Simplicial Reaction Networks and Dynamics on Graphs

By

Rachel S. Lawrence

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alistair Sinclair, Chair
Professor Jelani Nelson
Professor Sylvie Corteel

Fall 2023

Simplicial Reaction Networks and Dynamics on Graphs

Abstract

Simplicial Reaction Networks and Dynamics on Graphs

by

Rachel S. Lawrence

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Alistair Sinclair, Chair

Reaction networks are a powerful tool for modeling the behavior of a wide variety of real-world systems, including population dynamics and chemical processes, as well as algorithms for sampling combinatorial objects. While many such systems have well-understood equilibrium states, the long-standing conjecture that these states will always be achieved remains open. This thesis presents the class of simplicial reaction networks, which includes a wide variety of natural combinatorial examples of use in theoretical computer science. It shows how simplicial structures can be used to understand and control the equilibrium behavior of the network as a whole, and discusses related progress towards the Global Attractor Conjecture. Finally, this thesis presents additional work exploring combinatorial approaches to the Inverse Eigenvalue Problem on graphs, including the randomized Zero Forcing algorithm and a lower bound for the Minimum Rank problem.

To mom and dad, who made it possible.

To my EECS cohort, who made it *bear*-able (*go bears!*).

To Hannah and David, who inspired me to keep going.

And to Yuri, who crossed the finish line with me.

Contents

List of Figures

List of Tables

Acknowledgments

It's said that it takes a village to raise a child; I can attest that it also takes a village to raise a Ph.D thesis.

I'd first like to thank my advisor, Alistair Sinclair, for his guidance, patience, and support throughout the past five years, as well as the members of my qualifying exam and thesis committees — Jelani Nelson, Sylvie Corteel, Murat Arcak, and Bernd Sturmfels — for their time and effort in support of this dissertation. I could not have asked for a more supportive department than Berkeley EECS, and this thesis was made possible in large part due to the support and infectious intellectual curiosity of my friends and colleagues here.

I'd also like to recognize my collaborators for their permission to include co-authored material. In particular, Chapters 2 and 3 are based on joint work with Alistair Sinclair; Chapter 4 is based on joint work with Kevin Grace, H. Tracy Hall, and Alathea Jensen; and Chapter 5 is based on joint work with Bryan Curtis, Luyining Gan, Jamie Haddock, and Sam Spiro. My co-authors and I also extend sincere gratitude to the organizers of the Mathematics Research Community on Finding Needles in Haystacks: Approaches to Inverse Problems Using Combinatorics and Linear Algebra, and the AMS for funding the program through NSF grant 1916439.

Most importantly, this thesis would not exist without the love and support of those closest to me. To my parents, Cindy and Kevin, I am forever grateful for all the opportunities they provided, as well as the support for my interests, mathematical and otherwise, and the love of learning they taught me from a young age. And to my siblings, Hannah and David, thank you for always being my fiercest supporters and sharpest adversaries (often on the same day!). I also want to thank all the extended family and friends who showed me love and confidence throughout the process, with a special thanks to those who went to the extra length of attending the dissertation talk (and even pretending to enjoy it!). I especially want to acknowledge a very special group of humans who shared countless late nights, theory retreats, international adventures, and thought-provoking ideas with me over the past few years: Elizabeth, Chinmay, Arun, Tarun, Nick, Efe, Seri, Orr, Grace, Sidhanth, Siqi, Lydia, Katie, Ben, Utkarsha, Arya, Chanthia, Jessica, Mitchell, Chi, Una, Dariya, Elijah, Justin, and Jess, to name a few. And to Alex, in particular: I never would've dreamed of attempting this without your encouragement, and I'm so grateful that we could walk together on this ascent from the very beginning.

I couldn't end this section without giving a special shout-out to the West Berkeley marina, the Hidden Cafe, matcha lattes, the backyard cats at Eighth Street, and everyone who ran long distances around the East Bay with me — all of which were instrumental in maintaining sanity and optimism throughout a pandemic, lockdown, wildfires, and other global strangeness in 2020 and beyond.

And finally, my unending love and gratitude to Yuri, who met me in what should have been the worst of it, and brought out the best. Thank you for showing me the stars, the duplicated spanning tree, endless patience, and the proper way to assemble a tent—and for understanding the importance of doing things right.

Chapter 1

Introduction

My brain is open!

- Paul Erdős, *standard greeting*

This thesis describes work on two primary topics: the theory of chemical reaction networks and the minimum rank problem on graphs. In both areas, we bring ideas from theoretical computer science to bear on problems arising outside the field's traditional scope. In the case of chemical reaction networks, we consider their application in a combinatorial context, and prove that this context guarantees the convergence properties – including a resolution of the long-standing global attractor conjecture in this setting – necessary for new computational sampling applications. Similarly, we examine the minimum rank problem on graphs from a computational perspective, completely characterizing its computational complexity and also studying a more efficient constraint propagation process which provides bounds on its value. Of particular focus will be the examination of convergence properties – the question of whether a given process, over a long enough time, tends towards an equilibrium state – in graphs that arise from problems of combinatorial and computational significance.

Reaction Network Theory

The subject of Chapters 2 and 3 is reaction network theory, a branch of applied mathematics inspired by the behavior of real-world biochemical systems. In these systems, changes to the state are governed by a nonlinear probabilistic process on a graph. Specifically, the system consists of a probability distribution over different categories of items, akin to chemical elements or molecules, whose concentrations change over time. These changes adhere to the law of mass action: molecules react with each other, at a rate determined by the product of concentrations of all inputs (*reactants*), to create a new set of output (*product*) molecules. The system's possible reactions are represented by a network structure, where a directed edge between two sets of molecules signifies their roles as reactants and products in some reaction. For an audience familiar with computer science, these systems are reminiscent of

Markov chains, but with a critical departure from linearity: the change in state is no longer a linear function of the current state, but rather depends on a product of concentrations of any molecules participating in each reaction. Even restricted to the quadratic case, involving only two molecules per reactant and product, these networks prove far more challenging to analyze than linear systems.

Recent studies have brought to light profound connections between chemical reaction network theory and computer science, with the full scope of these relationships still emerging. As Marta Dueñas-Díez and Juan Pérez-Mercader note in [38], "Computation takes place not only in the myriad of electronic devices we use daily, but also in living systems... via chemical reaction mechanisms." In this view, natural systems function as automata, performing computation by transforming initial conditions into steady-state behavior following dynamical rules. Indeed, chemical reaction networks form the foundation of the biological and chemical world, and provably give rise to an expressive model of computation. These networks of chain reactions follow the law of mass action, but the behaviors arising from those simple rules can model all the complexity of Turing Machines and perform general computation [88, 91, 22, 30]. Harnessing the power of mass action dynamics to perform computation opens the door to a world of potential next-generation computing paradigms. In the words of Quanta Magazine's Charlie Wood, "The universe constantly pulls off tasks far beyond the limits of computers' meager bookkeeping abilities"; where simulating a chemical reaction network at scale could quickly become computationally intractable, the process itself can nonetheless be observed to play out through countless examples in nature [92].

In order to harness the power of chemical reaction networks, their mathematical underpinnings require rigorous analysis. Large systems with complex, interconnected networks of interactions have typically proved difficult, if not impossible, to pin down with existing mathematical methods. Even among systems with relatively well understood equilibrium states, such as complex balancing systems, a proof that they behave as expected – converging to the single "obvious" equilibrium on the interior of the state space, rather than to some other collection of boundary states – remains elusive. Without eliminating the possibility of certain pathological behaviors in the limit, it is hard to argue that these systems form a sound basis for a computational paradigm. In particular, the central Global Attractor Conjecture has remained open for more than 50 years, despite a widespread expectation that it will ultimately be resolved in the positive. In its absence, the field lacks guarantees of well-controlled limiting behaviors and equilibria for most systems of interest.

