfy the hypothesis of Proposition 5.2.2 and it follows that $\cap_{v \in V(H)} \nu(v)$ is a set of size $m-1$, a contradiction.

□

It would appear, then, that, if we build up a maximal clique in this way, we end with one whose intersection set $S = \emptyset$ is of size zero. This being the smallest possible intersection set, $\mathcal{M}_{min}$ could again denote the set of such maximal cliques in $J_n(m, m-1)$.

Returning to the intuition followed in the figure of Lemma 5.1.2, consider how vertices were added when taking the righthand choice. The label set of any third vertex would be formed from the intersection $\nu(v_1) \cap \nu(v_2)$, necessarily of size $m-1$ in $J_n(m, m-1)$, joined by any element of $[n]$ not already appearing in the union $\nu(v_1) \cup \nu(v_2)$. Following the same logic, vertices could be added providing the intersection set remained of size $m-1$ until all remaining elements of $[n]$ were exhausted, that is, until $\nu(v_1) \cup \nu(v_2) \cup \cdots \cup \nu(v_r) = [n]$. For a Johnson $J_n(m, m-1)$, the number of vertices for such a clique would be $r = n - (m-1)$ (e.g., the thick solid line maximal cliques shown in Figures 5.1). Cliques formed in this fashion, would have largest possible intersection set of size $m-1$ and so could, again, be denoted $\mathcal{M}_{max}$.

The next proposition shows that building a maximal clique H of a Johnson graph $G = J_n(m.m-1)$ in this way can only lead to one having largest intersection set of size $m-1$ and, hence, to $H \in \mathcal{M}_{max}$.

Proposition 5.2.2. *Let H be a maximal clique in $G = J_n(m, m-1)$. If $|\nu(v_1) \cap \nu(v_2) \cap \nu(v_3)| = m-1$ for some distinct nodes $v_1, v_2, v_3 \in V(H)$, then*

$$\bigcap_{v \in V(H)} \nu(v) = \nu(v_1) \cap \nu(v_2).$$

Proof. Suppose that $I := \nu(v_1) \cap \nu(v_2) \cap \nu(v_3)$ for some distinct nodes v_1, v_2 and v_3 in H, and write $\nu(v_j)$ in the form $\{x_j\} \cup I$ for $j = 1, 2, 3$. If $\bigcap_{v \in V(H)} \nu(v) \neq I$, then there must be some $i \in I$ for which $i \notin \nu(u_i)$ for some $u_i \in V(H)$.

Then since $u_i \in V(H)$ and H is a clique, $u_i \sim v_1, v_2, v_3$. Therefore, $\nu(u_i)$ must contain x_j and an $(m-2)-$subset I_j from I, for $j = 1, 2, 3$. In other words,

$$\nu(u_i) = I_1 \cup I_2 \cup I_3 \cup \{x_1, x_2, x_3\}.$$

Since I is disjoint from $\{x_1, x_2, x_3\}$, so are the I_1, I_2, I_3. Since the nodes v_1, v_2, v_3 are distinct, the variables x_1, x_2, x_3 are distinct. Consequently,

$$|\nu(u_i)| = |I_1 \cup I_2 \cup I_3| + 3 \geq m - 2 + 3 = m + 1,$$

and such a node cannot exist in $J_n(m, m-1)$. □

Should a maximal clique $H \in J_n(m, m-1)$ belong to one of the two classes $\mathcal{M}_{min}$ and $\mathcal{M}_{max}$, Propositions 5.2.1 and 5.2.2 provide information about H which is summarized in the following remark.

Remark 5.2.3. *Maximal cliques $H \in J_n(m, m-1)$ have the following characteristics unique to which class, $\mathcal{M}_{min}$ or $\mathcal{M}_{max}$, they belong.*

- *$H \in \mathcal{M}_{min}$:*
  - *The intersection set $S = \cap_{v \in V(H)} \nu(v) = \emptyset$ with $|S| = 0$.*
  - *Node labels of all vertices $v_j \in V(H)$ are distinct and of the form $\nu(v_j) = B \backslash \{x_j\}$ for some $B \subset [n]$ of size $|B| = m+1$ and all $x_j \in B$.*
  - *The size of the maximal clique H is $|H| = |B| = m+1$.*
  - *For any distinct v_i, v_j in $V(H)$, $\nu(v_i) \cup \nu(v_j) = B$.*
  - *The number of distinct maximal cliques in $\mathcal{M}_{min}$ is $\binom{n}{|B|} = \binom{n}{m+1}$.*
- *$H \in \mathcal{M}_{max}$:*
  - *The intersection set $S = \cap_{v \in V(H)} \nu(v)$ has size $|S| = m-1$.*
  - *Node labels of all vertices $v_j \in V(H)$ are distinct and of the form $\nu(v_j) = A \cup \{x_j\}$ for some $A \subset [n]$ of size $|A| = m-1$ and all $x_j \in [n] \setminus A$.*
  - *The size of the maximal clique H is $|H| = |[n]| - |A| = n - m + 1$.*
  - *For any distinct v_i, v_j in $V(H)$, $\nu(v_i) \cap \nu(v_j) = A = S$.*

– *The number of distinct maximal cliques in* $\mathcal{M}_{max}$ *is* $\binom{n}{|A|} = \binom{n}{m-1}$.

Following the logic of Lemma 5.1.2, Propositions 5.2.1 and 5.2.2 demonstrate at least two distinct classes of maximal cliques exist in $J_n(m, m-1)$, $\mathcal{M}_{min}$ and $\mathcal{M}_{max}$. That these are the *only* types of maximal cliques in a Johnson graph $G = J_n(m, m-1)$ is proved in Theorem 5.2.4 (by induction on m, with Lemma 5.1.2 providing the initial case).

Theorem 5.2.4. *Let H be a maximal clique in the Johnson graph $G = J_n(m, m-1)$ for $n > m \geq 2$ and let $S = \cap_{v \in H} \nu(v)$ be the intersection set of the node labels of H. Then, $|S| \in \{0, (m-1)\}$.*

Proof. The proof proceeds by induction on m.

Suppose, that there exists $m_0 \geq 2$ such that, for all $n_0 > m_0$, any maximal clique H_0 in $J_{n_0}(m_0, m_0 - 1)$ with intersection set $S_0 = \cap_{v \in H_0} \nu(v)$ has $|S_0| \in \{0, m_0 - 1\}$. For a proof by induction, we need to show that this implies that for $m_1 = m_0 + 1$, any maximal clique H_1 in $J_{n_1}(m_1, m_1 - 1)$, for all $n_1 > m_1$, must have intersection set of size $|S_1| \in \{0, m_1 - 1\}$.

The inductive step is proved by contradiction. The size of the intersection set of a maximal clique H_1 in $J_{n_1}(m_1, m_1 - 1)$ is at most $m_1 - 1$, so we assume the intersection set size $|S_1| = s$ with $0 < s < m_1 - 1 = m_0$. Without loss of generality, we may take the intersection S_1 to be

$$S_1 = \{(n_1 - s + 1), (n_1 - s + 2), \ldots, n_1\}.$$

Knowing that the node label sets of H_1 have intersection S_1 of size s and that H_1 induces a maximal clique in $J_{n_1}(m_1, m_1 - 1)$, the label set for any vertex $v_i \in H_1$ can be expressed as $\nu(v_i) = I_i \cup S_1$ where $I_i \subset [n_1]$ with $I_i \cap S_1 = \emptyset$ and $|I_i \cap I_j| = m_1 - 1 - s$ for $i \neq j$ and $v_i, v_j \in H_1$.

This knowledge allows us to construct vertices $v_j^- \in J_{n_1-1}(m_1-1, m_1-2) = J_{n_0}(m_0, m_0-1)$ with $n_0 = n_1 - 1 > m_1 - 1 = m_0$ by simply removing $n_1 \in S_1$ from the node label sets of all vertices in H_1. That is, for every $v_i \in V(H_1)$, define $v_i^- \in V(J_{n_0}(m_0, m_0 - 1))$ to have node label set $\nu(v_i^-) = \nu(v_i) \setminus \{n_1\}$.

Now consider the graph, H_0, induced in $J_{n_0}(m_0, m_0 - 1)$ by the vertices v_i^-, so defined. It is easy to see that H_0 is a clique in $J_{n_0}(m_0, m_0 - 1)$ – the label sets of its vertices are $\nu(v_i^-) = I_i \cup S_1 \setminus \{n_1\}$ with $\nu(v_i^-) \cap \nu(v_j^-) = (I_i \cap I_j) \cup S_1 \setminus \{n_1\}$ yielding cardinalities of $(m_1 - s) + (s - 1) = m_1 - 1 = m_0$ and $(m_1 - s - 1) + (s - 1) = m_1 - 2 = m_0 - 1$, respectively.

To see that H_0 is also maximal, suppose that it is not. Then, there is some vertex $v \in J_{n_0}(m_0, m_0 - 1)$ which is not in H_0 but is adjacent to every vertex $v_i^- \in H_0$. This adjacency implies

$$|\nu(v) \cap (I_i \cup S_1 \setminus \{n_1\})| = m_0 - 1$$

for all $v_i^- \in H_0$.

We construct a vertex $v^+ \in J_{n_1}(m_1, m_1 - 1)$ having label set $\nu(v^+) = \nu(v) \cup \{n_1\}$. Since v is adjacent to every $v_i^- \in H_0$, its intersection with each is of size $m_0 - 1 =$

$|\nu(v) \cap (I_i \cup S_1 \setminus \{n_1\})|$. It follows that for $v_i \in H_1$,

$$\begin{aligned}
\nu(v^+) \bigcap \nu(v_i) &= \nu(v^+) \bigcap (I_i \cup S_1) \\
&= \nu(v^+) \bigcap [(I_i \cup (S_1 \setminus \{n_1\})) \bigcup \{n_1\}] \\
&= [\nu(v^+) \bigcap (I_i \cup (S_1 \setminus \{n_1\}))] \bigcup [\nu(v^+) \cap \{n_1\}] \\
&= [\nu(v^+) \bigcap (I_i \cup (S_1 \setminus \{n_1\}))] \bigcup [\{n_1\}] \\
&= [(\nu(v) \cup \{n_1\}) \bigcap (I_i \cup (S_1 \setminus \{n_1\}))] \bigcup \{n_1\} \\
&= [(\nu(v) \bigcap (I_i \cup (S_1 \setminus \{n_1\}))) \bigcup (\{n_1\} \bigcap (I_i \cup (S_1 \setminus \{n_1\})))] \bigcup \{n_1\} \\
&= [(\nu(v) \bigcap (I_i \cup (S_1 \setminus \{n_1\}))) \bigcup \emptyset] \bigcup \{n_1\} \\
&= [\nu(v) \bigcap (I_i \cup (S_1 \setminus \{n_1\}))] \bigcup \{n_1\}.
\end{aligned}$$

Now, in square brackets, the left set of the union does not contain n_1 and is of known cardinality $m_0 - 1$. It follows, then, that

$$\left|\nu(v^+) \bigcap \nu(v_i)\right| = \left|\nu(v) \bigcap (I_i \cup (S_1 \setminus \{n_1\}))\right| + |\{n_1\}| = (m_0 - 1) + 1 = m_1 - 1.$$

Hence, v^+ is adjacent to every vertex $v_i \in H_1$, and H_1 can be extended to a larger clique in $J_{n_1}(m_1, m_1 - 1)$ – a contradiction since H_1 was assumed to be maximal. It follows, then, that no such node, $v \in J_{n_0}(m_0, m_0 - 1)$, exists which extends the clique H_0 and, hence, that H_0 must be maximal.

By construction, the intersection set of H_0 is $S_0 = S_1 \setminus \{n_1\}$ and is of size $|S_0| = s - 1$. And, because H_0 is a maximal clique in $J_{n_0}(m_0, m_0 - 1)$, by the inductive hypothesis $|S_0|$ is either 0 or $m_0 - 1$. If the latter, then $s = m_0$ is outside the bounds assumed and we have a contradiction. If the former, then, $s = 1$ and, by Proposition 5.2.1, $H_0 \in \mathcal{M}_{min}$ so $S_0 = \emptyset$ and $|S_0| = 0$ – again, a contradiction.

It follows that intersecting sets of a maximal clique $H_1 \in J_n(m_1, m_1 - 1)$ with $m_1 = m_0 + 1$ must have size 0 or $m_1 - 1$, if it is the case that intersecting sets of a maximal clique $H_0 \in J_n(m_0, m_0 - 1)$ must be of size 0 or $m_0 - 1$. The proof by induction is complete by noting that, by Lemma 5.1.2, the inductive hypothesis holds for $m_0 = 2$.

□

Theorem 5.2.4 proved that every maximal clique $H \in J_n(m, m - 1)$ has either the minimal or maximal intersection set possible. That is, either $H \in \mathcal{M}_{min}$, or $H \in \mathcal{M}_{max}$. If the former, then $|H| = m + 1$; if the latter, then $|H| = n - m + 1$ (see Remark 5.2.3). As a consequence, we obtain the clique number of $J_n(m, m - 1)$ for all $n \geq m + 1$.

Corollary 5.2.5. *The clique number $\omega(J_n(m, m - 1))$ of the Johnson graph $J_n(m, m - 1)$ is given by*

$$\max(m + 1, n - m + 1),$$

whenever $n \geq m+1$.

Rewriting, it follows from Corollary 5.2.5, that the clique number is

$$\omega(J_n(m, m-1)) = \begin{cases} m+1 & \text{if} \quad m+1 \leq n \leq 2m \\ \\ n-m+1 & \text{if} \quad 2m \leq n \end{cases}$$

and is undefined otherwise.

5.3 Extending an r-clique

Given some clique $C_r \subset J_n(m, m-1)$ of size $|C_r| = r$, what can be said about the maximal cliques $H \subset J_n(m, m-1)$ that contain it?

We begin with edges ($r = 2$). As the figure in the proof of Lemma 5.1.2 suggests, every edge in $J_n(m, m-1)$ can appear in one clique from $\mathcal{M}_{min}$ and one from $\mathcal{M}_{max}$. Proposition 5.3.1 shows that each edge can appear in *only one* maximal clique in each of $\mathcal{M}_{min}$ and $\mathcal{M}_{max}$.

Proposition 5.3.1. *Each edge of $J_n(m, m-1)$ will belong to precisely one maximal clique $H_{min} \in \mathcal{M}_{min}$ and to precisely one maximal clique $H_{max} \in \mathcal{M}_{max}$.*

Proof. Select any edge e_{ij} connecting vertices v_i and v_j and let $A = \nu(v_i) \cap \nu(v_j)$ denote the intersection of their node label sets and $B = \nu(v_i) \cup \nu(v_j)$ their union.

Define H_{max} to be the subgraph of $J_n(m, m-1)$ induced by the vertex set

$$V(H_{max}) = \{v \in V(J_n(m, m-1)) : A \subset \nu(v)\}$$

and H_{min} that induced by the vertex set

$$V(H_{min}) = \{v \in V(J_n(m, m-1)) : \nu(v) \subset B\}.$$

Vertices v_i and v_j belong to both sets, so $e_{ij} \in H_{max}$ and $e_{ij} \in H_{min}$.

By construction, $H_{max} \in \mathcal{M}_{max}$ and, by Proposition 5.2.2, there can be no other maximal clique in $\mathcal{M}_{max}$ containing both v_i and v_j.

Similarly, $H_{min} \in \mathcal{M}_{min}$ and, by Proposition 5.2.1, any maximal clique in $\mathcal{M}_{min}$ containing both v_i and v_j must consist of precisely all of the $m-$subsets surrounding the union of the label sets $B = \nu(v_i) \cup \nu(v_j)$. Again there is no other such maximal clique in $\mathcal{M}_{min}$.

Theorem 5.2.4 completes the proof by guaranteeing that there are no other possible maximal cliques containing both vertices. □

Although every edge, or 2-clique, appears in one maximal clique from each of $\mathcal{M}_{min}$ and $\mathcal{M}_{max}$, this is not the case for any other clique of size $r > 2$. Proposition 5.3.2 shows that any r-clique, for $r > 2$, can only appear in one maximal clique, which can only be from one of $\mathcal{M}_{min}$ or $\mathcal{M}_{max}$.

Proposition 5.3.2. *Let $C_r \subset J_n(m, m-1)$ be a clique of size $r \geq 2$ with $C_r \subset H$ and H a maximal clique in $J_n(m, m-1)$. Then,*

- $H \in \mathcal{M}_{min}$ *only if* $\left|\bigcup_{v \in V(C_r)} \nu(v)\right| = m+1$ *and* $\left|\bigcap_{v \in V(C_r)} \nu(v)\right| = m+1-r$.
- $H \in \mathcal{M}_{max}$ *only if* $\left|\bigcup_{v \in V(C_r)} \nu(v)\right| = m-1+r$ *and* $\left|\bigcap_{v \in V(C_r)} \nu(v)\right| = m-1$.