This thesis presents a class of chemical reaction networks called *simplicial reaction networks*, applicable to a rich class of computational problems, and identifies conditions sufficient to guarantee they adhere to well-controlled, predictable behaviors. Simplicial reaction networks are particularly interesting because they model processes on objects familiar to computer scientists and combinatorialists: spanning trees or matchings in a graph, vector spaces, matroids, and most generally, abstract simplicial complexes and the polytopes they define. The resulting reaction networks can thus be used to generate a distribution over large families of combinatorial objects that would often prove difficult to sample otherwise. In this way, we establish a method to harness the computational power of reaction networks to directly

produce a desired distribution, in analogy with well-known Markov chain Monte Carlo methods.

In Chapter 2, we first present a concise background on chemical reaction networks and outline the current state of the art, and then build on prior work to generalize the known convergence results for so-called single linkage [7] and strongly endotactic [50] networks to a broader class of reaction networks. This improvement is accomplished in part through the organization of limiting behaviors into a partial order known as *tiers*. An idea from measurement theory provides a conceptual link to earlier works, showing that a per-species rather than per-complex analysis of tiers yields a stronger version of analogous results. We observe that, while earlier results necessarily fail to apply to reaction networks under downward closure, the new method presented in this thesis in combination with an invariant-based linear programming algorithm succeeds in proving convergence results for some such networks.

Following this, Chapter 3 introduces and establishes the basic properties of simplicial reaction networks, including the space of linear invariants for all such networks. In particular, Section 3.5 introduces a geometric method to analyze simplicial reaction networks based on their associated polytopes, through the observation that points in these polytopes correspond precisely to different settings of invariants in the underlying reaction network. The chapter concludes with convergence results and proofs of the Global Attractor Conjecture for matroid, matching, and spanning tree reaction networks in Sections 3.4, 3.6 and 3.7 respectively.

Minimum Rank and Zero Forcing

The second half of the thesis shifts focus to explore networks from a linear algebraic perspective, motivated by the *minimum rank* and *inverse eigenvalue* problems on graphs. These problems address the following goal: Given an unweighted graph, determine whether there exists a weighting of edges consistent with a given structural condition on the graph's adjacency matrix. In the case of the minimum rank problem, the structural condition is a particular matrix rank, and the problem asks whether there exists a choice of edge weights such that the specified rank is achieved; and furthermore, what the minimum achievable rank is for the given graph [57]. Minimum rank also provides a bound on the related Inverse Eigenvalue Problem (or *IEP-G*), where the structural condition is instead a set of real numbers, and the IEP-G asks whether there exists an edge-weighting such that this set forms the spectrum of the graph's adjacency matrix.

The minimum rank problem is of interest both as a step towards understanding the more general IEP-G, and also as a question of independent interest. Indeed, finding a minimum rank matrix with a given sparsity structure can be understood as determining which edge weights provide the lowest-dimensional explanation for an observed network structure, revealing an application to matrix completion problems. In Chapter 4, we present minimum rank and the IEP-G, and derive new results characterizing the computational complexity of minimum rank via a reduction from the existential theory of the reals. We show that, while finite lists of forbidden subgraphs are sufficient to identify graphs with minimum rank 2 and below, it is

impossible for this technique to apply for any higher ranks. We further show that for $d = 3$ and above, determining whether a real, symmetric matrix has minimum rank at most d is computationally hard—and, in fact, is equivalent to the problem of solving general polynomial systems over the reals. Section 4.5 provides the details of this reduction, giving an explicit construction reducing any polynomial system to a 3-RANK problem. The hardness results presented in this chapter reveal a departure in the $d = 3$ case from the efficient algorithms known for $d \leq 2$, as well as from those known for the equivalent problem over finite fields.

In light of this fact, in Chapter 5 we turn our attention to more readily computable bounds on minimum rank, rather than exact computation. In a return to network dynamics, this chapter investigates the bound on minimum rank obtained from a graph infection process known as *zero forcing*. The zero forcing process can be understood as an initial set of blue vertices propagating through an otherwise white graph; white vertices turn blue when they are the only white neighbor of a blue vertex. Whether the process ultimately reaches all vertices is dependent on the structure of the underlying graph as well as the choice of the initial set of blue vertices, and the distribution of such initial sets reveals surprising connections to other properties of the graph. Towards this end, we introduce the problem of *randomized zero forcing*, analyzing the probability that a randomly chosen initial set of blue vertices is zero forcing – that is, whether the infection process starting with this set ultimately colors the entire graph blue. Our results include a proof that for large n, the probability of selecting a zero forcing set for any tree is upper bounded by the corresponding probability for a path graph; additional bounds on this probability based on vertex degrees; and the resolution of a conjecture due to Boyer et al. [20] regarding the number of zero forcing sets of a given size that any graph can have.

Chapter 2

Reaction Networks and the Global Attractor Conjecture

And the touch of a hand lit the fuse
Of a chain reaction of countermoves

- Taylor Swift, *Mastermind*

2.1 Introduction to Reaction Network Theory

The study of reaction networks originates in the field of mathematical chemistry, particularly the foundational work of Horn, Jackson and Feinberg [59, 43, 46, 44], whose research sought a rigorous understanding of the emergent dynamics of chemical chain reactions. In these systems, chemical reactions enact changes to a distribution over chemical species, mediated by the *law of mass action*: a nonlinear mechanism in which reaction rates are proportional to the product of the current concentrations of their input species. Of particular interest is the convergence of these systems (or lack thereof) to stable equilibrium states. From this line of research, the Global Attractor Conjecture was quickly identified: Any system with the seemingly natural *complex balancing* condition should converge to a unique globally-attracting equilibrium. Yet, despite significant efforts and progress, the conjecture has resisted proof for over fifty years since its inception [50, 33].

Several classically studied systems were found to fit into the reaction network framework, including the Boltzmann equation for ideal gases [51, 19] and Hardy-Weinberg population genetics [55, 2]; as well as the more general context of quadratic dynamical systems [77]. In subsequent years, the theory of reaction networks continued to develop as a model of independent interest to mathematicians, computer scientists, and biologists alike. Applications have ranged from models of drug interaction and analysis of biochemical signal transduction [49, 89, 69] to design of molecular control circuitry [91] to pure mathematical interest in the context of Petri nets and toric dynamical systems [10, 86]. The expressiveness and power of this model are further evident from a computability perspective: Certain forms of

reaction networks have been found to be sufficiently expressive to use as a general-purpose programming language which, when "compiled" into physical systems obeying the law of mass action, produce a state that encodes the outcome of an arbitrary computation [88, 91, 22, 30]. Other works build reaction network-based computational schemes while restricting to more specific domains, including signal processing, machine learning, and computing functions [71, 3, 23, 80]. Chapter 3 will explore a new direction of computation achievable by reaction networks, in which a reaction network is designed to carry out a sampling task by generating a desirable stationary distribution, in analogy to a Markov chain. In contrast to a Markov chain, however, the introduction of a nonlinear dependence on the current state allows for an explosive increase in the complexity of possible behaviors – both a feature (allowing for expressive computation) and a challenge (in proving the convergence properties required to make such a system useful).

Approaching computational problems using reaction networks is a promising new frontier; yet critical mathematical properties of these networks, and specifically convergence to equilibrium, remain far from well understood. What conditions are necessary or sufficient to achieve convergence to a single, stable global attractor? What limiting behaviors are possible, and how quickly do they occur? Which areas in the state space are off-limits for a given system, and which might be visited after enough time has passed? Each of these questions is intimately related to the Global Attractor Conjecture, as well as the question of whether a given dynamical system's trajectory stays sufficiently far away from the boundary of the state space [87, 44]. Using tools spanning nonlinear differential equations, experimental simulation, algebraic geometry, graph theory, and more, incremental progress on these questions continues; this chapter and the next present new work toward that goal. In this chapter, we take a broad view of reaction networks, with the main result providing a sufficient condition for convergence based on network structure – independent of specific values of parameters such as rate constants. This follows and expands on other recent work following a dynamical approach, with some interesting insights applied from the field of measurement theory. The following chapter takes a new perspective, approaching the problem from a combinatorial and geometric angle. There, we define the class of *simplicial reaction networks* to encompass desirable properties for sampling applications in theoretical computer science, and prove analogous convergence results in this setting.