Proof. Note that every r-clique, C_r, in a graph G is extendible to a maximal clique $H \subseteq G$. When $G = J_n(m, m-1)$, Theorem 5.2.4 shows that either, $H \in \mathcal{M}_{min}$, or, $H \in \mathcal{M}_{max}$ – there are no other possibilities.

Each has implications for the labels on the nodes in C_r. Without loss of generality, take $V(C_r) = \{v_1, \ldots, v_r\}$ as the vertices of C_r.

First, consider the case that $H \in \mathcal{M}_{min}$. In Remark 5.2.3, we note that every node $v_j \in V(H)$ has node label of the form $\nu(v_j) = B \setminus \{x_j\}$ for some set $B \subset [n]$, $x_j \in B$, and $|B| = m+1$. Further, $B = \nu(v_i) \cup \nu(v_j)$ for *every* pair of distinct nodes $v_i, v_j \in V(H)$. In particular, if C_r extends to $H \in \mathcal{M}_{min}$, then

$$\left|\bigcup_{v \in V(C_r)} \nu(v)\right| = |B| = m+1$$

and, for some $x_1, \ldots, x_r \in B$ (x_j peculiar to the node label set of each $v_j \in V(C_r)$),

$$\left|\bigcap_{v \in V(C_r)} \nu(v)\right| = \left|\bigcap_{j=1}^{r} (B \setminus \{x_j\})\right| = \left|B \setminus (\bigcup_{j=1}^{r} \{x_j\})\right| = m+1-r,$$

characterize the union and intersection sizes of the label sets for nodes in $V(C_r)$ when it is extendible to a maximal clique $H \in \mathcal{M}_{min}$.

Similarly, from Remark 5.2.3, if C_r extends to $H \in \mathcal{M}_{max}$, then

$$\left|\bigcup_{v \in V(C_r)} \nu(v)\right| = \left|\bigcup_{j=1}^{r} (A \cup \{x_j\})\right| = \left|A \cup \left(\bigcup_{j=1}^{r} \{x_j\}\right)\right| = m-1+r$$

for some set $A \subset [n]$, $x_j \in [n] \setminus A$ (x_j peculiar to the node label set of each $v_j \in V(C_r)$), and

$$\left|\bigcap_{v \in V(C_r)} \nu(v)\right| = |A| = m-1.$$

□

When $r = 2$, C_r is an edge and, by Proposition 5.3.1, there is both a maximal clique in $\mathcal{M}_{min}$ and one in $\mathcal{M}_{max}$ which extend C_r. This is corroborated by the matching set union sizes $(m+1)$ and set intersection sizes $(m-1)$ in Proposition 5.3.2 when $r = 2$. However, when $r > 2$ these sizes cannot match, and proving that for $r > 2$ any r-clique C_r extends to a unique maximal clique in $J_n(m, m-1)$ which must be a member of one of $\mathcal{M}_{min}$ or $\mathcal{M}_{max}$.

Corollary 5.3.3. *Let $C_r \subset J_n(m, m-1)$ be a clique of size $r > 2$, then C_r can be extended to only one maximal clique $H \subset J_n(m, m-1)$.*

Proof. By Proposition 5.3.2, C_r satisfies either $\left|\bigcup_{v \in V(C_r)} \nu(v)\right| = m+1$ or $\left|\bigcap_{v \in V(C_r)} \nu(v)\right| = m-1$, but not both. Thus, there is at least one maximal clique H which extends C_r.

If H is in $\mathcal{M}_{max}$, then by Proposition 5.2.2, the intersection of H must be the intersection of C_r. Similarly, if H is in $\mathcal{M}_{min}$, then by Proposition 5.2.1, the union of H must be the union of C_r.

In either case, the intersection and union of labels determine the maximal clique H uniquely. □

Corollary 5.3.3 shows that any clique C_r of size $r > 2$ can be extended to a maximal clique belonging to only one of $\mathcal{M}_{min}$ or $\mathcal{M}_{max}$. Proposition 5.3.2 provides the means for telling which one by examining the size of the intersection or of the union of the node labels of C_r.

5.3.1 The clique partition number

Should interest lie in the minimum number of cliques needed to partition the edges of $J_n(m, m-1)$, that is, its clique partition number $cp(J_n(m, m-1)$ (Erdős et al., 1988), then the maximal cliques produced by the edges in $J_n(m, m-1)$ provide the solution.

Proposition 5.3.1 showed that each edge in $J_n(m, m-1)$ led to a unique maximum clique in each of $\mathcal{M}_{min}$ and $\mathcal{M}_{max}$. That each edge will appear in only one element of each set and that the elements are maximal cliques, means that the cliques in either set partition the edges and that they are the fewest possible of that type. It remains only to determine which set, $\mathcal{M}_{min}$ or $\mathcal{M}_{max}$, is smaller – its size will be the clique partition number. Proposition 5.3.1 thus yields the following corollary.

Corollary 5.3.4. *The clique partition number of $J_n(m, m-1)$ is given by*

$$cp(J_n(m, m-1)) = \min\{|\mathcal{M}_{min}|, |\mathcal{M}_{max}|\}.$$

Or, to be precise, referring to the sizes of these sets in Remark 5.2.3, with minimal rewriting, exact expressions may be had as follows.

Corollary 5.3.5. *The clique partition number of* $J_n(m, m-1)$ *is given by*

$$cp(J_n(m,m-1)) = \begin{cases} \binom{n}{m-1} & n < 2m \\ \\ \binom{n}{m+1} & n \geq 2m \end{cases}$$

5.4 Discussion

While the results obtained above apply directly to the cliques and maximal cliques of a Johnson graph, $J_n(m, m-1)$, they may also be expressed in terms of families of intersecting subsets of $[n]$ – an *intersecting family*, $\mathcal{F}$, is a subset of the power set, $\mathcal{P}(n)$, whose elements are pairwise non-disjoint, that is, $A \cap B \neq \emptyset$ for every $A, B \in \mathcal{F}$ (e.g., see Erdős & Kleitman, 1974).

Numerous results have been found for intersection families (e.g., see Gerbner & Patkós, 2018), including the celebrated Erdős-Ko-Rado (EKR) theorem Erdős et al. (1961) which showed that any intersecting family having elements of size $k \leq m \leq \frac{1}{2}n$, and having no element contained in another (i.e., is an *antichain*, or *Sperner family*), had size at most $\binom{n-1}{m-1}$. The EKR bound is attained by the trivially intersecting family $\mathcal{F} \subset \binom{[n]}{m}$ defined by $\mathcal{F} = \{A \in \binom{[n]}{m} \ : \ x \in A \in [n]\}$ for some choice of x. Similarly, Hilton & Milner (1967) showed that when $\mathcal{F} \subset \binom{[n]}{m}$ is further restricted to have non-null intersection, $\cap_{A \in \mathcal{F}} A \neq \emptyset$, across all sets in $\mathcal{F}$, for $2 \leq m \leq \frac{1}{2}n$, there exists another class of maximal intersecting families within $\binom{[n]}{m}$ whose size is bounded above by $\binom{n-1}{m-1} - \binom{n-m-1}{m-1} + 1$.

The results of the present chapter are restricted to intersecting families $\mathcal{F}_{n,m,m-1} \subset \binom{[n]}{m}$ having $|A \cap B| = m - 1$ for distinct $A, B \in \mathcal{F}_{n,m,m-1}$. The set of node label sets of the vertices from any maximal clique in $J_n(m, m-1)$ corresponds to a maximally intersecting family $\mathcal{F} \subset \mathcal{F}_{n,m,m-1}$.

Theorem 5.2.4 shows that there are only two possible types of such maximal intersecting families, say $\mathcal{F}_{min}$, $\mathcal{F}_{max} \subset \mathcal{F}_{n,m,m-1}$, corresponding to the two types of maximal cliques, $\mathcal{M}_{min}$, $\mathcal{M}_{max} \subset J_n(m, m-1)$. Expressing Remark 5.2.3 in terms of intersecting families, there are exactly $\binom{n}{m+1}$ distinct families $\mathcal{F} \in \mathcal{F}_{min}$ and the following hold for each family $\mathcal{F}$:

- $|\cap_{A \in \mathcal{F}} A| = 0$,
- $\exists B \subset [n]$ of size $m+1$ with every $A \in \mathcal{F}$ having the form $B \setminus \{x\}$ for all $x \in B$,
- $A_i \cup A_j = B$ for all $A_i, A_j \in \mathcal{F}$, $i \neq j$, and
- there are $m+1$ elements in $\mathcal{F}$,

and there are $\binom{n}{m-1}$ distinct families $\mathcal{F} \in \mathcal{F}_{max}$ for each of which the following hold:

- $|\cap_{A\in\mathcal{F}} A| = m-1$,
- $\exists B \subset [n]$ of size $m-1$ with every $A \in \mathcal{F}$ having the form $B \cup \{x\}$ for all $x \in [n] \setminus B$,
- $A_i \cap A_j = \cap_{A\in\mathcal{F}} A$ for all $A_i, A_j \in \mathcal{F}$, $i \neq j$, and
- there are $n-m+1$ elements in $\mathcal{F}$.

Corollary 5.2.5 gives the size of the maximal intersecting $\mathcal{F} \subset \binom{[n]}{m}$ restricted to every pair intersection being of size $m-1$.

Similarly, Proposition 5.3.2 gives conditions on the size of the union and intersection of sets in a family $\mathcal{F}_r \subset \mathcal{F}_{n,m,m-1}$ of size $r \geq 2$ for $\mathcal{F}_r$ to be extended to a maximally intersecting family of type $\mathcal{F}_{min}$ or $\mathcal{F}_{max}$ – only one of which is possible for $r > 2$.

We again note that Chapter 4 showed how intersecting families of sets are related to the partition of a set of cliques defining a clique cover for any graph.

Consider again the problem of visualizing high dimensional statistical data which served as our initial motivation. The Johnson graph $J_n(2,1)$ has been used successfully in visual data analysis as shown by Oldford & Waddell (2011) and Hofert & Oldford (2017).

For a $J_n(2,1)$ graph, the maximal cliques are either a triangle representing a $3d$-space defined by the three variables in the union of the node labels ($\mathcal{M}_{min}$)(e.g., see Hurley & Oldford, 2011b, Figs. 2a and 3), or, an $(n-1)$-clique representing an n-dimensional space which privileges one of the n variables to appear in every $2d$ node (subspace) and swaps one of the remaining variables for another whenever an edge is followed ($\mathcal{M}_{max}$). The latter is natural in statistics, for example, when the privileged variable might be regressed upon the second variable at each node (or vice versa).

More generally, for $J_n(m, m-1)$, traversing maximal cliques in $\mathcal{M}_{min}$ is an exploration of an $m+1$ dimensional space via swapping one of the variables for another with every movement along an edge. Cliques in $\mathcal{M}_{max}$ now privilege $m-1$ variables (e.g. as regressors) while exploring the effect of changing one variable with another (e.g. as response variables in a regression model) for the remaining $n-m+1$ variables with every movement along an edge.

6

Variable graphs, random graphs and navigation graphs

Under some assumptions, the variable graphs from Chapter 2 behave like Erdős-Rényi random graphs. By carefully examining how cliques arise from the line graph operator, we can specialize the theory developed in Chapters 3, 4 and 5 to the clique centric study of navigation graphs. In addition to obtaining closed-form expressions for the moments of clique counts, we analyze the asymptotic nature of cliques in Johnson graphs and the potential implications for navigation graphs.

6.1 On variable graphs and random graphs

We begin by outlining the assumptions we make regarding the collection of random variables and the measure of interest w. Recall that $\mathcal{V}$ is the collection of random variables representing the underlying distributions of the data observed. Let $w : \binom{\mathcal{V}}{2} \to \mathbb{R}$ be the function quantifying the peculiarity of a space according to a predetermined measure of interest.

Let G be the simple, weighted complete variable graph on $\mathcal{V}$, where $E(G) = \binom{\mathcal{V}}{2}$ consists of all 2-spaces of interest, where an edge between X and Y is weighted by $w(\{X, Y\})$. Upon choosing one of the two mechanisms for pruning the navigation graph, the analyst obtains a subgraph of the complete variable graph.

We make the following simplifying assumptions regarding the joint distribution of the random variables $w(\{X, Y\})$:

A1 The random variables $w(\{X, Y\})$ are all independent, identically distributed random variables with some known distribution F.

A2 The random variables $w(\{X, Y\})$ and $W(\{X, Z\})$ are independent for all $X, Y, Z \in \mathcal{V}$.

While the distribution and independence assumptions are unrealistic in practice, they allow for a rough model of the behaviour of the variable graphs, and enable us to apply the insights and tools we have obtained in Chapters 3, 4 and 5. More broadly, since the Bernoulli sums framework of Chapter 3 applies to indicator variables with any form of dependence structure, we remark that expressions for the clique count moments may still be derived if provided with the joint distribution of $(w(\{X, Y\} : \{X, Y\} \in \binom{\mathcal{V}}{2}))$.

Now, we recall on M1 from Section 1.1.1 in the case where $m = 2, k = 1$:

M1 Fix a cutoff value t. Let G be the induced navigation subgraph produced by keeping the nodes $\{X, Y\}$ with $w(\{X, Y\}) > t$.

M2 Fix a proportion $q \in (0, 1)$. Let G be the induced navigation subgraph produced by keeping the nodes $\{X, Y\}$ with $w(\{X, Y\}) > t_q$, where t_q is the $q-$th quantile of the empirical distribution of w: i.e., the value x satisfying that the proportion of spaces $\{X, Y\}$ in $\mathcal{V}$ with $w(\{X, Y\}) < x$ is q.

Methods **M1** and **M2** can be restated completely in terms of the pruning of the underlying complete variable graph. In the case of **M1**, an edge e remains whenever $w(e) > t$. Similarly, under **M2**, an edge e remains if its weight is in the top q-th percentile of edge weights with respect to w.

Proposition 6.1.1. *Let V be the underlying variable graph V produced by method **M1** with fixed cutoff t and weight function w. If w satisfies **A1** and **A2**, then V is an Erdős-Rényi random graph with parameters $G(n, Pr(F > t))$.*

Proof. We demonstrate that the edge inclusion probabilities are independent and identically distributed Bernoulli random variables. Without loss of generality, suppose that V was generated by Method **M1**. Let $e = \{X, Y\}$ be an edge in G and consider the probability $Pr(w(e) > t)$ of e remaining after pruning:

$$Pr(w(\{X, Y\}) > t) = Pr(F > t) := p_t,$$

where the equality follows from **A1**. Similarly, by **A1** and **A2**,

$$Pr(w(\{X_1, Y_1\}) > t, w(\{X_2, Y_2\}) > t) = Pr(w(\{X_1, Y_1\}) > t)Pr(w(\{X_2, Y_2\}) > t)$$

holds for all distinct 2-subsets $\{X_1, Y_1\}, \{X_2, Y_2\}$ in $\binom{\mathcal{V}}{2}$.

We note that if the cutoff value t is selected such that $Pr(F > t) = p$, the resulting graph is an Erdős-Rényi random graph $G(n, p)$.

□

We note that under **M2**, the number of edges maintained in the variable graph is constant. In other words, if the proportion $q \in (0, 1)$ is fixed, the number of edges remaining in the complete variable graph after pruning is always $\lceil q\binom{n}{2} \rceil$. On the other hand, under the fixed cutoff method **M1**, the number of edges remaining is a random variable as independent, identically distributed realizations of the same dataset could produce different scores on the scagnostics.

The fact that all realizations of variable graphs under **M2** have the same number of edges and each realization is independent and identically distributed according to **A1** and **A2** immediately proves the following result.

Proposition 6.1.2. *Let V be the underlying variable graph V produced by method **M2** with fixed proportion q and weight function w. If w satisfies **A1** and **A2**, then V is an Erdős-Rényi random graph with parameters $G(n, M = \lceil q\binom{n}{2} \rceil)$.*

By viewing navigation graphs as realizations of line graphs of a subgraph of a complete variable graph, we can exploit the clique structure of random graphs under suitable assumptions.

6.1.1 Cliques and the line graph operator

Proposition 6.1.1 states that under special circumstances, we can model a pruned variable graph as an Erdős-Rényi random graph $G(n, p)$. The following Proposition describes precisely when a subgraph of a graph becomes a clique under the line graph operator.

Proposition 6.1.3. *Let G be a graph and let H be a subgraph of $L(G)$. Then $L(H)$ is a clique in $L(G)$ if and only if H is K_3 or H is a star.*

Proof. We first show that the star $K_{1,n}$ becomes a clique under the line graph operator.

If H is the star $K_{1,n}$, it is easy to show that every edge in H has the form $\{x, y\}$ where x is the vertex of degree n in $K_{1,n}$ and y is one of the n vertices of degree 1. Therefore, image of H under the line graph operator would result in nodes of the form $\{x, y\}$ and since the corresponding edges are incident in H, the nodes are pairwise adjacent in $L(H)$. Thus, $L(H)$ is a clique.

If H is the complete graph K_3 with the node set $\{x, y, z\}$, then the line graph operator applied to H produces the nodes $\{x, y\}, \{x, z\}, \{y, z\}$, all of which are pairwise adjacent in $L(G)$ because their corresponding edges incident in G. Therefore, $L(H)$ is a clique.

Now, suppose $L(H)$ is a clique in $L(G)$. Then $L(H) \simeq K_r$ for some $r \geq 1$. If $r \neq 3$, by Whitney's graph isomorphism theorem (Theorem 2.1.3), since $L(H)$ is connected and isomorphic to $L(K_{1,n})$, H and $K_{1,n}$ must be isomorphic.

On the other hand, if $r = 3$, then either H is either $K_{1,3}$ or K_3. □

It immediately follows that the maximal cliques of $L(G)$ are categorized into those induced by triangles and those induced by maximal stars.

Corollary 6.1.4. *Let G be a graph. Every maximal clique H of $L(G)$ correspond to either a triangle in G or a vertex of degree at least 3 in G.*

Therefore, to capture the moments of the $r-$clique counts in navigation graphs under the proposed model, we need to know the distribution of star counts, and in the special case when $r = 3$, we need to also know the distribution of triangles.

The distribution of $r-$stars in $G(n, p)$ is a challenging problem because of the dependence structure among node degrees. Even though edge inclusions are pairwise independent under our model assumptions, the presence of large degree nodes make it increasingly likely for other high degree nodes to be present. For instance, it is impossible for $G \sim G(n, p)$ to have a node of degree $n - 1$ and a node of degree 0.

In the following section, we derive the joint distribution of the degrees of vertices in $G(n, p)$. We use this in tandem with our previous results to approximate the distribution of counts of nodes in $G(n, p)$ of a particular, fixed degree.

6.1.2 Degree counts in $G(n,p)$

Let D_i denote the degree of vertex i in a realization of $G(n,p)$. Since all vertices have the same degree distribution, without loss of generality we consider the distribution of the first m vertices $\{1,2,\ldots,m\}$ when deriving the joint distribution of any m nodes. Before describing the joint degree distribution of any number of nodes, we examine a simpler case: the bivariate degree distribution of two vertices.

The following illustrates that the degree of one vertex is independent of the degree of another, given their adjacency.

Proposition 6.1.5. *Suppose that $n \geq 2$ and $G = G(n,p)$. Then*

$$P(D_1|E_{12}, D_2) = P(D_1|E_{12}).$$

Proof.

$$P(D_1 = d_1|D_2 = d_2, E_{12} = e) = \frac{P(D_1 = d_1, E_{12} = e, D_2 = d_2)}{P(D_2 = d_2, E_{12} = e,)}$$

$$= \frac{P(\sum_{j\neq 1,2} E_{j1} = d_1 - e, \quad \sum_{j\neq 1,2} E_{j2} = d_2 - e, E_{12} = e)}{P(\sum_{j\neq 1,2} E_{j2} = d_2 - e, E_{12} = e)}$$

$$= \frac{\binom{n-2}{d_1-e}p^{d_1-e}(1-p)^{n-2-(d_1-e)}\binom{n-2}{d_2-e}p^{d_2-e}(1-p)^{n-2-(d_2-e)}}{\binom{n-2}{d_2-e}p^{d_2-e}(1-p)^{n-2-(d_2-e)}}$$

$$= \binom{n-2}{d_1-e}p^{d_1-e}(1-p)^{n-2-(d_1-e)}$$

$$= \frac{P(\sum_{j\neq 1,2} E_{j1} = d_1 - e, E_{12} = e)}{P(E_{12} = e)}$$

$$= P(D_1 = d_1|E_{12} = e).$$

□

Using Proposition 6.1.5, we derive the bivariate degree distribution of an Erdős-Rényi graph.

Proposition 6.1.6. *The joint distribution (D_1, D_2) of $G = G(n,p)$ is given by*

$$\left(\left[\frac{n-1-d_1}{n-1}\right]\binom{n-2}{d_2}p^{d_2}(1-p)^{n-2-d_2} + \left[\frac{d_1}{n-1}\right]\binom{n-2}{d_2-1}p^{d_2-1}(1-p)^{n-2-d_2+1}\right)$$
$$\times \binom{n-1}{d_1}p^{d_1}(1-p)^{n-1-d_1}.$$

Proof.

$$\begin{aligned}
P(D_2 = d_2 | D_1 = d_1) &= \sum_{e \in \{0,1\}} P(D_2 = d_2, E_{12} = e | D_1 = d_1) \\
&= \sum_{e \in \{0,1\}} \frac{P(D_2 = d_2, E_{12} = e, D_1 = d_1)}{P(D_1 = d_1)} \\
&= \sum_{e \in \{0,1\}} \frac{P(D_2 = d_2 | E_{12} = e, D_1 = d_1) P(E_{12} = e, D_1 = d_1)}{P(D_1 = d_1)} \\
&= \sum_{e \in \{0,1\}} \frac{P(D_2 = d_2 | E_{12} = e) P(E_{12} = e, D_1 = d_1)}{P(D_1 = d_1)},
\end{aligned}$$

by Proposition 6.1.5. Now,

$$\begin{aligned}
\frac{P(E_{12} = e, D_1 = d_1)}{P(D_1 = d_1)} &= \frac{p^e (1-p)^{1-e} \binom{n-2}{d_1 - e} p^{d_1 - e} (1-p)^{n-2-(d_1-e)}}{\binom{n-1}{d_1} p^{d_1} (1-p)^{n-1-d_1}} \\
&= \frac{\binom{n-2}{d_1 - e} p^{d_1} (1-p)^{n-1-d_1}}{\binom{n-1}{d_1} p^{d_1} (1-p)^{n-1-d_1}} \\
&= I_{[e=0]} \left[\frac{n-1-d_1}{n-1} \right] + I_{[e=1]} \left[\frac{d_1}{n-1} \right].
\end{aligned}$$

Moreover,

$$\begin{aligned}
P(D_2 = d_2 | D_1 = d_1) = \sum_{e \in \{0,1\}} &\binom{n-2}{d_2 - e} p^{d_2 - e} (1-p)^{n-2-(d_2-e)} \left(I_{[e=0]} \left[\frac{n-1-d_1}{n-1} \right] \right. \\
&\left. + \left[I_{[e=1]} \frac{d_1}{n-1} \right] \right) \\
= &\left[\frac{n-1-d_1}{n-1} \right] \binom{n-2}{d_2} p^{d_2} (1-p)^{n-2-d_2} \\
&+ \left[\frac{d_1}{n-1} \right] \binom{n-2}{d_2 - 1} p^{d_2 - 1} (1-p)^{n-2-d_2+1}.
\end{aligned}$$

Therefore, the joint is given by

$$\begin{aligned}
\binom{n-1}{d_1} p^{d_1} (1-p)^{n-1-d_1} \left(\left[\frac{n-1-d_1}{n-1} \right] \binom{n-2}{d_2} p^{d_2} (1-p)^{n-2-d_2} \right. \; + \\
\left. \left[\frac{d_1}{n-1} \right] \binom{n-2}{d_2 - 1} p^{d_2 - 1} (1-p)^{n-2-d_2+1} \right)
\end{aligned}$$

□

To derive a similar expression in more generality for the joint distribution of node degrees, we first generalize Proposition 6.1.5.

Proposition 6.1.7. *Let $\mathcal{I}$ be a collection of vertices in $G(n,p)$ and let $\mathcal{D} := \{D_i : i \in \mathcal{I}\}$. If $j \notin \mathcal{I}$ and $\mathbf{E}_j$ is the random vector recording the edges between j and $\mathcal{I}$ then*

$$P(D_j|\mathbf{E}_j, \mathcal{D}) = P(D_j|\mathbf{E}_j).$$

Proof.

$$\begin{aligned} P(D_j = d_j|\mathbf{E}_j = \mathbf{e}_j, \mathcal{D} = \mathbf{d}) &= \frac{P(D_j = d_j, \mathbf{E}_j = \mathbf{e}_j, \mathcal{D} = \mathbf{d})}{P(\mathbf{E}_j = \mathbf{e}_j, \mathcal{D} = \mathbf{d})} \\ &= \frac{P\left(\sum_{\ell \neq j} E_{\ell j} = d_j - \sum_{i \in I} e_{ij}, \mathbf{E}_j = \mathbf{e}_j\right)}{P\left(\mathbf{E}_j = \mathbf{e}_j, \left[\sum_{k \neq i} E_{ki} = d_i - e_{ij} : i \in \mathcal{I}\right]\right)}, \end{aligned}$$

where the event in the numerator is within the set $\left[\sum_{k \neq i} E_{ki} = d_i - e_{ij} : i \in \mathcal{I}\right]$ and therefore the ratio is equal to

$$\begin{aligned} P\left(\sum_{\ell \neq j} E_{\ell j} = d_j - \sum_{i \in I} e_{ij}\right) &\times \frac{P\left(\mathbf{E}_j = \mathbf{e}_j, \left[\sum_{k \neq i} E_{ki} = d_i - e_{ij} : i \in \mathcal{I}\right]\right)}{P\left(\mathbf{E}_j = \mathbf{e}_j, \left[\sum_{k \neq i} E_{ki} = d_i - e_{ij} : i \in \mathcal{I}\right]\right)} \\ &= P(D_j|\mathbf{E}_j). \end{aligned}$$

□

Next, we obtain an expression for the joint distribution of the degrees any number of nodes in $G(n,p)$.

Theorem 6.1.8. *Let $G = G(n,p)$ and fix $m \leq n$. If*

$$\mathcal{E}_n = \{\mathbf{e} : \mathbf{e} = (e_{ij})_{\{i,j\} \subset [n]}, e_{ij} \in \{0,1\}\},$$

the joint distribution of $(D_1, D_2, \ldots D_m) = (d_1, \ldots, d_m)$ is

$$\sum_{\mathbf{e} \in \mathcal{E}_m} p^{|\mathbf{e}|}(1-p)^{\binom{m}{2} - |\mathbf{e}|} \prod_{i=1}^{m} \binom{n-m}{d_i - \sum_{j \in [n] \setminus \{i\}} e_{ij}} p^{d_i - \sum_{j \in [n] \setminus \{i\}} e_{ij}} (1-p)^{n-m-d_i + \sum_{j \in [n] \setminus i} e_{ij}},$$

where

$$|\mathbf{e}| = \sum_{\{i,j\} \subset [n]} e_{i,j}.$$

Proof. We proceed by induction on m, the number of node degrees in the random vector.

The base case is $m = 2$. Proposition 6.1.6 shows that $P(D_1 = d_1, D_2 = d_2)$ is

$$\begin{aligned} \left(\left[\frac{n-1-d_1}{n-1}\right]\binom{n-2}{d_2} p^{d_2}(1-p)^{n-2-d_2} + \left[\frac{d_1}{n-1}\right]\binom{n-2}{d_2-1} p^{d_2-1}(1-p)^{n-2-d_2+1}\right) \\ \times \binom{n-1}{d_1} p^{d_1}(1-p)^{n-1-d_1}. \end{aligned} \tag{6.1}$$

Our goal is to demonstrate that this distribution is equal to

$$(1-p)\left[\binom{n-2}{d_1}p^{d_1}(1-p)^{n-2-d_1}\binom{n-2}{d_2}p^{d_2}(1-p)^{n-2-d_2}\right] + p\left[\binom{n-2}{d_1-1}p^{d_1-1}(1-p)^{n-1-d_1}\binom{n-2}{d_2-1}p^{d_2-1}(1-p)^{n-1-d_2}\right]. \tag{6.2}$$

Dividing Equation 6.1 by $P(D_1 = d_1) = \binom{n-1}{d_1}p^{d_1}(1-p)^{n-1-d_1}$ yields

$$\begin{aligned}(*) &= \frac{\binom{n-2}{d_1}}{\binom{n-1}{d_1}}p^{d_1-d_1}(1-p)^{n-1-d_1-(n-1-d_1)}P(D_2 = d_2, E_{12} = 0) \\ &\quad + \frac{\binom{n-2}{d_1-1}}{\binom{n-1}{d_1}}p^{d_1-d_1}(1-p)^{n-1-d_1-(n-1-d_1)}P(D_2 = d_2, E_{12} = 1) \\ &= \frac{n-1-d_1}{n-1}\binom{n-2}{d_2}p^{d_2}(1-p)^{n-2-d_2} + \frac{d_1}{n-1}\binom{n-2}{d_2-1}p^{d_2-1}(1-p)^{n-1-d_2},\end{aligned}$$

which agrees with Proposition 6.1.6 after dividing by $P(D_1 = d_1)$. This shows the statement is true in the base case $m = 2$.

Now, suppose that the claim holds for some $m \geq 2$. Recall that Proposition 6.1.7 states that

$$P(D_j|\mathbf{E}_j, \mathcal{D}) = P(D_j|\mathbf{E}_j),$$

where $j \in [n]$ is a node and $\mathcal{D}$ is the random vector recording the degrees of all vertices i in some set of vertices $\mathcal{I}$. Thus,

$$\begin{aligned}P([D_i = d_i, 1 \leq i \leq m+1]) &= \sum_{\mathbf{e}_{m+1}} P([D_i = d_i, 1 \leq i \leq m+1], \mathbf{E}_{m+1} = \mathbf{e}_{m+1}) \\ &= \sum_{\mathbf{e}_{m+1}} P(D_{m+1} = d_{m+1}|\mathbf{E}_{m+1} = \mathbf{e}_{m+1}, [D_i = d_i, 1 \leq i \leq m]) \\ &\qquad \times P(\mathbf{E}_{m+1} = \mathbf{e}_{m+1}, [D_i = d_i, 1 \leq i \leq m])) \\ &= \sum_{\mathbf{e}_{m+1}} P(D_{m+1} = d_{m+1}|\mathbf{E}_{m+1} = \mathbf{e}_{m+1}) \\ &\qquad \times P(\mathbf{E}_{m+1} = \mathbf{e}_{m+1}, [D_i = d_i, 1 \leq i \leq m])).\end{aligned}$$

Simplifying terms we have

$$P(D_{m+1} = d_{m+1}|\mathbf{E}_{m+1} = \mathbf{e}_{m+1}) = \binom{n-(m+1)}{d_{m+1} - |\mathbf{e}_{m+1}|}p^{d_{m+1}-|\mathbf{e}_{m+1}|}(1-p)^{n-(m+1)-(d_{m+1}-|\mathbf{e}_{m+1}|)}.$$

Now, by the inductive hypothesis

$$\begin{aligned}P(\mathbf{E}_{m+1}=\mathbf{e}_{m+1},[D_i=d_i,1\le i\le m]))&=P(\mathbf{E}_{m+1}=\mathbf{e}_{m+1},[D_i=d_i-e_{i,m+1},1\le i\le m]\\&\qquad\text{in } G\setminus\{m+1\})\\&=p^{|\mathbf{e}_{m+1}|}(1-p)^{m+1-|\mathbf{e}_{m+1}|}\\&\quad\times\sum_{\mathbf{e}\in\mathcal{E}_m}\prod_{i=1}^{m}p^{|\mathbf{e}|}(1-p)^{\binom{m}{2}-|\mathbf{e}|}\\&\quad\times\binom{n-1-m}{d_i-e_{i,m+1}-\sum_{j\neq\{i,m+1\}}e_{ij}}\\&\quad\times(1-p)^{n-m-1-d_i+e_{i,m+1}+\sum_{j\neq\{i,m+1\}}e_{ij}}\\&\quad\times p^{d_i-e_{i,m+1}-\sum_{j\neq\{i,m+1\}}e_{ij}}.\end{aligned}$$