In the current chapter: Section 2.2 introduces the formal definitions of a reaction network, its dynamics, and the trajectory that a system defined by a reaction network follows as it evolves over time, and introduces some mathematical preliminaries for studying their behavior, culminating in the statement of the Global Attractor Conjecture in Section 2.3. Section 2.4 describes the main results of this chapter: a new structural condition based on equivalence classes of complexes and species which is sufficient to guarantee convergence to a unique interior equilibrium point, and an algorithm to determine whether this condition holds. This work both provides a simpler framework for understanding prior results, and expands the class of networks known to satisfy the Global Attractor Conjecture based purely on network structure to include certain quadratic dynamical systems which could not be analyzed with prior techniques.

2.2 Preliminaries

2.2.1 Reaction Networks

A reaction network is a dynamical system in which the concentrations of different *species* evolve over time. The change in concentrations is enacted by *reactions*: operations taking a set of "input" (or "reactant") species, and replacing them with a new set of "output" (or "product") species. We typically refer to the elements of species set $\mathcal{S}$ with capital letters $A, B, C \ldots$, and represent reactions with an arrow, evoking the chemical reactions which initially inspired the model. For example:

$$A + B \xrightarrow{\beta} C + D$$

represents a reaction taking the reactant set $\{A, B\}$ as input, and replacing it with product set $\{C, D\}$. The sets (or, more generally, multi-sets) $\{A, B\}$ and $\{C, D\}$, when used as the products or reactants of a reaction, are known as *complexes*. Each reaction is also equipped with an edge weight β, known as a *rate constant*, which will later be used to indicate the relative speed at which the reaction occurs.

Reactions may also have bidirectional arrows, indicating that the reaction can occur in both the forward and backward directions; and may contain more than one unit of the same species, as demonstrated in the following chemistry-inspired reaction on species set $\{N_2, H_2, NH_3\}$:

$$N_2 + 3H_2 \leftrightarrow 2NH_3$$

A reaction network is composed of species, complexes, and reactions, as well as a differential equation describing the change over time of a state vector x, which describes how much of each species is currently present in the system. To put this formally:

Definition 2.1 (Reaction Network)**.** Let $\mathcal{S}$ be a finite set of species, and $\mathcal{C}$ a finite set of complexes, each of which is a multi-set on ground set $\mathcal{S}$. Let the reaction set $\mathcal{R} \subseteq \mathcal{C} \times \mathcal{C}$ define a set of ordered pairs of complexes. A *reaction network* is defined by the triple $\{\mathcal{S}, \mathcal{C}, \mathcal{R}\}$, along with dynamics $\dot{x} = f(x)$ for $x \in \mathbb{R}^{|\mathcal{S}|}_{\geq 0}$.

For a given species set $\mathcal{S}$ with $|\mathcal{S}| = N$, we index $\mathcal{S}$ by $[N]$, and refer to a species either by name or by its index. Where clear from context, $y \in \mathcal{C}$ may also refer to the incidence vector of species in the complex (in $\{0, 1\}^N$ if each species appears at most once per complex; or in $\mathbb{Z}^N_{\geq 0}$ otherwise).

The complexes and reactions of a reaction network can also be viewed as defining the vertex and edge sets, respectively, of a *reaction graph* $G = (\mathcal{C}, \mathcal{R})$.

For example, Figure 2.1 depicts a reaction graph with three connected components on species set

$$\mathcal{S} = \{A, B, C, D, E, F, G, H, I\}$$

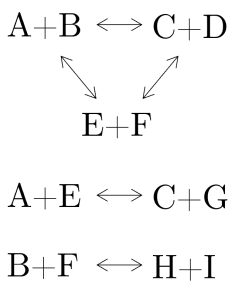


Figure 2.1: Reaction graph for an example system on nine species and seven complexes.

with complex set

$$\mathcal{C} = \{\{A, B\}, \{C, D\}, \{E, F\}, \{A, E\}, \{C, G\}, \{B, F\}, \{H, I\}\}$$

and reactions

$$\begin{aligned}\mathcal{R} = \{&A + B \leftrightarrow C + D,\\ &C + D \leftrightarrow E + F,\\ &E + F \leftrightarrow A + B,\\ &A + E \leftrightarrow C + G,\\ &B + F \leftrightarrow H + I\}.\end{aligned}$$

For many of the results in this thesis, we will restrict attention to *quadratic* reaction networks – that is, networks where each complex is composed of exactly two species, rather than an arbitrary number – and networks which are *reversible*[1], meaning each edge is bidirectional. The network in Figure 2.1, for example, is both quadratic and reversible. While many results discussed in this thesis will be seen to hold in full generality, we will primarily restrict attention to the reversible, quadratic case, following the work on quadratic dynamical systems in [77]. Most of the fundamental open questions and complex behaviors of reaction network theory already appear in this setting, as well as its utility for sampling applications, so it will provide a useful entry point for thinking about reaction networks.

2.2.2 Mass Action Kinetics

The reactions between complexes in the reaction graph act on a system by changing the quantity of each species in the system. The state of a reaction network system at any given

[1]Note that the names *reversible* and *weakly reversible* originate from chemical reaction network theory, and are not exactly analogous to the use of "reversible" in the theory of Markov Chains. Because these terms are deeply ingrained in the reaction network literature, we use them here as defined above, and will avoid their use in the Markov Chain context.

Reversible	For every edge $(y, y') \in \mathcal{R}$, the reverse edge $(y', y) \in \mathcal{R}$.
Symmetric	For every edge $(y, y') \in \mathcal{R}$, the reverse edge $(y', y) \in \mathcal{R}$, and the corresponding rate constants $\beta_{y,y'} = \beta_{y',y}$.
Weakly Reversible	Every edge $(y, y') \in \mathcal{R}$ belongs to a strongly connected component.
Quadratic	For every $y \in \mathcal{C}$, $\|y\| = 2$.
Mass-preserving	For some $k \in \mathbb{N}$, every $y \in \mathcal{C}$ has $\|y\| = k$.

Table 2.1: Definitions for some common subclasses of reaction graphs

time t is fully described by the following vector $x(t)$:

Definition 2.2 (State)**.** A *state* is a vector $x \in \mathbb{R}^N_{\geq 0}$ representing the quantity of each species in $\mathcal{S}$. For a state x, we also call x_I the *concentration* of species I at state x.

The definition of a reaction network is not complete without a choice of dynamics $\dot{x} = f(x)$ describing the rate at which such changes occur. But what are the most useful dynamics to impose on such a system? A natural choice, from the perspectives of both chemistry and abstract probabilistic models, is that of *mass-action kinetics*. Under mass-action kinetics, the rate of each reaction is proportional to the product of concentrations of each of its reactants; in a network that preserves the total concentration, or *mass* of the species, this can be thought of as proportional to the probability of selecting the necessary set of reactants when the species to include are chosen independently at random from the current probability distribution $x \in \mathbb{R}^{|\mathcal{S}|}_{\geq 0}$[2]. Furthermore, the rate constant of each reaction modifies the overall reaction rate by a linear factor. Formally, and using the notation x^y to represent $\prod_{I \in y} x_I$ with $y \in \mathcal{C}$, $x \in \mathbb{R}^N$:

Definition 2.3 (Mass-action kinetics)**.** A reaction network has *mass-action kinetics* if, for each reaction $(y, y') \in \mathcal{R}$, there exists a constant $\beta_{y,y'} > 0$ such that the rate of the reaction $y \to y'$ at state x is $\beta_{y,y'} x^y$.

A reaction network with this property is also known as a *mass-action system*, and its dynamics are given by

$$\dot{x}(t) = \sum_{(y,y') \in \mathcal{R}} \beta_{y,y'} x(t)^y (y' - y) \tag{2.1}$$

In this thesis, unless otherwise stated, all reaction networks will be assumed to use mass-action kinetics. See Figure 2.2 for a breakdown of the terms in this equation.

[2]Note the role of mass preservation in this analogy: a network being mass-preserving is equivalent to the constraint that $\sum_{i \in \mathcal{S}} x_i$ is invariant under all reactions, so with the appropriate normalization to $\sum_{i \in \mathcal{S}} x_i = 1$, a state x can be thought of as a probability distribution over $\mathcal{S}$.

Each term of the summation corresponds to a single unidirectional reaction (y, y'), and the contribution of the term depends nonlinearly on the product of the current concentrations of all reactants in complex y, scaled by the rate constant $\beta_{y,y'}$. A change of precisely this magnitude is applied to the corresponding reaction vector $(y' - y)$, which reduces the concentration of each reactant and increases the concentration of each product by the same amount.

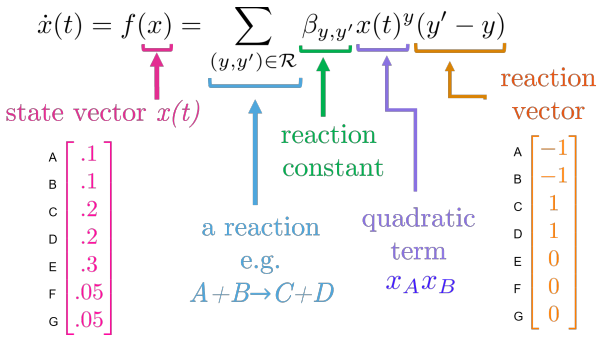


Figure 2.2: Annotated mass action dynamics for an example system on species set $\mathcal{S} = \{A, B, C, D, E, F, G\}$, at state $x(t) = [.1, .1, .2, .2, .3, .05, .05]$, with the reaction $A + B \to C + D$.

In particular, note that for reactions with two species per complex, the corresponding $x(t)^y$ term is quadratic, whereas in a system with only one species in each complex, this term would become linear, and the overall dynamics would then reduce to that of a continuous-time Markov chain. In this sense, quadratic and higher-order reaction networks can be thought of as a nonlinear generalization of a Markov chain.

2.2.3 Autonomous and Non-Autonomous Dynamics

When not otherwise specified, all systems discussed in this thesis are *autonomous*, meaning that the edge weights $\beta_{y,y'}$ are constant. In some special cases, we may wish to generalize to the more general class of *non-autonomous* kinetics, in which the rate coefficient varies as a function of time. We will restrict this study to only those networks where each rate coefficient is bounded above and away from zero.

Definition 2.4 (Non-autonomous, bounded-rate mass action system)**.** A reaction network has *non-autonomous*, *bounded-rate* mass-action kinetics if, for each reaction $(y, y') \in R$, there is some function $\beta_{y,y'}(t)$ not dependent on x, and some $\epsilon > 0$ such that $\epsilon < \beta_{y,y'}(t) < \frac{1}{\epsilon}$ for all $t \geq 0$, and the rate of the reaction (y, y') at state $x(t)$ is $\beta_{y,y'}(t)$.

Then the dynamics of the chemical reaction system are given by Equation (2.1) with $\beta_{y,y'}$ replaced by $\beta_{y,y'}(t)$. Here, *bounded-rate* refers to the fact that $\beta_{y,y'}(t)$ is bounded above and below; *non-autonomous* refers to the fact that $\beta_{y,y'}$ is a function of t. Unless otherwise

specified, we will assume that all mass action systems are autonomous. If a theorem is intended to hold even when a reaction network is non-autonomous, it will be explicitly noted.

Where do non-autonomous mass-action systems arise? For our purposes, they may appear as a result of a technique known as *projection*, which is used to remove some species from a mass action system by updating the rate constants to rate functions $\beta_{y,y'}(t)$ for the remaining reactions to mimic the behavior of the initial system. This technique makes it possible to replace an existing autonomous mass action system with a new non-autonomous system on a strictly smaller species set, possibly creating advantageous graph structures in the process, without changing the overall behavior of the system; see Section 2.4.5 for more on this point.

2.2.4 The Invariant Class

A reaction network naturally gives rise to a linear subspace describing all the possible linear combinations of the reaction vectors; we call this the *stoichiometric subspace*. This subspace describes all the possible changes that may befall the initial state, regardless of the choice of dynamics for the reaction network:

Definition 2.5 (Stoichiometric Subspace)**.** The *stoichiometric subspace* of a reaction network is defined as

$$\mathcal{H} := \operatorname{span}\{y' - y \mid (y, y') \in \mathcal{R}\}$$

Definition 2.6 (Stoichiometric Compatibility Class)**.** Let $\mathcal{H}$ be the stoichiometric subspace of a reaction network. For an initial condition x_0, the *stoichiometric compatibility class* is the set

$$\mathcal{SC}(x_0) := (x_0 + \mathcal{H}) \cap \mathbb{R}^N_{\geq 0}$$

and the *positive stoichiometric compatibility class* is the set

$$\mathcal{SC}^+(x_0) := (x_0 + \mathcal{H}) \cap \mathbb{R}^N_{>0}$$

It is straightforward to see that a trajectory starting at x_0 must always stay in $\mathcal{SC}(x_0)$. Moreover, any vector orthogonal to the stoichiometric subspace does not change no matter which reactions occur, since $x \cdot (y' - y) = 0$ for all $(y, y') \in \mathcal{R}$; another way to say this is that all such vectors lie in the space of *linear invariants*.

Definition 2.7 (Linear Invariant)**.** A *linear invariant* of a chemical reaction network is a linear function $q(x) = \sum_{I \in \mathcal{S}} x_I \alpha_I$ for some $\alpha \in \mathbb{R}^N$, such that for any trajectory $x(t)$, $q(x(t)) = q(x(t'))$ for all $t, t' > 0$.

Definition 2.8 (Invariant Class)**.** The *invariant class* $\mathcal{IC}(x_0)$ is the set of points $x \in \mathbb{R}^N_{\geq 0}$ such that for all linear invariant functions q, $q(x) = q(x_0)$. The *positive invariant class* $\mathcal{IC}^+(x_0)$ is defined analogously for $x \in \mathbb{R}^N_{>0}$.

While $\mathcal{IC} \subset \mathcal{SC}$, depending on the dynamics of the reaction network, there may be additional linear invariants arising from the dynamics of a system which cannot be inferred solely from the stoichiometric subspace. In this case, there are increased restrictions on what states can be reached along any given trajectory, and $\mathcal{IC} \subsetneq \mathcal{SC}$.

Example 2.1. *We now see an example where $\mathcal{IC} \neq \mathcal{SC}$.*

$$A \xrightarrow{1} C$$
$$A \xrightarrow{1} B$$
$$C + B \xrightarrow{1} 2A$$

The stoichiometric subspace $\mathcal{SC} = \text{span}\{(B - A), (C - A), 2A - C - B\}$ which clearly has dimension 2. However, because the first two reactions occur at an identical rate at all times, they act identically to the single reaction $2A \xrightarrow{1} B + C$ – just the reverse of the third reaction. Thus the invariant class $\mathcal{IC}$ is spanned by the single vector $\{2A - C - B\}$; so $\mathcal{IC}$ has dimension 1, and $\mathcal{IC} \subset \mathcal{SC}$.

However, upon further investigation, this situation is often impossible. In particular, (1) $\mathcal{IC} \subsetneq \mathcal{SC}$ cannot be achieved robustly, in the sense that there always exists a small perturbation of reaction constants after which point $\mathcal{IC}$ and $\mathcal{SC}$ coincide (see [44]); and (2) as we will see in Lemma 2.2, $\mathcal{SC} = \mathcal{IC}$ for all reversible reaction networks (which are the type of networks of most interest in this thesis).