Thus,

$$\begin{aligned}P([D_i=d_i,1\le i\le m+1])&=\sum_{\mathbf{e}_{m+1}}\binom{n-(m+1)}{d_{m+1}-|\mathbf{e}_{m+1}|}p^{d_{m+1}-|\mathbf{e}_{m+1}|}(1-p)^{n-(m+1)-(d_{m+1}-|\mathbf{e}_{m+1}|)}\\&\quad\times p^{|\mathbf{e}_{m+1}|}(1-p)^{m+1-|\mathbf{e}_{m+1}|}\sum_{\mathbf{e}\in\mathcal{E}_m}p^{|\mathbf{e}|}(1-p)^{\binom{m}{2}-|\mathbf{e}|}\\&\quad\times\prod_{i=1}^{m}\binom{n-1-m}{d_i-e_{i,m+1}-\sum_{j\neq\{i,m+1\}}e_{ij}}\\&\quad\times p^{d_i-e_{i,m+1}-\sum_{j\neq\{i,m+1\}}e_{ij}}\\&\quad\times(1-p)^{n-m-1-d_i+e_{i,m+1}+\sum_{j\neq\{i,m+1\}}e_{ij}},\end{aligned}$$

which can be further simplified into

$$\begin{aligned}&=\sum_{\mathbf{e}_{m+1},\mathbf{e}}p^{|\mathbf{e}|+|\mathbf{e}_{m+1}|}(1-p)^{m+1+\binom{m}{2}-|\mathbf{e}_{m+1}|-|\mathbf{e}|}\binom{n-m-1}{d_{m+1}-\sum_{i\neq m+1}e_{i,m+1}}p^{d_{m+1}-\sum_{i\neq m+1}e_{i,m+1}}\\&\qquad\times(1-p)^{n-(m+1)-(d_{m+1}-\sum_{i\neq m+1}e_{i,m+1})}\prod_{i=1}^{m}\binom{n-1-m}{d_i-e_{i,m+1}-\sum_{j\neq\{i,m+1\}}e_{ij}}\\&\qquad\times p^{d_i-e_{i,m+1}-\sum_{j\neq\{i,m+1\}}e_{ij}}(1-p)^{n-m-1-d_i+e_{i,m+1}+\sum_{j\neq\{i,m+1\}}e_{ij}}\end{aligned}$$

$$\begin{aligned}&=\sum_{\mathbf{f}=(\mathbf{e},\mathbf{e}_{m+1})\in\mathcal{E}_{m+1}}p^{|\mathbf{f}|}(1-p)^{\binom{m+1}{2}-|\mathbf{f}|}\prod_{i=1}^{m+1}\binom{n-1-m}{d_i-\sum_{j\in[n]\setminus\{i\}f_{ij}}}\\&\qquad\times p^{d_i-\sum_{j\in[n]\setminus\{i\}}f_{ij}}(1-p)^{n-(m+1)-d_i+\sum_{j\in[n]\setminus i}f_{ij}},\end{aligned}$$

as claimed. Therefore, by induction, the joint distribution of $(D_1,\ldots,D_m)$ is given by

$$\sum_{\mathbf{e}\in\mathcal{E}_m}p^{|\mathbf{e}|}(1-p)^{\binom{m}{2}-|\mathbf{e}|}\prod_{i=1}^{m}\binom{n-m}{d_i-\sum_{j\in[n]\setminus\{i\}}e_{ij}}p^{d_i-\sum_{j\in[n]\setminus\{i\}}e_{ij}}(1-p)^{n-m-d_i+\sum_{j\in[n]\setminus i}e_{ij}},$$

□

Recall that our goal lies in approximating the distribution of vertices of degree exactly ℓ in $G(n,p)$. Let Z_ℓ denote the total count and $Y_i^{(\ell)}$ be the indicator random variable recording if vertex i has degree ℓ. Therefore, $Y_i^{(\ell)}$ is Bernoulli($\binom{n-1}{k}p^\ell(1-p)^{n-1-\ell}$) and Z_ℓ is a Bernoulli sum.

Proposition 6.1.9. *The $k-$th moment of Z_ℓ is*

$$E(Z_\ell^k) = \sum_{m=1}^{k} \binom{k}{m} S(k,m) \sum_{\mathbf{e}\in\mathcal{E}_m} p^{|\mathbf{e}|}(1-p)^{\binom{m}{2}-|\mathbf{e}|} \times \prod_{i=1}^{m} \binom{n-m}{\ell - \sum_{j\in[n]\setminus\{i\}} e_{ij}} p^{\ell-\sum_{j\in[n]\setminus\{i\}} e_{ij}} (1-p)^{n-m-\ell+\sum_{j\in[n]\setminus i} e_{ij}}.$$

Proof. By Proposition 3.2.1,

$$E(Z_\ell^k) = \sum_{m=1}^{k} \sum_{i_1\neq\cdots\neq i_m} S(k,m) P(Y_{i_1}^{(\ell)} = 1, \cdots, Y_{i_m}^{(\ell)} = 1)$$

Since $P(Y_{i_1}^{(\ell)} = 1, \cdots, Y_{i_m}^{(\ell)} = 1) = P(D_{i_1} = \ell, \ldots, D_{i_m} = \ell)$,

$$E(Z_\ell^k) = \sum_{m=1}^{k} \binom{k}{m} S(k,m) \sum_{\mathbf{e}\in\mathcal{E}_m} p^{|\mathbf{e}|}(1-p)^{\binom{m}{2}-|\mathbf{e}|} \times \prod_{i=1}^{m} \binom{n-m}{\ell - \sum_{j\in[n]\setminus\{i\}e_{ij}}} p^{\ell-\sum_{j\in[n]\setminus\{i\}} e_{ij}} (1-p)^{n-m-\ell+\sum_{j\in[n]\setminus i} e_{ij}},$$

by Theorem 6.1.8. □

Specializing to $k = 1, 2$, we obtain the first two moments.

Proposition 6.1.10. *The mean and the variance of Z_ℓ are given by*

$$E(Z_\ell) = n\binom{n-1}{\ell} p^\ell (1-p)^{n-1-\ell},$$

and

$$V(Z_\ell) = n\binom{n-1}{\ell} p^\ell (1-p)^{n-1-\ell} + 2!\binom{n}{2}\left[\left((n-1-\ell)\binom{n-2}{\ell}p + \ell\binom{n-2}{\ell-1}(1-p)\right) \times \frac{p^{\ell-2}(1-p)^{n-2-\ell}}{n-1}\right]\binom{n-1}{\ell}p^\ell(1-p)^{n-1-\ell} - \left[n\binom{n-1}{\ell}p^{\ell-1}(1-p)^{n-1-\ell}\right]^2$$

Proof. Follows immediately from Proposition 3.2.1 and 6.1.6. □

This result allows us to capture the moments of cliques induced by a single variable from $\mathcal{V}$ – a special case of the $\mathcal{M}_{max}$ cliques we examined in Chapter 5, where $m = 2, k = 1$. We now turn our attention to the other class of cliques a navigation subgraph and how they arise: triangles in $G(n,p)$.

6.1.3 Clique counts in $G(n,p)$

As discussed throughout this work, closed-form expressions for $r-$clique distributions for random graphs where $r \geq 3$ thus far have been elusive. In this section, we achieve our initial objective that led to the work in Chapters 3 and 4: we derive expressions for the moments of cliques in random graphs.

Proposition 6.1.11. *Let X_r denote the number of cliques of size r in $G \sim G(n,p)$, and let $\mathcal{I}_r$ be the set of all $r-$subsets of $[n]$. For every $r-$subset of $[n]$, let Y_i be the indicator variable recording whether the $r-$subset i forms an $r-$clique in G. Then the raw, central and factorial moments of X_r are*

$$E(X_r^k) = \sum_{m=1}^{k} \sum_{\{i_1,\cdots,i_m\} \subseteq \mathcal{I}_r} S(k,m) p^{e(i_1,\ldots,i_m)},$$

$$E\left((X_r - \mu_r)^k\right) = (-\mu_r)^k + \sum_{\ell=1}^{k} \binom{k}{\ell} (-\mu_r)^{k-\ell} \sum_{m=1}^{\ell} S(\ell,m) \sum_{\{i_1,\ldots,i_m\} \subseteq \mathcal{I}_r} p^{e(i_1,\ldots,i_m)},$$

$$E([X_r]_k) = k! \sum_{\{i_1,\ldots,i_m\} \subseteq \mathcal{I}_r} p^{e(i_1,\ldots,i_m)},$$

where $\mu_r = \binom{n}{r} p^{\binom{r}{2}}$ and $e(i_1,\cdots,i_m)$ is the number of edges induced by the collection of cliques $\{i_1,\ldots,i_m\}$.

Proof. As X_r records the number of $r-$cliques, $X_r = \sum_{i \in \mathcal{I}_r} Y_i$ and

$$X_r = \sum_{i \in \mathcal{I}_r} Y_i$$

by Proposition 3.2.1. By Corollary 4.5.6, the number of edges induced by the collection of $r-$cliques with the underlying node labels given by the collection of $r-$sets $\{i_1, i_2, \ldots, i_m\}$ is

$$e(i_1,\ldots,i_m) = \sum_{J \subseteq [m]} \binom{\gamma_J}{2} + \frac{1}{2} \sum_{J \subseteq [m]} \gamma_J \sum_{I \neq J:\ |I \cap J| \geq 1} \gamma_I,$$

where γ_J is the cardinality of $\Gamma_J = \bigcap_{j \in J} i_j \setminus \left(\bigcup_{j \in \overline{J}} i_j\right)$, as defined in Proposition 4.4.1. Since G is a homogeneous Erdős-Rényi random graph, $E(Y_{i_1} \cdots Y_{i_m}) = p^{e(i_1,\ldots,i_m)}$ and the claims follow. □

Proposition 6.1.11 readily specializes to the work of Bollobas & Erdős (1976) when $r = 1, 2$. For instance, the variance is evaluated as follows.

Corollary 6.1.12. *The variance $V(X_r)$ of the number of cliques of size $r \geq 3$ in $G(n,p)$ is*

$$V(X_r) = \binom{n}{r} p^{\binom{r}{2}} \left[1 - \binom{n}{r} p^{\binom{r}{2}}\right] + \sum_{s=0}^{r-1} \binom{n}{s} \binom{n-s}{r-s} \binom{n-r}{r-s} p^{2\binom{r}{2} - \binom{s}{2}}.$$

Proof. By Proposition 6.1.11, we have

$$\begin{aligned} V(X_r) &= E((X_r - \mu_r))^2) \\ &= (-\mu_r)^2 + \sum_{\ell=1}^{2} \binom{2}{\ell} (-\mu_r)^{2-\ell} \sum_{m=1}^{\ell} S(\ell, m) \sum_{\{i_1,\ldots,i_m\} \subseteq \mathcal{I}} p^{e(i_1,\ldots,i_m)}. \end{aligned}$$

When $\ell = 1$, the summation is

$$-\binom{2}{1}\binom{n}{r} p^{\binom{r}{2}} S(1,1) \left[\sum_{i_1 \in \mathcal{I}_r} p^{\nu(v_1)} \right] = -2\binom{n}{r}^2 p^{2\binom{r}{2}}.$$

When $\ell = 2$, the summation is

$$\binom{n}{r} p^{\binom{r}{2}} + \sum_{s=0}^{r-1} \binom{n}{s}\binom{n-s}{r-s}\binom{n-r}{r-s} p^{2\binom{r}{2} - \binom{s}{2}},$$

where we note that $\binom{n}{s}\binom{n-s}{r-s}\binom{n-r}{r-s}$ is the number of ways to construct an ordered pair (i_1, i_2) of distinct $r-$sets with intersection size s. Therefore, the variance is

$$V(X) = \binom{n}{r} p^{\binom{r}{2}} + \sum_{s=0}^{r-1} \binom{n}{s}\binom{n-s}{r-s}\binom{n-r}{r-s} p^{2\binom{r}{2} - \binom{s}{2}} - \binom{n}{r}^2 p^{2\binom{r}{2}}.$$

□

We note that here that the first two moments of clique counts allow the application of the Lindeberg-Levy Central Limit Theorem to approximate the distribution of the average number of clique counts in a (large) sequence of Erdős-Rényi graphs. More generally, it has been shown (Erdős & Rényi, 1960) that if H is a subgraph of K_n and $Y_n(H)$ is the random variable counting the number of isomorphic copies of H appearing in a realization of $G(n, p)$, then Y_n is asymptotically normal. The original proof relies on the method of moments: if a distribution is uniquely determined by its moments, then the convergence of random variables to the moments of the distribution implies convergence in distribution.

Other researchers have shown several similar results, each with a unique approach to bounding the asymptotic error in the normal approximation. For example, Gilmer & Kopparty (2016) used bounds on the characteristic function of the triangle counts to derive a local limit theorem. More recently, Temcinas et al. (2021, Section 6) present a more nuanced result by using $U-$statistics in tandem with Stein's method to derive a limit theorem that asymptotically bounds the error of their approximation.

For our purposes, the normal approximation $N(\mu_t, \sigma_t^2)$ suffices, where

$$\begin{aligned}
\mu_t &= \binom{n}{3}p^3 \\
\sigma_t^2 &= \binom{n}{3}p^{\binom{3}{2}}\left[1-\binom{n}{3}p^{\binom{3}{2}}\right] + \sum_{s=0}^{3-1}\binom{n}{s}\binom{n-s}{3-s}\binom{n-r}{3-s}p^{2\binom{3}{2}-\binom{s}{2}} \\
&= \binom{n}{3}p^3\left[1-\binom{n}{3}p^3\right] + \sum_{s=0}^{2}\binom{n}{s}\binom{n-s}{3-s}\binom{n-3}{3-s}p^{6-\binom{s}{2}} \\
&= \binom{n}{3}p^3\left[1-\binom{n}{3}p^3\right] + \binom{n}{0}\binom{n}{3}\binom{n-3}{3}p^6 + \binom{n}{1}\binom{n-1}{2}\binom{n-3}{2}p^6 + \\
&\quad \binom{n}{2}\binom{n-2}{1}\binom{n-3}{1}p^5.
\end{aligned}$$

Clearly, the approximation is more accurate if there are more variables present in the dataset.

6.1.4 Clique counts in navigation subgraphs

In this section, we combine the results of the preceding sections to derive closed-form expressions for cliques in navigation subgraphs under Model **M1** and assumptions **A1**, **A2** on the measure of interest w.

Theorem 6.1.13. *Let G be the complete variable graph obtained from model **M1** under the assumptions **A1** and **A2**, where the cutoff value is chosen so that $Pr(F > t) = p$. Let H be the corresponding navigation subgraph. Let C_r, X_r, Z_r be the random variables where*

1. *C_r is recording the number of maximal $r-$cliques in H,*
2. *X_r is recording the number of $r-$cliques in $G(n,p)$, and*
3. *Z_r is recording the number of nodes with degree exactly r in G.*

Then the moments of C_r are given by

$$E(C_r^k) = \begin{cases} E(Z_r^k), & r \geq 4 \\ E((Z_3 + X_3)^k), & r = 3 \end{cases},$$

where the expressions for Z_r and X_3 are as in Propositions 6.1.10 and 6.1.11.

Proof. Since H is the line graph of G, by Whitney's isomorphism theorem (Theorem 2.1.3), maximal cliques in H correspond to exactly one of two cases: either they are induced by a star $K_{1,r}$ from H (hence, a vertex of degree exactly r) or $r = 3$ and they are induced by a triangle in H. □

The results of this section shed light on the moments and behaviour of cliques in navigation subgraphs under some assumptions, and in particular, they demonstrate that maximal cliques arise as a result of one of two occurrences in the variable graph: the existence of a triangle or the existence of a node with nonzero degree. In the following section, we see the ramifications of the results of Chapter 5 in the context of cliques in navigation graphs that do not require the assumptions of pruning or the distribution of the measure of interest w, assumptions **A1** and **A2**.

6.2 On the clique structure of Johnson graphs

In this section, we explore the consequences of Chapter 5 for the study of cliques in Johnson graphs. In particular, we derive the asymptotic distribution of cliques in a Johnson graph and illustrate that for large values of n, almost all cliques are of type $\mathcal{M}_{max}$. We identify the condition for equality of clique counts of the two types of cliques and discuss two different mechanisms for sampling cliques from a Johnson graph.

We begin with a straightforward result that shows that for $m \in \mathbb{N}$ and $r \geq 3$ fixed, almost all $r-$cliques in $J_n(m, m-1)$ are of type $\mathcal{M}_{max}$, as $n \to \infty$.

Proposition 6.2.1. *Fix $m \in \mathbb{N}$, and $r \geq 3$. Then the number of $r-$cliques in $J_n(m, m-1)$ is dominated by cliques of type $\mathcal{M}_{max}$. That is,*

$$\frac{\binom{n}{m-1}\binom{n-m+1}{r}}{\binom{n}{m+1}\binom{m+1}{r} + \binom{n}{m-1}\binom{n-m+1}{r}} \to 1,$$

as $n \to \infty$.