Lemma 2.2. *For a reversible reaction network, the linear invariants are exactly given by*

$$\left\{ \sum_{I \in \mathcal{S}} x_I \alpha_I \mid (y - y') \cdot \alpha = 0 \ \forall \ (y, y') \in \mathcal{R} \right\}$$

and accordingly, $\mathcal{IC} = \mathcal{SC}$.

Proof. Let q be a linear invariant, such that $q = \sum_{I \in \mathcal{S}} x_I \alpha_I$. Then for all x we must have

$$\dot{q} = \frac{1}{2} \sum_{\substack{(y,y') \in \mathcal{R} \text{ s.t.} \\ \sum_{I \in y'} \alpha_I \geq \sum_{J \in y} \alpha_J}} \left(\sum_{I \in y'} \alpha_I - \sum_{J \in y} \alpha_J \right) \left(x^{y'} - \frac{\beta_{y,y'}}{\beta_{y',y}} x^y \right) \beta_{y',y} = 0. \tag{2.2}$$

Our goal is to show that for some choice of x, $\dot{q}(x) > 0$ unless $(y - y') \cdot \alpha = 0$. Let k be a constant such that $k > \ln\left(\frac{\beta_{y,y'}}{\beta_{y',y}}\right)$ for all $(y, y') \in \mathcal{R}$. Let

$$m := \min_{(y,y') \in \mathcal{R}} \left| \sum_{I \in y'} \alpha_I - \sum_{J \in y} \alpha_J \right|$$

and $n > \min(1, \frac{k}{m})$. Now for any $(y, y') \in \mathcal{R}$ s.t. $\sum_{I \in y'} \alpha_I \geq \sum_{J \in y} \alpha_J$,

$$n \sum_{I \in y'} \alpha_I \geq mn + n \sum_{J \in y} \alpha_J > k + n \sum_{J \in y} \alpha_J.$$

Letting $\ln x_I := n\alpha_I$, we have

$$\sum_{I \in y'} \ln x_I = n \sum_{I \in y'} \alpha_I > k + n \sum_{J \in y} \alpha_J = k + \sum_{J \in y} x_J > \ln \frac{\beta_{y,y'}}{\beta_{y',y}} + \sum_{J \in y} \ln x_J$$

and hence $x^{y'} > \beta_{y,y'} \beta_{y',y} x^y$.

This implies by Equation (2.2) that $\dot{q}(x) > 0$ unless $\sum_{I \in y'} \alpha_I = \sum_{J \in y} \alpha_J$ for all $(y, y') \in \mathcal{R}$ (in which case $\dot{q}(x) = 0$). Thus, the linear invariants are exactly those $q = \sum_{I \in \mathcal{S}} x_I \alpha_I$ such that $(y - y') \cdot \alpha = 0$ for all $(y, y') \in \mathcal{R}$. □

This theorem implies that the set of invariants is independent of the reaction constants for all reversible reaction networks. Although the proof method provided here is different, Lemma 2.2 also follows as a consequence of a more general theorem proved in [45]:

Theorem 2.3 (Feinberg and Horn [45])**.** *For any mass action system in which the underlying reaction network is weakly reversible, $\mathcal{IC} = \mathcal{SC}$.*

That is, the class of points reachable from x_0 by applying reaction vectors is exactly the set of points whose invariants match those of x_0: All states which are reachable from one another are also invariant compatible.

The original statement of the above theorem, and the corresponding discussion by its author in [44], refer to the "kinetic subspace" rather than $\mathcal{IC}$, but upon inspection, the two describe identical subspaces[3].

Corollary 2.4. *For any weakly reversible mass action system, the linear invariants are exactly given by $\left\{\sum_{I \in \mathcal{S}} x_I \alpha_I \mid (y - y') \cdot \alpha = 0 \ \forall \ (y, y') \in \mathcal{R}\right\}$*

In fact, the invariant class $\mathcal{IC}$ of a weakly reversible reaction network defines a polyhedron, known as the *invariant polytope*; see, for example, [86]. This polytope will be of central importance in Chapter 3.

[3]Feinberg [44] defines the kinetic subspace K as the smallest linear subspace of $\mathbb{R}^N$ containing $\mathrm{Im}(\dot{x})$. This is equivalent to our definition of $\mathcal{IC}$, since we take the largest linear subspace not in $\mathrm{Im}(\dot{x})$ (i.e., the largest subpace of invariant vectors), and define $\mathcal{IC}$ to be the space orthogonal to that.

2.2.5 Trajectories, Limits, and Equilibria

Like Markov Chains and other dynamical systems, one of the primary questions to ask about a reaction network is how $x(t)$ changes over time. Does it approach a steady state over sufficiently long time horizons? If so, which state does it approach? If not, how does it behave instead: Does it exhibit multistationarity or cyclic behavior? In either case, are these dynamics the same for all trajectories within the same invariant compatibility class (asymptotic stability), or are they more sensitive to the choice of initial condition? Answering these questions is a crucial prerequisite for building systems based on reaction network dynamics. To describe the question more formally, we use the following definition:

Definition 2.9 (Trajectory)**.** A *trajectory* $x(t)$ represents the solution of the initial value problem $\dot{x} = f(x)$, $x_0 = x(t_0)$ at any given time $t \in [t_0, \infty)$.

One can think of the trajectory of a reaction network as the sequence of states reached over time, describing the evolution of the system's state as reactions occur and the concentrations of species fluctuate. This trajectory describes a path through the *state space* $\mathbb{R}^N_{\geq 0}$, parameterized by time, noting that all states along a given trajectory exist somewhere within this nonnegative orthant. It is often meaningful to distinguish between the *interior* and the *boundary* of this space; a state $x \in \mathbb{R}^N_{\geq 0}$ is in the interior if it has full support, and on the boundary otherwise.

With this in mind, one of the critical questions to answer will be: does a particular trajectory $x(t)$, with an initial condition on the interior of the state space, always remain in the interior? Could it reach the boundary at some finite time, and even if not, does it have any limit points on the boundary? We will see in Section 2.3 that for "well-behaved" networks which converge to a single equilibrium distribution, that point will fall in the interior of the state space; while other, more "pathological" behaviors (e.g., oscillation or multi-stability) can only arise on a trajectory which reaches or approaches arbitrarily close to the boundary. For now, we establish a few definitions to talk about these possible behaviors.

Definition 2.10 (ω-limit point)**.** Given a dynamical system in $\mathbb{R}^n$ with initial condition x_0, with trajectory $x(t)$, an *ω-limit point* is any point $z \in \mathbb{R}^n$ such that $x(t_n) \to z$ for some sequence of times $t_n \to \infty$. Similarly, the *ω-limit set* of a dynamical system for a given initial condition x_0 is the set consisting of all ω-limit points for trajectory $\phi(t, x_0)$.

Definition 2.11 (Stationary point)**.** A point π is *stationary*, or equivalently, an *equilibrium* if $\dot{\pi} = 0$. That is, a trajectory passing through $x(t) = \pi$ remains at that point at all subsequent times, so that $x(t') = \pi$ for all $t' \geq t$.

A single mass-action system without a specified initial condition may have multiple equilibria within an invariant class, and a single trajectory governed by a mass-action system may have multiple ω-limit points. In fact, for a given reaction network and initial conditions, there may well be stationary points which are not ω-limit points of that particular trajectory (indeed, this will be the case for all but the most trivial reaction networks). For instance, see Example 2.5 below. Furthermore, there might be ω-limit points of a trajectory which are

not stationary. This behavior can be observed in the cyclic limiting behavior exhibited by Example 2.6.

Example 2.5. *The following system, studied in [59, 44], with rate constants* $\beta_{3A,A+2B} = \beta_{3B,2A+B} = 0.1$ *and* $\beta_{2A+B,3A} = \beta_{A+2B,3B} = 1$*, has three stationary points in each invariant class, and converges to a single one of them in the limit, determined by the initial conditions. In all cases, two of these equilibria are asymptotically stable, and one is unstable; and all are in the interior of the state space.*

$$\begin{array}{ccc} 3\mathrm{A} & \longrightarrow & \mathrm{A}{+}2\mathrm{B} \\ \uparrow & & \downarrow \\ 2\mathrm{A}{+}\mathrm{B} & \longleftarrow & 3\mathrm{B} \end{array}$$

Example 2.6. *The following system, known as the Lotka–Volterra predator–prey model [68], has a single, interior equilibrium point at* $[x_A, x_B] = \left[\frac{\beta_{B,0}}{\beta_{A+B,2B}}, \frac{\beta_{A,2A}}{\beta_{A+B,2B}}\right]$*. However, for any initial condition other than the equilibrium point itself, the trajectory forms a cycle, which amounts to a continuum of* ω*-limit points. Note that* $\dot{x} \neq 0$ *at all of these points, so they are not equilibria.*

$$\begin{aligned} \mathrm{A} &\longrightarrow 2\mathrm{A} \\ \mathrm{A}{+}\mathrm{B} &\longrightarrow 2\mathrm{B} \\ \mathrm{B} &\longrightarrow 0 \end{aligned}$$

We also recall the theorem, as stated in [67], which applies to any positive semi-orbit of a continuous, autonomous, bounded dynamical system $\dot{x} = f(x)$, and apply it to reaction networks under autonomous mass-action kinetics:

Theorem 2.7 (Follows from Lebovitz [67])**.** *A set of states* S *is called* permanent *if a trajectory through a point* $x(t) \in S$ *remains in* S *for all* $t' \geq t$*. The* ω*-limit set for any bounded, autonomous mass-action system is connected, closed, and permanent.*

We further reduce the space of possible boundary ω-limit points using a condition called *stationary support*[4]:

[4]The concept of stationary support has been previously identified in [10] and [9] using the term "semi-locking set". A point has stationary support exactly when the complement of its support is a semi-locking set.

Definition 2.12 (Stationary Support)**.** A point π in state space Δ_N has *stationary support* if for all complexes $y, y' \in \mathcal{C}$ with $\beta_{y,y'} > 0$,

$$\prod_{I \in y} \pi_I = 0 \Leftrightarrow \prod_{J \in y'} \pi_J = 0$$

Intuitively, stationary supported boundary points are those x where, if the reaction $y \to y'$ cannot continue because $x_I = 0$ for some $I \in y$, we can guarantee that no reverse reaction $y' \to y$ occurs to replenish x_I. This is formalized in the following theorem, proved by both Angeli, De Leenheer, and Sontag [10] and Anderson [9]:

Theorem 2.8 (Angeli, De Leenheer, and Sontag [10] and Anderson [9])**.** *For any autonomous mass-action system, all ω-limit points have stationary support.*

2.2.6 Balancing Properties

In addition to simply finding equilibria at the species level (ie., $\dot{x} = 0$), are there additional criteria that produce "balancing" behavior at the reaction or complex level? To answer this question, this section introduces two classes of reaction networks first investigated by [59, 43, 58], which are defined by their *kinetic*, rather than *structural* properties. Structural properties, such as a reaction network being reversible, symmetric, or quadratic, are determined solely by the structure of the reaction graph, independent of the rate constants. In contrast, kinetic properties of a mass action system require the rate constants to satisfy particular relationships, and these properties are not necessarily preserved for different rate constants on the same reaction graph.

Two types of balancing – complex balance and detailed balance – will be our primary objects of study within the landscape of kinetic properties. These properties are derived from two possible properties of positive equilibrium states. Under *detailed balance*, the rates of forward and backward reactions are equal at equilibrium, in analog to the Markov Chain definition of reversibility. However, a more apt analog is found in *complex balance*, in which a similar balancing property holds at the *complex level* rather than the reaction level: the total flow of mass in and out of each complex is equal. Note that the total flow of mass in and out of each *species* is equal for equilibria in any type of reaction network; detailed and complex balancing can be viewed as two generalizations of this property.

It is worth noting also that detailed balanced mass action systems are a subset of reversible networks; and similarly, complex balanced systems are a subset of weakly reversible networks. This correspondence will take on extra interest in our study of persistence and the Global Attractor Conjecture in a later section.

We now proceed with the formal definitions:

Definition 2.13 (Detailed balanced state)**.** A state x of a mass action system is *detailed balanced* if $\beta_{y,y'} x^y = \beta_{y',y} x^{y'}$ for all $(y, y') \in \mathcal{R}$.

It is immediate from applying this definition to the dynamics of mass-action kinetics that if detailed balancing obtains at x, then x is an equilibrium.

Detailed balanced reaction networks are exactly those reversible reaction networks for which, at some positive equilibrium, the rates of forward and reverse reactions are identical.

Definition 2.14 (Detailed balanced network)**.** A reversible mass action system is *detailed balanced* if there exists a positive equilibrium at which detailed balancing obtains.

In general, detailed balance is a condition that depends critically on the specific reaction rates; but if the set of reaction vectors is linearly independent, it is always possible to solve for a detailed balanced, positive equilibrium:

Theorem 2.9 (Feinberg [44])**.** *Given a mass action system, if the set of reaction vectors* $\{y_i - y_j \mid (y_i, y_j) \in \mathcal{R}, i < j\}$ *is linearly independent, detailed balancing obtains at every equilibrium (including any boundary equilibria).*

In fact, detailed balanced networks have this property at *all* positive equilibria, not just one; this is a fundamental theorem of detailed balance:

Theorem 2.10 (Horn and Jackson [59])**.** *If detailed balancing obtains at one positive equilibrium, then it obtains at all positive equilibria of the reaction network.*

Our other balancing condition, *complex balance*, describes a much larger class of reaction networks, of which detailed balanced networks are a subset. Complex balanced networks are a subset of weakly reversible reaction networks, not necessarily reversible ones; and in contrast to detailed balancing, complex balance requires only that the *total* flows in and out of a complex are equal, not the necessarily along each individual reaction.

Definition 2.15 (Complex balanced state)**.** A state x of a mass action system is *complex balanced* if for all complexes y,

$$\sum_{y' \in \mathcal{C}} \beta_{y',y} x^{y'} = \sum_{y' \in \mathcal{C}} \beta_{y,y'} x^{y}$$

Definition 2.16 (Complex balanced network)**.** A weakly reversible mass action system is complex balanced if there exists a positive equilibrium at which complex balance obtains.

As with detailed balance, complex balance at a single equilibrium is sufficient to determine that all equilibria are complex balancing.

Theorem 2.11 (Horn and Jackson [59])**.** *If complex balancing obtains at one positive equilibrium, then it obtains at all positive equilibria of the reaction network.*

Finally, positive complex balanced equilibria are unique within a given invariant compatibility class:

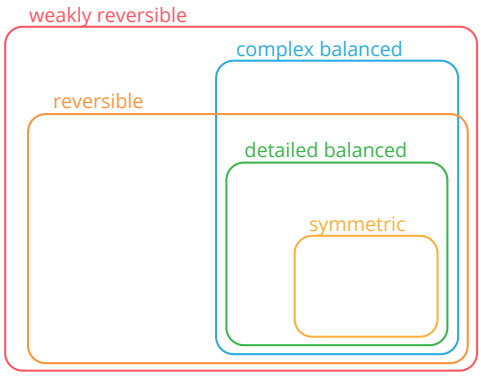


Figure 2.3: Relationships between reversibility, balancing, and symmetry under mass action kinetics

Theorem 2.12 (Birch's theorem, [59])**.** *For a complex balanced mass action system, there is exactly one positive equilibrium in each invariant class. Furthermore, this equilibrium is asymptotically stable and there is no nontrivial cyclic trajectory residing entirely in the interior of the state space.*

Note that this theorem does not preclude the existence of additional equilibria if they reside on the boundary of the state space; and similarly, there may be cyclic trajectories or unstable equilibria which intersect the boundary. The question of whether the unique interior complex balance point is in fact an asymptotically stable global attractor is the subject of the Global Attractor Conjecture in Section 2.3.