Proof. First, we note that

$$1 \geq \frac{\binom{n}{m-1}\binom{n-m+1}{r}}{\binom{n}{m+1}\binom{m+1}{r} + \binom{n}{m-1}\binom{n-m+1}{r}}$$

and

$$\begin{aligned}\frac{\binom{n}{m-1}\binom{n-m+1}{r}}{\binom{n}{m+1}\binom{m+1}{r} + \binom{n}{m-1}\binom{n-m+1}{r}} &= 1 - \frac{\binom{n}{m+1}\binom{m+1}{r}}{\binom{n}{m+1}\binom{m+1}{r} + \binom{n}{m-1}\binom{n-m+1}{r}} \\ &\geq 1 - \frac{\binom{n}{m+1}\binom{m+1}{r}}{\binom{n}{m-1}\binom{n-m+1}{r}}.\end{aligned}$$

Therefore, it suffices to show that

$$\frac{\binom{n}{m+1}\binom{m+1}{r}}{\binom{n}{m-1}\binom{n-m+1}{r}} = o(1).$$

Recall that for $k \leq n$ fixed, we have

$$\binom{n}{k} = \Theta(n^k)$$

and so

$$\frac{\binom{n}{m+1}\binom{m+1}{r}}{\binom{n}{m-1}\binom{n-m+1}{r}} = \frac{\Theta(n^{m+1})\Theta(1)}{\Theta(n^{m-1})\Theta(n^r)} = \Theta(n^{2-r}) = O(n^{-1}) = o(1),$$

as claimed above. □

Proposition 6.2.1 shows that as the size of a base set grows, $\mathcal{M}_{max}$ cliques will constitute the majority of cliques in $J_n(m, m-1)$. This leads to a related question: when are the distributions of $\mathcal{M}_{max}$ and $\mathcal{M}_{min}$ cliques equal? In other words, for which values of n and m are the counts of $\mathcal{M}_{min}$ and $\mathcal{M}_{max}$ cliques equal? We begin by solving a related but easier problem which will motivate our technique for finding the solution to this problem.

Proposition 6.2.2. *The only solution to the system*

$$\binom{n}{m-1}\binom{n-m+1}{r} = \binom{n}{m+1}\binom{m+1}{r} \tag{6.3}$$

for $r \geq 0$ is given by $n = 2m$.

Proof. Let $F_{max}(q)$ and $F_{min}(q)$ denote the generating series for the left and right hand sides of Equation 6.3, respectively. By Theorem 2.3.3, these generating series can be written compactly as follows

$$\begin{aligned} F_{max}(q) &= \binom{n}{m-1}\sum_{r=0}^{\infty}\binom{n-m+1}{r}q^r \\ &= \binom{n}{m-1}(1+q)^{n-m+1}, \\ F_{min}(q) &= \binom{n}{m-1}\sum_{r=0}^{\infty}\binom{m+1}{r}q^r \\ &= \binom{n}{m-1}(1+q)^{m+1}. \end{aligned}$$

Therefore, the ratio of the two series is

$$\frac{F_{max}(q)}{F_{min}(q)} = \frac{\binom{n}{m-1}}{\binom{n}{m+1}}(1+q)^{n-2m}.$$

For this ratio to equal to 1, it must be that $[q^s]\frac{F_{max}(q)}{F_{min}(q)} = 0$ for all $s > 0$. If $n > 2m$, then

$$(1+q)^{n-2m} = \sum_{r=0}^{n-2m} \binom{n-2m}{r} q^r,$$

and at least one positive power of q has a non zero coefficient since

$$[q^1](1+q)^{n-2m} = \frac{\binom{n}{m-1}}{\binom{n}{m+1}} \binom{n-2m}{1} > 0,$$

whenever $n > 2m$. On the other hand, if $n < 2m$, we set $\ell = 2m - n$ and note that by Theorem 2.3.3,

$$(1+q)^{n-2m} = \frac{1}{(1+q)^\ell} = \sum_{r \geq 0} \binom{r+\ell-1}{r} (-1)^r q^r.$$

Since ℓ is an integer and $\ell > 0$, we see that

$$\binom{r+\ell-1}{r} \neq 0,$$

for all $r \geq 0$. Therefore,

$$\frac{F_{max}(q)}{F_{min}(q)} = \sum_{r \geq 0} \binom{r+\ell-1}{r} (-1)^r q^r,$$

has a non-zero coefficient for all q^s with $s \geq 0$. If $\ell > 1$, then

$$[q^s]\frac{F_{max}(q)}{F_{min}(q)} = \frac{\binom{n}{m-1}}{\binom{n}{m+1}} \binom{s+\ell-1}{s} (-1)^s \neq 0,$$

and thus the ratio cannot be equal to 1. So, we can conclude that the only candidate for which we may see equality in the generating series is when $n = 2m$.

Next, we prove that $n = 2m$ ensures for equality of the two generating series.

If $n = 2m$, then since $2m - (m+1) = m - 1$, we have that $\binom{n}{m-1} = \binom{n}{m+1}$ and

$$\frac{F_{max}(q)}{F_{min}(q)} = \frac{\binom{n}{m-1}}{\binom{n}{m+1}} (1+q)^{2m-2m} = (1+q)^0 = 1,$$

as needed. □

Corollary 6.2.3. *Let $G = J_n(m, m-1)$. The distribution of cliques in G of type $\mathcal{M}_{max}$ is equal to the distribution of type $\mathcal{M}_{min}$ if and only if $n = 2m$.*

Proof. We remark that in order to solve the system

$$\binom{n}{m-1}\binom{n-m+1}{r} = \binom{n}{m+1}\binom{m+1}{r}$$

for all $r \geq 3$, it is sufficient to solve the system

$$\binom{n}{m-1}\binom{n-m+1}{r}r(r-1)(r-2) = \binom{n}{m+1}\binom{m+1}{r}r(r-1)(r-2), \tag{6.4}$$

for all $r \geq 3$. The advantage of the latter system is that it has a generating function with a factorization as a product of generating series, as we show below.

Let $F_{max}(q)$ and $F_{min}(q)$ be as in the proof of 6.2.2. Let $f_{max}(q)$ and $f_{min}(q)$ be the generating series defined by

$$\begin{aligned}
f_{max}(q) &= q^3\frac{\partial^3}{\partial q^3}F_{max}(q) \\
&= q^3(n-m+1)(n-m)(n-m-1)\binom{n}{m-1}(1+q)^{n-m-2} \\
&= \binom{n}{m-1}\sum_{r=3}^{\infty}\binom{n-m+1}{r}r(r-1)(r-2)q^r, \\
f_{min}(q) &= q^3\frac{\partial^3}{\partial q^3}F_{min}(q) \\
&= q^3(m+1)m(m-1)\binom{n}{m+1}(1+q)^{m-2} \\
&= \binom{n}{m+1}\sum_{r=3}^{\infty}\binom{m+1}{r}r(r-1)(r-2)q^r,
\end{aligned}$$

We note that n and m are solutions to system (6.4) if and only if $f_{max}(q) = f_{min}(q)$. So, dividing $f_{max}(q)$ by $f_{min}(q)$ and examining the conditions under which this generating series equals 1 we see that

$$\begin{aligned}
\frac{f_{max}(q)}{f_{min}(q)} &= \frac{q^3(n-m+1)(n-m)(n-m-1)\binom{n}{m-1}(1+q)^{n-m-2}}{q^3(m+1)m(m-1)\binom{n}{m+1}(1+q)^{m-2}} \\
&= \frac{(n-m+1)(n-m)(n-m-1)\binom{n}{m-1}(1+q)^{n-2m}}{(m+1)m(m-1)\binom{n}{m+1}}
\end{aligned}$$

We may reuse the argument from Proposition 6.2.2 and note that since $[q^s]\dfrac{f_{max}(q)}{f_{min}(q)} = 0$, it must be that $(1+q)^{n-2m} = 1$ and $n = 2m$.

It is straightforward to verify that the other terms yield the proper cancellation when $n = 2m$. □

Additionally, when the distribution of $\mathcal{M}_{min}$ and $\mathcal{M}_{max}$ cliques are identical, we are immediately able to both identify the count and the size of the most common class of clique.

Proposition 6.2.4. *Let $G = J_{2m}(m, m-1)$. Then the largest count of $r-$cliques occurs when $r = \frac{m+1}{2}$, for m odd and for m even, the mode*

$$r = \begin{cases} \frac{m+1}{2} & \text{When } m \text{ is odd.} \\ \lceil \frac{m+1}{2} \rceil, \lfloor \frac{m+1}{2} \rfloor, & \text{Else} \end{cases}$$

Proof. Since $n = 2m$, the count of cliques of size r is given by

$$2\binom{2m}{m+1}\binom{m+1}{r}.$$

This is maximized when $\binom{m+1}{r}$ is maximized which occurs at $r = \frac{m+1}{2}$ for m odd and $\frac{m}{2}, \frac{m}{2}+1$, for m even. □

This section illustrates that as the number of variables in a dataset grows and the dimension of projection is held fixed, the underlying unfiltered navigation graph would consist of almost entirely $\mathcal{M}_{max}$ cliques. Thus, for example, if $m = 2$, the majority of cliques one would encounter on the graph would have an intersection of size 1, hence highlighting a particular variable. On the other hand, if n is fixed and the larger the dimension of projection, the larger of a proportion of cliques encountered are $\mathcal{M}_{min}$ cliques. Thus, cliques would typically highlight $m+1$ collections of variables from the dataset. Finally, when $n = 2m$, the clique counts for all sizes of nontrivial cliques ($r \geq 3$) would be equal among the two clique types.

6.3 Discussion

This chapter has examined the nature of cliques under M1 and M2 of generating subgraphs of navigation graphs. In particular, when the navigation graph has nodes of dimension $m = 2$ and edges correspond to nodes sharing one variable in common, i.e. the navigation graph is isomorphic to the Johnson graph $J_n(2, 1)$, we captured closed-form expressions for moments of clique counts under construction of M1 and assumptions **A1** - **A2**. More generally, we illustrated that whenever $m = k+1$, cliques appearing in subgraphs generated by either M1 or M2 can only have one of two types: the $\mathcal{M}_{min}$ type or $\mathcal{M}_{max}$ type.

In the case when $m = k+1$, the connection established between subgraphs of the complete graph and navigation graphs indicates that cliques are either induced by a collection of k variables (as is the case for $\mathcal{M}_{max}$ cliques) or are generated by a collection of $k+1$ variables that pairwise possess a relationship of interest (as is the case for $\mathcal{M}_{min}$ cliques). In the former case, we speculate that the analyst should investigate the collection of variables in the intersection of the $\mathcal{M}_{max}$ clique. In the latter case, the collection of the variables in the union is of interest and should be examined closely. This intuition stems from our understanding of how cliques in navigation graphs form as a result of two structures in the variable graph: a star and a triangle.

Moreover, we saw that as m is held fixed and n tends to infinity, the proportion of cliques of type $\mathcal{M}_{max}$ in $J_n(m, m-1)$ tends to 1. On the other hand, if $n = 2m$, the distribution of cliques in a Johnson graph is balanced: there are equal number of $r-$cliques of each type for every size $r \geq 3$. While the asymptotic nature of the two clique types is interesting from a mathematical perspective, applying it to navigation graphs remains a challenge. It seems unclear why an analyst may choose to sample cliques uniformly at random from a navigation graph – one of the obvious settings in which these results could be interpreted for the purposes of data exploration.

6.3.1 Limitations

Finally, a number of limitations to this work should to be mentioned. These can be summarized as follows:

1. Generalizability to other parameter families of navigation graphs;
2. Computational considerations;
3. The assumptions underlying the measure of interest w;
4. The clique structure of navigation graphs under **M2**; and,
5. The interpretability of the significance of cliques.

First, the clique count moment expressions derived for navigation subgraphs where $m = 2, k = 1$ under M1 and assumptions **A1** and **A2** rely on the fact that the only subgraph that are mapped to cliques the line graph operator are the triangles and stars. A natural progression for this work is to examine clique count distributions for navigation graphs with $m = k+1$ and more generally, to generalized Johnson graphs $J_n(m, k)$ (where no additional assumptions on m and k are assumed). While we have made some progress on this problem in Section 7.1.2, deriving clique the moments of clique counts in this general case remains elusive. Thus, the following question arises:

Problem 6.3.1. *What is the distribution of clique counts for **M1** under assumptions **A1** and **A2** where $m \geq k$ are arbitrary?*

Second, while we derived closed-form expressions for the moments of $r-$clique counts and nodes of degree ℓ, these are computationally challenging in most practical cases. For instance, the expression in Proposition 6.1.8 has k summands adding up to $\sum_{\ell=0}^{k} 2^{\ell} = 2^{k+1} - 1$ terms. Therefore, there is an impetus to conduct research on how the expressions can be either simplified or approximated to develop reliable computational methods for quantifying the outlying nature of the cliques, either by size of clique or by quantity.

Problem 6.3.2. *How can the clique count moment expressions from Theorem 6.1.13 be simplified or approximated?*

Third, the assumptions we made regarding the metric of 'interestingness' in Assumptions **A1** and **A2** are strong. It appears to be unknown if such a measure exists. Nonetheless, the work here serves as a building block for future research on the models of graphs appropriate for various mechanisms of generating navigation subgraphs.

Problem 6.3.3. *Are there measures of interest satisfying **A1** and **A2**? If not, what are some reasonable assumptions that can be made regarding a measure of interest and what is the resulting distribution of the underlying navigation subgraph obtained from **M1** or **M2**?*

Fourth, Proposition 6.1.2 asserts that M2 follows the $G(n, M)$ model of random graphs. However, there appears to be very little in the literature regarding cliques in the $G(n, M)$ model. We summarize the challenge here as follows:

Problem 6.3.4. *What is the clique count distribution of $G(n, M)$ the Erdős-Rényi random graph model with fixed number of edges? What are other possible models for navigation subgraphs arising from M2 and what assumptions do they require?*

Next, further studies into the significance behind the variables appearing in cliques would be worthwhile. For example, in the case of $m = 2, k = 1$, we have shown there is a single variable Y in the intersection of a $\mathcal{M}_{max}$ clique – one might wonder under what circumstances and measures of interest could this be indicative of that Y would be suited to be modelled as a variable dependent on the other variables appearing in the clique.

As another example, what do the different types of cliques indicate when taking into account the measure of interest? Consider the monotonic measure of interest, where the square of the correlation indicates how interesting a relationship between two variates is. It is clear that highly correlated variables should form cliques in the variable graph under method **M1**. Therefore, collections of highly correlated variables will partition the navigation graph according to their respective classes.

Problem 6.3.5. *Given a measure of interest w, what specifically does the presence of different types of cliques indicate regarding the underlying variables?*

7

Related problems

This chapter describes problems related to our investigation and progress we have made in attacking them. The emphasis of this chapter is on interesting problems arising on Johnson graphs and network theory that are tangential to the aims of this thesis. The chapter concludes with a reflection on the past and future of this work.

7.1 Johnson graphs

In this thesis, we investigated the structure of cliques in the Johnson family of graphs $J_n(m,k)$, where $n \geq m = k+1$ and the relationship between variable graphs and Johnson graphs. Now, there are a few remaining open problems regarding the clique structure of generalized Johnson graphs and how larger Johnson graphs could be iteratively constructed from smaller ones.

In the first section, we discuss an approach we used to investigate the clique structure of generalized Johnson graphs using algebraic combinatorics. We introduce MacMahonian operators and describe their utility in tackling enumeration problems by filtering out objects which fail to satisfy desired constraints. In conjunction with the ideas revolving intersecting set families from Chapter 4, we explain how MacMahonian operators and a further development of Andrews et al.'s (2001) could lead to a solution to Godsil & Meagher's (2016) Johnson coclique problem as well as other related clique problems.

Next, we present progress made towards addressing remarks made by Hurley & Oldford (2011a) regarding a construction of Johnson graphs with line graph operators. We demonstrate that after applying a natural projection, the line graph of a generalized Johnson graph becomes the graph sum of smaller generalized Johnson graphs. This result is is then used to illustrate that the Johnson family of graphs can be obtained by iteratively applying the aforementioned projection and the line graph operators.

7.1.1 MacMahon operators and the generalized Johnson clique structure

Let $\mathcal{A}$ denote the set of functions

$$\sum_{s_1 \in \mathbb{Z}} \cdots \sum_{s_r \in \mathbb{Z}} A_{s_1,\ldots,s_r} \lambda_1^{s_1} \cdots \lambda_r^{s_r},$$

with absolutely convergent multisum expansions in an open neighbourhood of the complex circles $|\lambda_i| = 1$. The Omega operators $\underset{=}{\Omega}$ and $\underset{\geqslant}{\Omega}$ are operators on $\mathcal{A}$ which were popularized by Andrews et al. (2001) and are based on MacMahon[1]'s partition analysis.

The action of the operator $\underset{\geqslant}{\Omega}$ on members of $\mathcal{A}$ is given by

$$\underset{\geqslant}{\Omega} \sum_{s_1 \in \mathbb{Z}} \cdots \sum_{s_r \in \mathbb{Z}} A_{s_1,\ldots,s_r} \lambda_1^{s_1} \cdots \lambda_r^{s_r} = \sum_{s_1 \geq 0} \cdots \sum_{s_r \geq 0} A_{s_1,\ldots,s_r}.$$

[1]Percy A. MacMahon (1854 – 1929) was an influential British combinatoralist who made significant contributions to the study of integer partitions and symmetric functions MacMahon (2001)

Thus, $\underset{\geqslant}{\Omega}$ sends all $\lambda_i^{s_i}$ with a negative power to 0 and then evaluates the $\lambda_i^{s_i}$ with a non-negative power to 1. The action of the operator $\underset{=}{\Omega}$ on $\mathcal{A}$ is given by

$$\underset{=}{\Omega} \sum_{s_1 \in \mathbb{Z}} \cdots \sum_{s_r \in \mathbb{Z}} A_{s_1,\ldots,s_r} \lambda_1^{s_1} \cdots \lambda_r^{s_r} = A_{0,\ldots,0}.$$

Example 7.1.1. Consider the problem of finding nonnegative integer solutions (a_1, a_2, a_3) to the system

$$\begin{aligned} a_1 + a_2 + a_3 &= k \\ a_1 + a_2 - a_3 &\geq 0, \end{aligned}$$

for some $k \geq 0$ fixed. One approach to enumerating all such solutions (a_1, a_2, a_3) is by identifying the generating series of all tuples (a_1, a_2, a_3) which records both the sum $a_1 + a_2 + a_3$ and the constraint $a_1 + a_2 - a_3$.