Sufficient conditions for balancing

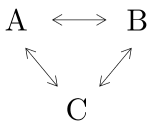
It is immediate from the definitions that if a mass action system is symmetric, it is detailed balanced; and if a system is detailed balanced, it is complex balanced; see Figure 2.3 for a summary of the relationships between some of the reaction network properties defined so far.

Whether a given reaction network admits a complex balanced solution can in some cases be determined by an invariant called the *deficiency* of the network, which was first defined and studied by Feinberg [43], Horn [58], and Horn and Jackson [59]. The deficiency of a reaction network is an integer $\delta := |\mathcal{C}| - l - \sigma$, where $|\mathcal{C}|$ is the cardinality of set of complexes, l represents the number of weakly connected components in the reaction graph, and σ is the dimension of the stoichiometric subspace $\mathcal{H}$. A deficiency $\delta = 0$ corresponds to a graph with reaction vectors that are maximally independent given the number of connected components in the graph, and is a sufficient condition for that network to admit a complex balanced equilibrium for any choice of rate constants [44].

Craciun et al. [33] further showed that the space of rate constants which support complex balancing in a given reaction network is a toric variety and characterized its combinatorial

structure, leading to the alternate name *toric dynamical system* to refer to complex balanced reaction networks.

Example 2.13 (From Feinberg [44])**.** *Consider the following reversible network on complex set* $\mathcal{C} = \{A, B, C\}$*:*



This example has $|\mathcal{C}| = 3$*,* $l = 1$*, and* $\sigma = 2$*, and so it has deficiency* $\delta = 0$ *and is complex balanced for any choice of nonzero rate constants on each edge. On the other hand, Example 2.13 is detailed balanced exactly when* $\beta_{A,B}\beta_{B,C}\beta_{C,A} = \beta_{A,C}\beta_{C,B}\beta_{B,A}$*. Any other class of rate constants does not admit any equilibrium state satisfying the detailed balancing condition.*

A note on detailed balanced equilibria

In contrast to the traditional view in reaction network theory – which tends to focus on the modeling and analysis of small, biological or chemical systems – this thesis takes the perspective that we are interested in *large* networks, with arbitrarily many species representing combinatorial objects, and equipped with a graph structure and dynamics that naturally arises from the objects that the species represent. Furthermore, in order to use these networks to sample from a distribution over a large species set, it is of high importance to prove that a given network converges to a known class of distributions.

The class of reaction networks which best fit this description are detailed balanced networks. The following theorem characterizes the rate constants required to give rise to detailed balance:

Theorem 2.14 (Feinberg [44])**.** *A reversible mass action system is detailed balanced if and only if the rate constants satisfy*

$$\prod_{(y,y')} (\beta_{y,y'})^{\alpha^{\theta}_{y,y'}} = \prod_{(y,y')} (\beta_{y',y})^{\alpha^{\theta}_{y,y'}}$$

for all $\theta \in 1, 2, \ldots, |\mathcal{C}| - \sigma$ *where* $\{\alpha^1, \ldots, \alpha^{|\mathcal{C}|-\sigma}\}$ *are a set of linearly independent solutions to* $\sum_{(y,y')} \alpha_{y,y'}(y - y') = 0$*.*

For applications where convergence to a known family of distributions is required, we typically assume the ability to set the rate constants on each edge in $\mathcal{R}$ to ensure that the desired detailed balanced equilibrium exists.

2.3 The Global Attractor Conjecture

A major motivation for the study of complex balanced and (weakly) reversible mass action systems comes in the form of a major unsolved problem in reaction network theory. In its simplest form, the problem asks: When is it possible for a mass action system to approach a state where the concentration of one or more species is zero? Even in the case of complex balancing systems, with their well-behaved reaction rates, this can still be a mystery – while there is a unique *positive* equilibrium for a given initial condition, this does not guarantee that the trajectory always approaches that equilibrium, nor does it rule out the possibility of boundary limit points. The Global Attractor Conjecture, however, theorizes exactly that: In the case of complex balancing systems, it posits that convergence to the unique interior equilibrium is the *only* possible limiting behavior.

2.3.1 Main Conjectures

Definition 2.17 (Global attractor condition)**.** A mass action system satisfies the *global attractor* condition if there exists a unique equilibrium in each positive invariant class, which is approached in the limit as $t \to \infty$ by any trajectory originating in that positive invariant compatibility class.

Conjecture 2.15 (Global Attractor Conjecture, Feinberg and Horn [46])**.** *Every complex balanced mass action system satisfies the global attractor condition.*

Definition 2.18 (Persistence condition)**.** A mass action system is *persistent* if, given a start state in $\mathbb{R}^N_{>0}$, all ω-limit points of the trajectory are in $\mathbb{R}^N_{>0}$.

Conjecture 2.16 (Persistence Conjecture, Feinberg [42])**.** *Every weakly reversible mass action system is persistent.*

In fact, Conjecture 2.16 is strictly stronger than Conjecture 2.15: If a complex balanced reaction network is persistent, it automatically satisfies the global attractor property as well. This fact is non-obvious, but follows from the following theorem which clarifies the relationship between the persistence and global attractor properties in the complex balanced setting:

Theorem 2.17 (Siegel and MacLean [87])**.** *Any complex balanced reaction network with a given initial condition has as its ω-limit set either the unique, interior complex balanced equilibrium point in its invariant class, or some set of boundary complex balanced equilibrium points.*

In other words, the only possible counterexamples to the Global Attractor Conjecture are boundary equilibria, and – for complex balanced reaction networks – these are mutually exclusive with convergence to the unique interior equilibrium for a given initial condition.

2.3.2 History and Recent Progress

Both the Global Attractor and Persistence Conjectures remain open, despite many attempts and partial proofs[5]. Some success has been found in special cases – of particular interest are structural conditions relating to the number of weakly connected components [7], as well as more technical conditions such as *concordance* [85] and *strong endotacticity* [50], which are found to be sufficient for the Global Attractor and Persistence Conjectures to hold. These will be discussed in Section 2.4 alongside a new structural condition which generalizes some of these results.

2.4 Structural Conditions for Convergence

Because reaction constants are rarely known precisely, and properties based on these constants may not be robust to small perturbations, theorems that use only structural properties are highly valuable; and much recent work has targeted expanding the list of structural conditions which are known to imply persistence.

2.4.1 Tier Methods

The following section highlights a method for analyzing mass action systems by creating an ordering of complexes for a given convergent subsequence of $x(t)$, based on how quickly each species' concentration approaches zero (if at all). The equivalence classes in this ordering will be known as *tiers*.

This technique, first introduced by Anderson [7], is sufficient to prove persistence for mass action systems with particular connectivity properties, notably those which have a *single linkage class*. Work original to this thesis, as well as work by Anderson et al. [6], further extends the technique to the much larger class of systems which possess an edge leaving the lowest tier, and in doing so, also simplifies an earlier proof of Gopalkrishnan, Miller, and Shiu [50] for *strongly endotatctic* systems. In particular, we show here for the first time that a significant simplification is possible for quadratic reaction networks, and the algebraic tool used to accomplish this leads to a generalization of the prior results about tiers to the broader family of systems with *realizable orderings* (as defined in Section 2.4.2).

The following definitions are adapted from Anderson et al. [6].

Definition 2.19 (Proper tier sequence)**.** A sequence $(x_n)_{n=0}^{\infty}$ of positive vectors in $\mathbb{R}_{>0}^{|\mathcal{S}|}$ is a *tier sequence* if $\lim_{n\to\infty} ||\ln(x_n)||_\infty = \infty$ and for all pairs of complexes $y, y' \in \mathcal{S}\times\mathcal{S}$, $\lim_{n\to\infty} x_n^{y'-y}$ exists. The tier sequence is a *proper tier sequence* if additionally, for all $n, m \in \mathbb{Z}_{\geq 0}$, $x_n - x_m$ is in the stoichiometric subspace.