Let $F(x, y, z, \lambda)$ denote the generating series for all integer three tuples (a_1, a_2, a_3) which records the difference $a_1 + a_2 - a_3 \geq 0$. Then

$$F(x, y, z, \lambda) = \sum_{a_1,a_2,a_3 \geq 0} x^{a_1} y^{a_2} z^{a_3} \lambda^{a_1+a_2-a_3}.$$

Upon simplifying,

$$\begin{aligned} F(x, y, z, \lambda) &= \sum_{a_1,a_2,a_3 \geq 0} x^{a_1} y^{a_2} z^{a_3} \lambda^{a_1+a_2-a_3} \\ &= \sum_{a_1 \geq 0} (\lambda x)^{a_1} \sum_{a_2 \geq 0} (\lambda y)^{a_2} \sum_{a_3 \geq 0} (\lambda^{-1} z)^{a_3} \\ &= \frac{1}{1 - x\lambda} \frac{1}{1 - y\lambda} \frac{1}{1 - \frac{z}{\lambda}}. \end{aligned}$$

Let $f(x, y, z)$ denote the generating series

$$f(x, y, z) = \sum_{\substack{a_1,a_2,a_3 \geq 0 \\ a_1+a_2-a_3 \geq 0}} x^{a_1} y^{a_2} z^{a_3},$$

and hence

$$f(x, y, z) = \underset{\geqslant}{\Omega} F(x, y, z, \lambda) = \underset{\geqslant}{\Omega} \frac{1}{1 - x\lambda} \frac{1}{1 - y\lambda} \frac{1}{1 - \frac{z}{\lambda}}.$$

By Theorem 2.1 (Andrews et al., 2001), when $n = 2, m = 1, a = 0$:

$$f(x, y, z) = \underset{\geqslant}{\Omega} \frac{1}{1 - x\lambda} \frac{1}{1 - y\lambda} \frac{1}{1 - \frac{z}{\lambda}} = \frac{1 - xyz}{(1 - x)(1 - y)(1 - xz)(1 - yz)}.$$

Finally, we find that the number of integer solutions to $a_1 + a_2 + a_3 = k$ for a fixed $k \geq 0$ where $a_1 + a_2 \geq a_3$ is given by

$$[q^n]f(q,q,q) = [q^k]\frac{1-q^3}{(1-q)^2(1-q^2)^2} = [q^k]\frac{1-q^3}{(1-q-q^2+q^3)^2}.$$

The result from the example above is a specialization of one of the major contributions of Andrews et al.'s (2001) work:

Theorem 7.1.2. *For any integer a,*

$$\underset{\geqslant}{\Omega}\frac{\lambda^a}{(1-x_1\lambda)(1-x_2\lambda)\cdots(1-x_n\lambda)} = \begin{cases} \frac{1}{(1-x_1)(1-x_2)\cdots(1-x_n)}, & a \geq 0 \\ \frac{1}{(1-x_1)(1-x_2)\cdots(1-x_n)} - \sum_{j=0}^{-a-1} h_j(x_1,\ldots,x_n), & a < 0 \end{cases},$$

where $h_j(x_1,\ldots,x_n)$ is the $j-$th complete homogeneous symmetric function and $\underset{\geqslant}{\Omega}$ is the corresponding MacMahon operator (Andrews et al., 2001).

Informally, this is an algebraic description of the simple idea that to solve a problem of interest, we can lump together all feasible configurations of a desired object and then filter out those that do not meet our constraint. Since generating functions can be viewed as the discrete analogues to parameterizations of polynomial systems, the $\underset{\geqslant}{\Omega}$ operators provide another tool for finding integer solutions to complicated systems of equations. Generating series that contain all possible configurations prior to refinement, such as the generating series on the left handside of the equality in Theorem 7.1.2 are known as *crude* generating series. Upon the application of the MacMahon operators, the generating series obtained are often called the *refined* generating series (Xin, 2004).

This has been used to solve complicated constrained integer composition enumeration problems with wide areas of application including discrete geometry (Beck et al., 2013), polyhedral combinatorics (Breuer & Zafeirakopoulos, 2017), diophantine systems of equations (Garsia et al., 2009) and even ODE's and matrix analysis (Neto, 2020).

Currently, the theory is limited to iterative eliminations of constraints; we only eliminate a single constraint at a time in our march towards a general solution. Although this is sufficient for the purposes of solving single constraint problems and verifying solutions in higher dimensional settings, this approach is insufficient in terms of generalization. It would be beneficial to extend the theory from the single constraint case to solve multiple constraints simultaneously by exploring the symmetric functions that appear when evaluating expressions of the form

$$\underset{\geqslant}{\Omega}\frac{\lambda_1^{a_1}\cdots\lambda_k^{a_k}}{(1-x_1\lambda_{i_1}\cdots\lambda_{i_{m_1}})\cdots(1-x_n\lambda_{i_1}\cdots\lambda_{i_{m_n}})}.$$

This would allow for simpler expressions of many problems, and perhaps could lead to a complete solution to open problems such as Godsil's Johnson coclique problem (Godsil & Meagher, 2016):

[Johnson coclique problem] *What is the size of the largest coclique in the Johnson graph $J_n(k,k-1)$ for all n and k?*

Our proposed approach for attacking this open problem is as follows. First, we generalize the notion of type of clique from Chapter 5, where there are only two types of cliques (those with maximal intersection and those with maximal union). More generally, the notion of *type* of clique is captured by the sizes of all of its possible intersections, an idea akin to signatures from Chapter 4. Since intersections are encapsulated through the partition discussed in Proposition 4.4.1, the following proposition follows the same argument as the one after Proposition 4.4.3.

Proposition 7.1.3. *Fix $r \in \mathbb{N}$, then the type of an $r-$clique corresponds to a unique solution to the system*

$$\sum_{J\subseteq[r]} \boldsymbol{\gamma}_J = n \tag{7.1}$$

$$\sum_{i\in J\subseteq[r]} \boldsymbol{\gamma}_J = m, \qquad \forall i \tag{7.2}$$

$$\sum_{i,j\in J\subseteq[r]} \boldsymbol{\gamma}_J = k, \qquad \forall i \neq j, \tag{7.3}$$

where $\boldsymbol{\gamma}_J \geq 0$ for all $J \subseteq [r]$.

Equation 7.1 ensures that each of the n variables in $[n]$ is present either in the $r-$clique or outside of it, and hence the base set n constraint is met. Equation 7.2 ensures that each node has size m while Equation 7.3 ensures that every two nodes in the clique have intersection k. To summarize, each of the three parameters (n, m, k) that appear in $J_n(m, k)$ has a corresponding linear diophantine equation which must be satisfied in order for a clique to be formed.

Proposition 7.1.3 allows us to write down the following expression for the crude, type generating series of *all* possible configurations of types for all parameters n, m, k and clique size r.

Theorem 7.1.4. *The crude generating series for all clique types in all generalized Johnson graphs and generalized Kneser graphs is given by*

$$\Phi(w, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}) = \sum_{r\geq 0} w^r \left(\frac{1}{1 - y_1 y_2 \prod_{i=1}^r \lambda_i \prod_{i,j\in[r] i\neq j} \varepsilon_{i,j}}\right)\left(\frac{1}{1 - y_1 \prod_{i=1}^r \lambda_i}\right) \prod_{J\subseteq[r]} \left[\frac{1}{1 - \frac{x_J}{\prod_{i\in J} \lambda_i \prod_{i,j\in J; i\neq j} \varepsilon_{i,j}}}\right],$$

where y_1 records the node size, y_2 records the intersection size constraint and w records the clique size of interest, $\boldsymbol{\lambda}$ tracks the node size feasibility constraint and $\boldsymbol{\varepsilon}$ tracks the intersection constraint size feasibility constraint.

We describe the two challenges remaining with this approach.

First, we would like to apply $\underset{=\boldsymbol{\varepsilon}}{\Omega} \circ \underset{=\boldsymbol{\lambda}}{\Omega}$ to $\Phi(w, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$. For if we had a closed form expression for

$$\phi_J(w, \boldsymbol{x}, \boldsymbol{y}) := \underset{=}{\Omega}_{\boldsymbol{\varepsilon}} \underset{=}{\Omega}_{\boldsymbol{\lambda}} \Phi(w, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}),$$

then $[q^n y_1^m y_2^k]\phi_J(w, \boldsymbol{x}|_{x=q}, \boldsymbol{y})$ is a finite degree polynomial in w. In particular, the degree of the polynomial corresponds to the clique number of $J_n(m, k)$, by construction. To our knowledge, closed-form expressions of the clique number of a generalized Johnson graph have yet to be discovered. Second, we would like to apply $\underset{\geqslant\boldsymbol{\varepsilon}}{\Omega} \circ \underset{=\boldsymbol{\lambda}}{\Omega}$ to $\Phi(w, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ since this refined generating series would hold the clique number of generalized Kneser graph $KG_n(m, k)$. In particular, let ϕ_K denote the series

$$\phi_K(w, \boldsymbol{x}, \boldsymbol{y}) := \underset{\geqslant}{\Omega}_{\boldsymbol{\varepsilon}} \underset{=}{\Omega}_{\boldsymbol{\lambda}} \Phi(w, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}),$$

then the term $[q^n y_1^m y_2^k]\phi_K(w, \boldsymbol{x}|_{x=q}, \boldsymbol{y})$ is a finite degree polynomial in w whose degree is the clique number of $KG_{n,m}$. We now explain how the clique number of $KG_{n,m}$ solves the Johnson coclique number problem.

Two vertices x, y in $J_n(m, m-1)$ are adjacent if and only if $|\nu(x) \bigcap \nu(y)| = m-1$. Therefore, x and y are not adjacent in $J_n(m, m-1)$ if and only if $|\nu(x) \bigcap \nu(y)| \leq m-2$. Thus, the complement of $J_n(m, m-1)$ is $KG_n(m, m-2)$. So, if H is coclique in the graph $J_n(m, m-1)$ if and only if H is a clique in $KG_n(m, m-2)$ and the coclique number of $J_n(m, m-1)$ is equal to the clique number of $KG_n(m, m-2)$.

7.1.2 Line graphs of Johnson graphs

Hurley & Oldford (2011a) investigated the graphs produced by applying the line graph operator (and a suitable projection) on Johnson graphs. After noting that $J_n(3, 2)$ is isomorphic to a graph obtained by deleting repeated elements from $L(L(K_n))$, the authors wrote:

> Many interesting graphs seem to be built up from a completre graph through the three operators $L(\cdot)$, $R(\cdot)$ and complement. It would be of interest to know what graph properties, if any, are preserved and/or created by these operations, singly and in composition.

In this section, we extend their findings by showing that after applying a suitable projection, the line graph of a generalized Johnson graph can be decomposed into a graph sum of $(m-k-1)$ generalized Johnson graphs on the system of all $(2m-k)$-subsets of n.

Recall that we call a tuple (A, f_A) a *multiset* if A is a set and $f_A : A \to \mathbb{N}$ is a function recording the number of times an element appears in A. We say that (A, f_A) is finite if A is a finite set. Throughout this section, we denote a finite multiset A using the notation

$$\{a_1^{n_1}, \ldots, a_r^{n_r}\},$$

where n_i is $f(a_i)$ the number of times a_i appears in the multiset A. We note that like a set, the order in which a multiset's elements are presented does not matter. Unlike a set, however, elements may appear multiple times. We say that two multisets (A, f_A) and (B, f_B) are equal if and only if for every $i \in A \cup B$, $f_A(i) = f_B(i)$.

For instance, consider the multisets

- (A, f_A) where $A = \{1, 2, 3\}$, $f_A(1) = 1, f_A(2) = 4, f_A(3) = 2$; and,
- (B, f_B) where $B = \{1, 2, 3\}$, $f_B(1) = 2, f_B(2) = 2, f_B(3) = 3$.

Although A and B are equal as sets, the two multisets (A, f_A) and (B, f_B) are not equal.

Definition 7.1.5. Let $\mathcal{R}$ denote the projection operator from the set of all multisets of $[n]$ onto the set of all subsets of $[n]$ which acts on a multiset by replacing multiplicities greater than 1 with 1. That is,

$$\mathcal{R}(\{i_1^{(\ell_1)}, i_2^{(\ell_2)}, \ldots, i_r^{(\ell_r)}\}) = \{i_1, i_2, \ldots, i_r\}.$$

This operator induces an action on a graph G by contracting vertices. That is, suppose one has a collection of vertices of the form

$$\{\ \{i_1^{(\ell_1)}, \ldots, i_r^{(\ell_r)}\} : \ \ell_j \geq 1, \forall j = 1, \ldots, r\}\ \},$$

then all such sets would contract into a single vertex corresponding to the $m-$set $\{i_1, i_2, \ldots, i_r\}$. Moreover, two vertices $\{i_1, i_2, \ldots, i_r\}$ and $\{j_1, j_2, \ldots, j_m\}$ are adjacent in the $\mathcal{R}(G)$ if there are two vertices

$$v_1 = \{i_1^{(\ell_1)}, \ldots, i_r^{(\ell_r)}\} : \ \ell_j \geq 1, \forall j = 1, \ldots, r\}$$

and

$$v_2 = \{j_1^{(s_1)}, j_2^{(s_2)}, \ldots, j_m^{(s_m)} : s_j \geq 1, \forall j = 1, \ldots, m\}$$

in G for which $v_1 \sim v_2$.

We note that while the line graph operator does not take into account any additional structure on the nodes, the reduction operator first reduces nodes down to their class representatives and then considers if any classes are adjacent. We investigate the application of these two operators in tandem to see how one may construct new Johnson graphs from old ones.

The following theorem describes the action of the line graph and projection operators as they apply unto generalized Johnson graphs and generalizes the remarks of Section 3.5 of Hurley & Oldford (2011a).

Theorem 7.1.6. *For all $n > m > k$ positive integers,*

$$\begin{aligned}\mathcal{R}(L(J_n(m,k)) &\cong J_n(2m-k, m) \bigcup J_n(2m-k, m+1) \bigcup \cdots \bigcup J_n(2m-k, 2m-k-1) \\ &\cong KG_n(2m-k, 2m-k-1) \bigcap KG_n(2m-k, m-1).\end{aligned}$$

Proof. Let H denote the graph $L(J_n(m,k))$. To prove the claim, we must show that $\mathcal{R}(H)$ consists of all of the $(2m-k)-$subsets of $[n]$ and that two nodes are adjacent if and only if the two corresponding sets have an intersection of size m.

First, we show that $\mathcal{R}(H)$ consists of nodes of the form $A \cup e \cup B$, where e is a subset of $[n]$ of size k and $A, B \subset [n] \setminus e$ are disjoint sets of size $m-k$. To this end, fix $e \subseteq [n]$ of size k let A, B be two disjoint sets $A, B \subset [n] \setminus e$ of size $m-k$. Then $\nu(v_1) = A \cup e$ and $\nu(v_2) = B \cup e$ are two adjacent nodes in $J_n(m,k)$. Applying the line graph operation to this edge produces a node in H which has the form

$$\nu(v) = \{x^{(1)} : x \in A\} \cup \{y^{(1)} : y \in B\} \cup \{i^{(2)} : i \in e\}.$$

Applying the reduction operator $\mathcal{R}$, we find

$$\mathcal{R}(\nu(v)) = A \cup e \cup B.$$

Since every node in $\mathcal{R}(H)$ is constructed through identifying it with an edge in $J_n(m,k)$, the first claim follows:

$$|\nu(v)| = |A \cup e \cup B| = |A| + |B| + |e| = 2(m-k) + k = 2m - k,$$

for all $v \in V(\mathcal{R}(H))$.

Next, we must show that v_1 and v_2 are adjacent in $\mathcal{R}(H)$ if and only if they intersect in at least m elements.

Suppose that $|\nu(v_1) \cap \nu(v_2)| \geq m$. Fix $e \subset \nu(v_1) \cap \nu(v_2)$ of size m and suppose that $v_1 = A_1 \cup e$, $v_2 = A_2 \cup e$, where A_i is the complement of e in $\nu(v_i), i = 1, 2$ and hence has cardinality $m-k$. Fix $f \subset e$ of size $(m-k)$. Since e is of size m, it corresponds to a unique node in $J_n(m,k)$. Now, consider the set $u_1 = (e \setminus f) \cup A_1$. This is a set of size m and hence also a node in $J_n(m,k)$. Since $|e \setminus f| = k$ and A_1 is disjoint from e, we know that the two nodes corresponding to e and u_1 are adjacent in $J_n(m,k)$. Similarly, the two nodes corresponding to e and $u_2 = (e \setminus f) \cup A_2$ are adjacent in $J_n(m,k)$. Therefore, the two edges that connect e with u_1 and e with u_2 must be adjacent in H. However, these edges are precisely v_1 and v_2 after applying the projection operator $\mathcal{R}$.

Conversely, suppose that v_1, v_2 are adjacent in $\mathcal{R}(H)$. Then there exists x, y, z some $m-$subsets of $[n]$ for which $x \sim y, x \sim z$ in $J_n(m,k)$ and $\nu(v_1) = x \cup y, \nu(v_2) = x \cup z$. Now, we claim that $|\nu(v_1) \cap \nu(v_2)| \geq m$. Clearly, $x \subseteq (x \cup y) \bigcap (x \cup z)$ and $|x| = m$ and hence we are done. □

Example 7.1.7. Starting with $J_4(2,1)$, we can apply to the line graph operation to transform edges into vertices. Since the labels of the resulting vertices have variables that appear multiple times, we apply the reduction operator $\mathcal{R}$ to produce a Johnson graph. In this case, since each node has a single variable appearing twice and two variables appearing once, applying $\mathcal{R}$ would convert labels into ones with exactly three variables. Moreover, as we will see in Corollary 7.1.8, two nodes in the resulting graph are adjacent if and only if they intersect in two variables. Thus, $\mathcal{R}(L(J_4(2,1)) = J_4(3,2)$.

A consequence of Theorem 7.1.6 is that the projection of the line graph of a Johnson graph is a Johnson graph with the set size and intersection parameters incremented by 1.

Corollary 7.1.8. *For all positive integers $n > m$,*

$$\mathcal{R}(L(J_n(m, m-1))) = J_n(m+1, m),$$

where we use the convention that $J_n(n, n-1)$ is the graph with a single node $[n]$ and no edges.

Proof. Follows immediately from Theorem 7.1.6 as when $k = m - 1$, the right handside becomes

$$J_n(2m - (m-1), m) = J_n(m+1, m).$$

□

Corollary 7.1.8 implies that starting from K_n, by applying $\mathcal{R} \circ L$ iteratively, we obtain a chain of all the Johnson graphs that exist on $[n]$.

$$K_n \xrightarrow{\mathcal{R}\circ L} J_n(2,1) \xrightarrow{\mathcal{R}\circ L} J_n(3,2) \xrightarrow{\mathcal{R}\circ L} \cdots \xrightarrow{\mathcal{R}\circ L} J_n(n-1, n-2) \xrightarrow{\mathcal{R}\circ L} K_1$$

Figure 7.1: A chain of $\mathcal{R} \circ L$ operations on the complete graph K_n.

This raises a natural question about cliques: how does the structure and count of cliques change as we apply the $\mathcal{R} \circ L$ operators?

Moreover, it would be interesting to investigate how clique count changes when we consider a chain of graphs of the form $(G_0 := G, G_1, G_2, \ldots, G_\ell)$ where $G_i = \mathcal{R}(L(G_{i-1}))$ and G is any graph. Even a mild modification to the original graph G_0 poses a challenge in tracking how the clique counts evolve through iterative applications of the reduction and line graph operators.

Another interesting thread that came from the work above was due to Theorem 7.1.6. In order to find what chains of iterative applications of $\mathcal{R} \circ L$ produce, it appears that one must investigate the action of the line graph operator on union of graphs. We believe this is tractable and intend to investigate the line graph of unions of graphs in the future.

7.2 Network theory and related problems

In this section, we suggest directions for future research related to the foundations laid out in Chapters 3 and 4. This thesis focused on cliques as proxies for the presence of interesting relationships between variables in data. Here, we propose other candidates for the subgraph structure of interest, describe a gap in the random graph with community structure literature which could be addressed by Bernoulli sums framework and advocate for the use of combinatorial methods in network theory.

First, as we noted in Chapter 2, the clique structure is rather restrictive and therefore, it is natural to wonder how our theory can be generalized to other structures. Thus, we discuss some of the graphic structures which generalize the notion of cliques in graphs which could capture other notions of community.

Next, we describe the connections between cycles in random graphs and pseudorandom graphs and motivate why capturing cycles through the methods of Chapter 4 is difficult.

Additionally, we propose two generalizations to our work which involve relaxation of the underlying graph assumptions. In the former, we turn to directed graphs and provide an overview of the opportunity for generalizing our methods to directed network motifs. In the latter, we discuss some of the success our Bernoulli sums framework has had in the random graphs with community structure setting.

Finally, we suggest broad directions for the application of combinatorial methods in the study of networks.

7.2.1 Generalizations of the notion of the clique

A natural avenue for future research is the choice of subgraph used for detecting interesting patterns. Although cliques exemplify the concept of a cohesive group, they are overly stringent in practical scenarios. In other words, one might not need all connections to exist between every pair of elements to determine that a group of nodes is cohesive.

Since Harary & Ross's (1957) work, other graph-theoretic notions have been introduced to address the rigidity of the clique model. For instance, Alba (1973) studied the notion of a sociometric clique of diameter n, which today is sometimes referred to as an $n-$clique, a subgraph of size n of the network where every pair of nodes are at most distance n from one another in G. Mokken et al. (1979) examined n-clans: sets of vertices in G that induce an $n-$clique in G. In other words, the clan members must satisfy the shortest path condition in the $n-$clique definition. As another example, Kitsak et al. (2010) studied k-cores on complex networks: maximal connected subgraphs of the network where all nodes have degree at least k – a form of a relaxation of the clique model. Surprisingly, Kitsak et al. (2010) illustrated the nodes with the most 'efficient' spreading capacity (for example, such as the spread of ideas or infectious disease) are not necessarily those with the highest connectivity, but rather those located within one of the k-cores of a network.

All of the aforementioned clique generalizations could lead to insights about the variables that they enclose. Thus, one direction of future research would be to examine the distributions of $k-$cliques, $k-$clans and $k-$cores in navigation graphs as well as what their members might suggest about the data.

7.2.2 Cycles and pseudorandom graphs

More broadly, there are other graphic structures that would be interesting to study using the ideas we developed in this work. In particular, we believe that extending of the

disentanglement idea from Chapter 4 to cycles in random graphs is compelling for several reasons. From an enumerative combinatorial point of view, there is a formidable challenge in the derivation of expressions for the number of edges present in a collection of cycles. The challenge is due to the encoding of a cycle, where unlike with the encoding of cliques, the order in which vertices appear matters. For graph and network theorists, this problem is intriguing due to the relationship between pseudographs and random graphs.

As mentioned in Chapter 2 and illustrated throughout this work, random graphs have received a lot of attention due to their simplicity in assumptions, the utility of their modelling applications and the alluring difficulty in solving seemingly simple problems regarding their structure. The ubiquity of random graphs motivates the following question: what are the essential properties of a random graph and how can we tell when a given graph appears to be pseudorandom? This is akin to the statistical problem of evaluating the 'randomness' of a random number generator.

In graph theory, a graph is called **pseudorandom** if it obeys certain properties that random graphs have with high probability. One of these properties is the number of 4-cycles in the graph (Chung et al., 1989). Thus, understanding the distribution of cycles on a random graph would shed light on when a graph is pseudorandom. Moreover, since cycles lie in the kernel of the boundary map ∂ of the graph chain complex, this problem is also interesting from a topological point of view (Carlsson & Vejdemo-Johansson, 2021).

7.2.3 Directed networks motifs

The idea of examining a particular subgraph structure, such as the clique in our investigation of random graphs, navigation graphs and Johnson graphs, is sometimes referred to in the literature by the name *network motif* (Alon, 2007). As touched upon in the previous sections, there are other intriguing candidates for network motifs, especially in the case where the simple graph assumption is relaxed.

In this thesis, we studied the distribution of cliques in the Erdős-Rényi random graph $G(n, p)$, we used two tools: Bernoulli sums framework (Chapter 3), and a carefully constructed partition that captures the clique intersections (Chapter 4). While the Bernoulli sums framework applies regardless of the assumptions made regarding the graph theoretic structure, the disentanglement idea from Chapter 4 implicitly relies on all edges being indistinguishable from one another. Therefore, to examine more nuanced network motifs, such as the different triad isomorphism classes, further research is needed to describe how collections of motifs overlap.

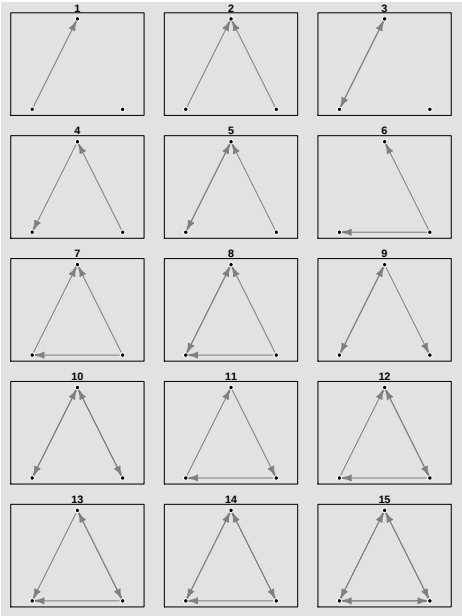


Figure 7.2: The 15 non-trivial triad network motifs.

To the best of our knowledge, the only nontrivial isomorphism classes with fully derived first and second order moments are the directed 3-cliques (Wasserman, 1977). Thus, the nature of even relatively simple motifs in directed graphs remains unknown.

The interpretation of the various subgraphs that could appear in navigation graphs that have a directed structure could be reminiscent of the one from graphical models, where graph-based representations are used to capture the conditional dependencies between random variables.

7.2.4 Random graphs with community structure

The framework from Chapter 3 appears to be a viable tool for attacking many problems on random graphs. For instance, in networks where nodes are classified according to a community structure, one might be interested in measuring the connectedness of nodes within the same class.

The *dyadicity* D_A of a class A in a network is a measure of the connectedness of nodes within the same class compared to what it should be in a random configuration of the network (that is, the average connectedness I would expect to see in a random graph generated from $G(n, N_E)$). So, if Y_{ij} records the adjacency of i and j, the dyadicity of A is

$$D_A = \frac{\sum_{\{i,j\}\subset A} Y_{ij}}{\binom{|A|}{2} c},$$

c is the network connectedness of G.

We derived the following expression for the first moment of edge dyadicity using the tools we developed in Chapter 3.

Theorem 7.2.1. *Suppose that G is an inhomogeneous Erdős-Rényi graph on n nodes with independent edge inclusions and where edge e is included in G with probability p_e. Let N_E denote the total number of edges in G. Suppose that the nodes are partitioned into communities A and B according to some sets $A \neq \emptyset \neq B$. The expected dyadicity of A is*

$$E(D_A) = \frac{\binom{n}{2}}{\binom{|A|}{2}} \Big[P(\textstyle\sum_{\{i,j\}\subset A} Y_{ij} = 0) P(\sum_{\{i,j\}\not\subset A} Y_{ij} = 0) + \\ \sum_{(m,\ell)\in\mathcal{I}} \frac{\ell}{m+\ell} P(\textstyle\sum_{\{i,j\}\subset A} Y_{ij} = \ell)\ P(\sum_{\{i,j\}\not\subset A} Y_{ij} = m) \Big],$$

where $\mathcal{I} = \{(m,\ell) : m, \ell \geq 0, \ell \leq \binom{|A|}{2}, m + \ell \leq \binom{n}{2}, (m,\ell) \neq (0,0)\}$.

Results of this nature indicate the expected behaviour of complex models for networks. Deviations from the expected behaviour could be examined to detect anomalies in networks using classic statistical inference.

Since homogeneous Erdős-Rényi graphs are a special case of inhomogeneous Erdős-Rényi graphs, this result readily specializes. Moreover, as this only assumes two possible classes for the group membership of the nodes, another direction for extending this result is generalizing the number of classes to any finite number k. Lastly, there are many other statistics of interest on these graphs, such as hetrophilicity and homophily and higher order moments that appear to be tractable for derivation.

7.2.5 Algebraic combinatorics, networks, and infectious disease modelling

There are many problems on random graphs that seem amenable to algebraic combinatorics approaches. For instance, consider the seminal work of Newman et al. (2001) on the probability generating series corresponding to various statistics on random graphs. By using Lagrange's celebrated implicit function theorem (Goulden & Jackson, 1983, Section 1.2), we overcame a computational challenge for the authors described below Equation (27) Newman et al. (2001). Subsequently, we discovered new expressions for the corresponding component size generating series of a network. We have since learned that this challenge had been addressed by Newman using complex analysis (Newman, 2007) .

Nevertheless, we believe there is opportunity for further extensions. For instance, one could examine these generating series through the lens of operations on probability generating functions as described by Miller (2018). This could lead to an algebraic approach for the investigation of the spread of infectious diseases on networks with prescribed degree distributions. In fact, this appears promising for practical applications due to the empirical probability generating function's flexibility as a tool for statistical inference of count data Nakamura & Pérez-Abreu (1993).

Additionally, it would be interesting to identify other network structures that can be uncovered by combinatorial decompositions. There are numerous combinatorial operations on generating functions developed by classic combinatorics (Flajolet & Sedgewick, 2009;

Goulden & Jackson, 1983) and the combinatorial theory of species (Bergeron et al., 1998) that have been applied to graph theory. To the best of our knowledge, the approaches have not yet been applied extensively to network theory.

7.3 Reflection

The research that motivated this dissertation began by experimenting with community detection methods and their potential applications to navigation graphs. However, we realized that the output of community detection algorithms could be too restrictive to identify interesting, possibly unanticipated patterns in data. Community detection methods separate variables into groups that can obfuscate patterns that may exist on overlapping cells of nodes.

Thus, we investigated the graph theoretic archetype of a community: a clique. Despite their simplicity, there are gaps in the research literature on their prevalence in random graphs as well as their behaviour in generalized Johnson graphs. Nonetheless, the relationships between variable graphs, Johnson graphs and certain models of navigation graphs were amenable to an algebraic combinatorial investigation of cliques. The simplicity of cliques and the relationship between navigation graphs and Johnson graphs led to several discoveries, all of which were facilitated through algebraic combinatorial techniques.

Our investigation shows that navigation graphs arising under the Johnson graph model, where $m = k+1$, have only two types of cliques and provide closed-form expressions for the moments of clique counts in special cases. Moreover, this research presents a framework for capturing the moments of count random variables, establishes a connection between clique covers and intersecting families of sets and provides a characterization of the clique structure of Johnson graphs.

In this research, we discovered connections between ostensibly different areas of mathematics. Statistics, algebraic combinatorics, graph theory, probability, and extremal set theory are all a part of the unifying shape of the mosaic of this work. The problems, and at times the tools, from these fields guided our investigation in a non-linear fashion, despite the efforts undertaken here to tell a linear story.

We hope that the tools and problems discussed in this work inspire other researchers to examine the wide variety of compelling problems in this rich area of research.

References

Adcock, A., Carlsson, E., & Carlsson, G. (2013). The ring of algebraic functions on persistence bar codes. *arXiv preprint arXiv:1304.0530*.

Agong, L. A., Amarra, C., Caughman, J. S., Herman, A. J., & Terada, T. S. (2018). On the girth and diameter of generalized Johnson graphs. *Discrete Mathematics*, 341(1), 138–142.

Aigner, M. & Axler, S. (2007). *A course in Enumeration*, volume 1. Springer.

Aktas, M. E., Akbas, E., & El Fatmaoui, A. (2019). Persistence homology of networks: Methods and applications. *Applied Network Science*, 4(1), 1–28.

Alba, R. D. (1973). A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, 3(1), 113–126.

Alon, N. & Spencer, J. H. (2016). *The Probabilistic Method*. John Wiley & Sons.

Alon, U. (2007). Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, 8(6), 450–461.

Andrews, G. E. (1998). *The Theory of Partitions*. Number 2. Cambridge university press.

Andrews, G. E. & Paule, P. (2012). MacMahon's dream. In *Partitions, q-series, and Modular Forms* (pp. 1–12). Springer.

Andrews, G. E., Paule, P., & Riese, A. (2001). Macmahon's partition analysis: The Omega package. *European Journal of Combinatorics*, 22(7), 887–904.

Ashourvan, A., Telesford, Q. K., Verstynen, T., Vettel, J. M., & Bassett, D. S. (2019). Multi-scale detection of hierarchical community architecture in structural and functional brain networks. *PLoS One*, 14(5), e0215520.

Asimov, D. (1985). The grand tour: A tool for viewing multidimensional data. *SIAM journal on scientific and statistical computing*, 6(1), 128–143.

Bannai, E. & Ito, T. (1984). *Algebraic Combinatorics I: Association Schemes*.

Beck, M., Braun, B., & Le, N. (2013). Mahonian partition identities via polyhedral geometry. In *From Fourier analysis and number theory to radon transforms and geometry* (pp. 41–54). Springer.

Bender, C. M., Brody, D. C., & Meister, B. K. (2002). Inverse of a Vandermonde matrix. *preprint*.

Bergeron, F., Bergeron, F., Labelle, G., & Leroux, P. (1998). *Combinatorial species and tree-like structures*. Number 67. Cambridge University Press.

Bick, C., Gross, E., Harrington, H. A., & Schaub, M. T. (2021). What are higher-order networks? *arXiv preprint arXiv:2104.11329*.

Biggs, N. (1993). *Algebraic Graph Theory*. Number 67. Cambridge university press.

Biggs, N., Lloyd, E. K., & Wilson, R. J. (1986). *Graph Theory*. Oxford University Press.

Bollobás, B. (1981). Degree sequences of random graphs. *Discrete Mathematics*, 33(1), 1–19.

Bollobás, B. (2001). *Random graphs*. Number 73. Cambridge university press.

Bollobas, B. & Erdős, P. (1976). Cliques in random graphs. *MPCPS*, 80(3), 419.

Bomze, I. M., Budinich, M., Pardalos, P. M., & Pelillo, M. (1999). The maximum clique problem. In *Handbook of combinatorial optimization* (pp. 1–74). Springer.

Bondy, J. A. & Murty, U. S. R. (2008). Graph Theory. *Graduate texts in Mathematics*.

Bose, R. C. & Shimamoto, T. (1952). Classification and analysis of partially balanced incomplete block designs with two associate classes. *Journal of the American Statistical Association*, 47(258), 151–184.

Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391), 580–598.

Breuer, F. & Zafeirakopoulos, Z. (2017). Polyhedral Omega: A new algorithm for solving linear diophantine systems. *Annals of Combinatorics*, 21(2), 211–280.

Brouwer, A. E., Cohen, A. M., & Neumaier, A. (1989). Distance-regular graphs. (Section 9.1).

Brouwer, A. E., Mills, C., Mills, W., & Verbeek, A. (2013). Counting families of mutually intersecting sets. *The Electronic Journal of Combinatorics*, 20(2), P8(pp.1–8).

Buja, A. & Asimov, D. (1986). Grand tour methods: An outline. In *Proceedings of the Seventeenth Symposium on the interface of computer sciences and statistics on Computer science and statistics* (pp. 63–67).

Carlsson, G. & Vejdemo-Johansson, M. (2021). *Topological Data Analysis with Applications*. Cambridge University Press.

Cavique, L., Mendes, A. B., & Santos, J. M. (2009). An algorithm to discover the k-clique cover in networks. In *Portuguese Conference on Artificial Intelligence* (pp. 363–373).: Springer.

Chakraborti, S., Jardim, F., & Epprecht, E. (2019). Higher-Order Moments Using the Survival Function: The Alternative Expectation Formula. *The American Statistician*, 73(2), 191–194.

Chen, S. X. & Liu, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica*, (pp. 875–892).

Chung, F. R. K., Graham, R. L., & Wilson, R. M. (1989). Quasi-random graphs. *Combinatorica*, 9(4), 345–362.

Cvetkovic, D. M. et al. (1980). Spectra of graphs.

Dang, T. N., Anand, A., & Wilkinson, L. (2012). Timeseer: Scagnostics for high-dimensional time series. *IEEE Transactions on Visualization and Computer Graphics*, 19(3), 470–483.

Dang, T. N. & Wilkinson, L. (2014). Scagexplorer: Exploring scatterplots by their scagnostics. In *2014 IEEE Pacific visualization symposium* (pp. 73–80).: IEEE.

de Montmort, P. R. (1713). *Essay d'analyse sur les jeux de hazard...* J. Quillau.

Diestel, R. (2005). Graph Theory. *Graduate texts in mathematics*, 173.

Erdős, P. & Kleitman, D. J. (1974). Extremal Problems among Subsets of a Set. *Discrete Mathematics*, 8, 281–294.

Erdős, P., Ko, C., & Rado, R. (1961). Intersection Theorems for Systems of Finite Sets.

Quarterly Journal of Mathematics, Second Series, 12, 313–320.
Erdős, P. & Rényi, A. (1959). On Random Graphs. I. *Publicationes Mathematicae*, 6(290-297), 18.
Erdős, P. (1959). Graph theory and probability. *Canadian Journal of Mathematics*, 11, 34–38.
Erdős, P., Faudree, R., & Ordman, E. T. (1988). Clique partitions and clique coverings. *Discrete Mathematics*, 72(1-3), 93–101.
Erdős, P. & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1), 17–60.
Erdős, P. & Szekeres, G. (1935). A combinatorial problem in geometry. *Compositio mathematica*, 2, 463–470.
Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2), 179–188.
Flajolet, P. & Sedgewick, R. (2009). *Analytic Combinatorics*. cambridge University press.
Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3-5), 75–174.
Fréchet, M. (1935). Généralisation du théoreme des probabilités totales. *Fundamenta mathematicae*, 1(25), 379–387.
Fréchet, M. (1943). Sur l'extension de certaines évaluations statistiques au cas de petits échantillons. *Revue de l'Institut International de Statistique*, (pp. 182–205).
Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American statistical association*, 82(397), 249–266.
Fu, L. (2009). Implementation of three-dimensional scagnostics. *Univ. of Waterloo, Dept. of*.
Garsia, A., Musiker, G., Wallach, N., & Xin, G. (2009). Invariants, kronecker products, and combinatorics of some remarkable diophantine systems. *Advances in Applied Mathematics*, 42(3), 392–421.
Gerbner, D. & Patkós, B. (2018). *Extremal Finite Set Theory*. Discrete Mathematics and Its Applications. New York: Chapman and Hall/CRC.
Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, 30(4), 1141–1144.
Gilmer, J. & Kopparty, S. (2016). A local central limit theorem for triangles in a random graph. *Random Structures & Algorithms*, 48(4), 732–750.
Godsil, C. & Meagher, K. (2016). *Erdős-Ko-Rado Theorems: Algebraic Approaches*. Number 149. Cambridge University Press.
Godsil, C. & Royle, G. F. (2001). *Algebraic Graph Theory*, volume 207 of *Graduate Texts in Mathematics*. New York, NY, USA: Springer Science & Business Media.
Godsil, C. D. (1993). *Algebraic Combinatorics*. Routledge.
Goulden, I. P. & Jackson, D. M. (1983). *Combinatorial Enumeration*. Wiley.
Gross, J. L. & Yellen, J. (2003). *Handbook of graph theory*. CRC press.
Hall, J. (1987). A local characterization of the johnson scheme. *Combinatorica*, 7(1), 77–85.
Harary, F. & Ross, I. C. (1957). A procedure for clique detection using the group matrix. *Sociometry*, 20(3), 205–215.
He, Y., Chen, Z., & Evans, A. (2008). Structural insights into aberrant topological patterns of large-scale cortical networks in Alzheimer's disease. *Journal of Neuroscience*, 28(18), 4756–4766.

Hilton, A. J. & Milner, E. C. (1967). Some intersection theorems for systems of finite sets. *The Quarterly Journal of Mathematics*, 18(1), 369–384.

Hofert, M. & Oldford, R. W. (2017). Visualizing Dependence in High-dimensional Data: An Application to S&P 500 Constituent Data. *Econometrics and Statistics*, 8(C), 161–183.

Hofert, M. & Oldford, W. (2020a). Zigzag Expanded Navigation Plots in R: The R Package zenplots. *Journal of Statistical Software, Articles*, 95(4), 1–44.

Hofert, M. & Oldford, W. (2020b). Zigzag expanded navigation plots in R: The R package zenplots. *Journal of Statistical Software*, 95, 1–44.

Hurley, C. & Oldford, R. (2011a). Graphs as navigational infrastructure for high dimensional data spaces. *Computational Statistics*, 26(4), 585–612.

Hurley, C. B. & Oldford, R. W. (2011b). Graphs as Navigational Infrastructure for High Dimensional Data Spaces. *Computational Statistics*, 26(4), 585–612.

Iyer, P. K. (1958). A theorem on factorial moments and its applications. *The Annals of Mathematical Statistics*, 29(1), 254–261.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM computing surveys (CSUR)*, 31(3), 264–323.

Jain, G. & Consul, P. (1971). A generalized negative binomial distribution. *SIAM Journal on Applied Mathematics*, 21(4), 501–513.

Jain, S. & Seshadhri, C. (2020). The power of pivoting for exact clique counting. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 268–276).

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate Discrete Distributions*, volume 444. John Wiley & Sons.

Kadane, J. B. et al. (2016). Sums of possibly associated Bernoulli variables: The Conway–Maxwell-binomial distribution. *Bayesian Analysis*, 11(2), 403–420.

Kahle, M. (2009). Topology of random clique complexes. *Discrete mathematics*, 309(6), 1658–1671.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature physics*, 6(11), 888–893.

Knoblauch, A. (2008). Closed-form expressions for the moments of the binomial probability distribution. *SIAM Journal on Applied Mathematics*, 69(1), 197–204.

Kobak, D. & Shpilkin, S. (2021). Legislative election 2021. GitHub.

Kobak, D., Shpilkin, S., & Pshenichnikov, M. S. (2016). Integer percentages as electoral falsification fingerprints. *The Annals of Applied Statistics*, 10(1), 54–73.

Lancichinetti, A. & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical review E*, 80(5), 056117.

Lerner, J. (2005). Role Assignments. In U. Brandes & T. Erlebach (Eds.), *Network Analysis*, volume LNCS 3418 chapter 9, (pp. 216–252). Berlin: Springer-Verlag.

Lo, C.-Y., Wang, P.-N., Chou, K.-H., Wang, J., He, Y., & Lin, C.-P. (2010). Diffusion tensor tractography reveals abnormal topological organization in structural cortical networks in alzheimer's disease. *Journal of Neuroscience*, 30(50), 16876–16885.

Luby, M. (2002). LT codes. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.* (pp. 271–271).: IEEE Computer Society.

Luce, R. D. & Perry, A. D. (1949). A method of matrix analysis of group structure. *Psychometrika*, 14(2), 95–116.

MacKay, D. J., Mac Kay, D. J., et al. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge university press.

MacMahon, P. A. (2001). *Combinatory Analysis, Volumes I and II*, volume 137. American Mathematical Soc.

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2019). *Robust statistics: Theory and methods (with R)*. John Wiley & Sons.

Martin, G. E. (2001). *Counting: The Art of Enumerative Combinatorics*. Springer.

Meyerowitz, A. (1995). Maximal intersecting families. *European Journal of Combinatorics*, 16(5), 491–501.

Miller, J. C. (2018). A primer on the use of probability generating functions in infectious disease modeling. *Infectious Disease Modelling*, 3, 192–248.

Mokken, R. J. et al. (1979). Cliques, clubs and clans. *Quality & Quantity*, 13(2), 161–173.

Moody, J. (2001). Race, school integration, and friendship segregation in America. *American journal of Sociology*, 107(3), 679–716.

Nakamura, M. & Pérez-Abreu, V. (1993). Empirical probability generating function: An overview. *Insurance: Mathematics and Economics*, 12(3), 287–295.

Neto, A. F. (2020). Matrix analysis and Omega calculus. *SIAM Review*, 62(1), 264–280.

Newman, M. (2018). *Networks*. Oxford University Press.

Newman, M. E. (2007). Component sizes in networks with arbitrary degree distributions. *Physical review e*, 76(4), 045101.

Newman, M. E. et al. (2003). Random graphs as models of networks. *Handbook of graphs and networks*, 1, 35–68.

Newman, M. E., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2), 026118.

Ogyanova, K. (2020). Network visualization with R.

Oldford, R. W. & Waddell, A. (2011). Visual clustering of high-dimensional data by navigating low-dimensional spaces. In *Proc. 58th World Statistical Congress*, volume LVIII (pp. 3294 – 3303). Dublin, Ireland: International Statistical Institute. Special Topics Session STS057.

Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017). Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert Systems with Applications*, 85, 204–220.

Östergård, P. R. (2002). A fast algorithm for the maximum clique problem. *Discrete Applied Mathematics*, 120(1-3), 197–207.

Ouyang, Q., Kaplan, P. D., Liu, S., & Libchaber, A. (1997). DNA solution of the maximal clique problem. *Science*, 278(5337), 446–449.

Pardalos, P. M. & Xue, J. (1994). The maximum clique problem. *Journal of global Optimization*, 4(3), 301–328.

Park, J. & Barabási, A.-L. (2007). Distribution of node characteristics in complex networks. *Proceedings of the National Academy of Sciences*, 104(46), 17916–17920.

Pattillo, J., Youssef, N., & Butenko, S. (2012). Clique relaxation models in social network analysis. In *Handbook of Optimization in Complex Networks* (pp. 143–162). Springer.

Plantié, M. & Crampes, M. (2013). Survey on social community detection. In *Social media*

retrieval (pp. 65–85). Springer.

Poisson, S. D. (1837). *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. Bachelier.

Riordan, J. (1937). Moment recurrence relations for binomial, Poisson and hypergeometric frequency distributions. *The Annals of Mathematical Statistics*, 8(2), 103–111.

Roberts, F. S. (1985). Applications of edge coverings by cliques. *Discrete Applied Mathematics*, 10(1), 93–109.

Ruciński, A. (1988). When are small subgraphs of a random graph normally distributed? *Probability Theory and Related Fields*, 78(1), 1–10.

Shah, B. (1973). Distribution of sum of independent integer valued random-variables. *The American Statistician*, 27(3), 123–124.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway–Maxwell–Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1), 127–142.

Shuldiner, P. & Oldford, R. (2021). Moments of Bernoulli Sums. *arXiv preprint arXiv:2110.02363*.

Shuldiner, P. & Oldford, R. W. (2022a). The Clique Structure of Johnson Graphs. *arXiv preprint arXiv:2208.12710*.

Shuldiner, P. & Oldford, R. W. (2022b). How many cliques can a clique cover cover? *arXiv preprint arXiv:2206.14895*.

Sporns, O. (2013). Network attributes for segregation and integration in the human brain. *Current opinion in neurobiology*, 23(2), 162–171.

Stam, C. J. & Reijneveld, J. C. (2007). Graph theoretical analysis of complex networks in the brain. *Nonlinear biomedical physics*, 1(1), 1–19.

Stanley, R. P. (2011). *Enumerative Combinatorics, Volume I*. 2nd edition.

Temcinas, T., Nanda, V., & Reinert, G. (2021). Multivariate Central Limit Theorems for Random Clique Complexes. *arXiv preprint arXiv:2112.08922*.

The Economist (2021). Russian elections once again had a suspiciously neat result.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2 (pp. 523–531).

Tukey, J. W. (1977). *Exploratory data analysis*, volume 2. Reading, MA.

Tukey, J. W. & Tukey, P. A. (1985). Computer graphics and exploratory data analysis: An introduction. In *Proceedings of the sixth annual conference and exposition: computer graphics*, volume 85 (pp. 773–785).

Waddell, A. (2016). Interactive visualization and exploration of high-dimensional data.

Waddell, A. & Oldford, R. (2018). loon: Interactive Statistical Data Visualization.

Waddell, A. & Oldford, R. W. (2011). Rnavgraph: A visualization tool for navigating through high-dimensional data. *Proc. 58th World Statistical Congress*, (pp. 1852–1860).

Waddell, A. & Oldford, R. W. (2022). *loon: Interactive Statistical Data Visualization*. R package version 1.4.0.

Waddell, A. R. & Oldford, R. W. (2014). *RnavGraph: Using graphs as a navigational infrastructure.* R package version 0.2.0.

Wasserman, S. S. (1977). Random directed graph distributions and the triad census in

social networks. *Journal of Mathematical Sociology*, 5(1), 61–86.

Whitney, H. (1992). Congruent graphs and the connectivity of graphs. In *Hassler Whitney Collected Papers* (pp. 61–79). Springer.

Wilf, H. S. (2005). *generatingfunctionology*. A K Peters/CRC Press.

Wilkinson, L., Anand, A., & Grossman, R. (2005). Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (pp. 157–164).: IEEE.

Xin, G. (2004). A fast algorithm for Macmahon's partition analysis. *The Electronic Journal of Combinatorics*, 11(1), 58.

Yao, Z., Zhang, Y., Lin, L., Zhou, Y., Xu, C., Jiang, T., & Initiative, A. D. N. (2010). Abnormal cortical networks in mild cognitive impairment and Alzheimer's disease. *PLoS computational biology*, 6(11), e1001006.