[5]A proof of the Global Attractor Conjecture was initially claimed by Horn and Jackson [59], but was found to be incomplete; the claim was withdrawn in their later paper [46] which first introduced the conjecture as an open question. Notably, Craciun [32] also recently claimed a proof of the Global Attractor Conjecture, which was later determined to be incomplete [31, 44].

Tiers, then, define a partition of $\mathcal{C}$ into subsets with different asymptotic behavior along a given tier sequence:

Definition 2.20 (Tiers)**.** Given a tier sequence $(x_n)_{n=0}^{\infty}$, define the *tier ordering* $\succeq_{(x_n)}$ such that $y \succ_{(x_n)} y'$ iff $\lim_{n\to\infty} x_n^{y-y'} = 0$ and $y \sim_{(x_n)} y'$ iff $\lim_{n\to\infty} x_n^{y-y'} \in (0, +\infty)$. The equivalence classes over the tier ordering form a partition of $\mathcal{C}$: we label the equivalence classes $k+1$ partitions $T_0, \ldots, T_k$, such that for any $y \in T_i, y' \in T_j$, $y \succeq_{(x_n)} y'$ iff $i \geq j$.

Note that the tier numbering $T_0, \ldots, T_k$ matches [6], but the tier ordering $\succeq_{(x_n)}$ is exactly reversed from that paper. The convention in Definition 2.20 will be followed in all places in this document; thus the tier ordering can be understood as describing a complex's speed of approach to zero concentration, with a higher position in the order corresponding to a faster approach to zero (and the tier with concentrations approaching positive constants, if any, is labeled T_0).

Anderson [7] also presents a more expansive definition of tiers, which allows for arbitrary finite sets of vectors in $\mathbb{R}^N$ in place of the complex set $\mathcal{C}$:

Definition 2.21 (Partition along a subsequence)**.** Let $\mathcal{C}$ denote any finite set of vectors in $\mathbb{R}^N$. Let $x_n \in \mathbb{R}^N_{>0}$ denote a sequence of points in the strictly positive orthant. We say that $\mathcal{C}$ is *partitioned along the sequence* $\{x_n\}$ if there exists a partition $T_0, \cdots, T_P$ of $\mathcal{C}$, called *tiers*, and a constant $C > 1$, such that

1. If $y_j, y_k \in T_i$ for some $i \in \{0, \ldots, P\}$, then for all n,

$$\frac{1}{C} x_n^{y_j} \leq x_n^{y_k} \leq C x_n^{y_j}, \text{ and}$$

2. If $y_j \in T_i$ and $y_k \in T_{i+m}$ for some $m \in \{0, \ldots, P-i\}$, then

$$\frac{x_n^{y_j}}{x_n^{y_k}} \to \infty \text{ as } n \to \infty$$

Lemma 2.18 (Anderson [7])**.** *Let $\mathcal{C}$ denote a finite set of vectors in $\mathbb{R}^N$. Let x_n be a sequence of points in $\mathbb{R}^N_{>0}$. Then, there exists a subsequence of $\{x_n\}$ along which $\mathcal{C}$ is partitioned.*

In particular, defining $\mathcal{C}^+$ to contain indicator vectors for not only the complexes of the system, but also all possible pairs of species, Lemma 2.18 guarantees that there exists a subsequence $x(q)$ of $x(t)$ along which $\mathcal{C}^+$ is partitioned into tiers. This ordering defines a total order over pairs of species in $\mathcal{S} \times \mathcal{S}$.

2.4.2 Realizable Orderings for Quadratic Systems

Anderson [8] proved that tiers along a subsequence approaching a boundary limit point admit a vector that indicates whether two complexes are in the same tier (in that paper's terminology, a "conservation relation"):

Theorem 2.19 (Anderson [8])**.** *Given any proper tier sequence of a reaction network, with boundary limit point z, then there must exist a vector $w \in \mathbb{R}_{\geq 0}$, such that for y, y' in the same tier, $w \cdot (y - y') = 0$; and $w_I > 0$ if and only if $I \notin supp(z)$.*

This fact has a useful consequence: $w \cdot x$ is invariant across reactions that exist within a single tier. This leads to the following observation:

Corollary 2.20. *If a trajectory $x(t)$ of the reaction network with positive initial condition approaches a boundary limit point z along a proper tier sequence, there exists an edge between two distinct tiers under the corresponding tier ordering.*

Proof. Suppose there exists no such edge. Then for all edges $y \leftrightarrow y'$, $w \cdot (y - y') = 0$. Because every reaction is invariant under w, we can write $w \cdot \frac{dx}{dt} = 0$. Observe also that $w \cdot z = 0$, and that $w \cdot x(t_0) > 0$ (by the assumption that the initial conditions at time t_0 satisfy $x_i(t_0) > 0$ for every i). This is a contradiction, so some edge between distinct tiers exists. □

In the quadratic case, we are able to extend Theorem 2.19 to something significantly stronger:

Theorem 2.21. *Given any proper tier sequence $(x_n)_{n=1}^{\infty}$ of a quadratic reaction network, with boundary limit point z, there exists $w \in \mathbb{R}_{\geq 0}$ such that for all $y, y' \in \mathcal{C}$,*

$$w \cdot y \geq w \cdot y' \Leftrightarrow y \succeq_{(x_n)} y'$$

(where $\succeq_{(x_n)}$ is the ordering over tiers along $(x_n)_{n=1}^{\infty}$); and $w_I > 0$ if and only if $I \notin supp(z)$.

That is: It is possible to assign each species $a \in \mathcal{S}$ a real number w_a, such that the tier ordering of two complexes (a, b) and $(c, d) \in \mathcal{C}$ can be determined simply by comparing $w_a + w_b$ to $w_c + w_d$. This amounts to *measuring* the convergence speed of each species along the trajectory rather than simply ordering the complexes.

To accomplish this, we borrow the idea of a *utility function* from measurement theory, as described by, e.g., Scott [84]. Theorem 2.22 below will show that a quadratic reaction structure makes it possible to quantify each species' contribution to the tier ordering individually via a utility function, and thus frees us to work with tiers *either* on a per-species or a per-complex basis, interchangeably. The proof of Theorem 2.21 will follow quite easily from this fact.

In what follows, let $\succeq$ denote the ordering over complexes into tiers along a subsequence for which the partition is well-defined.

Definition 2.22 (Realizability by a utility function)**.** For a given set $\mathcal{S}$ and an ordering $\succeq$ over $\mathcal{S} \times \mathcal{S}$, $\succeq$ is *realizable by a utility function* ϕ if and only if there exists a function $\phi : \mathcal{S} \to \mathbb{R}$ such that, for any pair (a, b) and $(c, d) \in \mathcal{S} \times \mathcal{S}$,

$$(a, b) \succeq (c, d) \Leftrightarrow \phi(a) + \phi(b) \geq \phi(c) + \phi(d)$$

Naturally, it would be useful if this definition applied to the tier ordering, and indeed it does:

Theorem 2.22. *Given a proper tier sequence $(x_n)_{n=0}^{\infty}$ of a quadratic reaction network, the ordering $\succeq_{(x_n)}$ of complexes into tiers along x_n is realizable by a utility function.*

Proof. The ordering of complexes over tiers can be formulated as a "problem of ordered differences" as described by Scott [84]. Note that $\succeq_{(x_n)}$ is a quaternary relation on $\mathcal{S}$ (that is, $\succeq_{(x_n)}$ compares $y, y' \in \mathcal{S} \times \mathcal{S}$). [84] shows that for a quaternary relation $\succeq$ to be realizable by a utility function, it is necessary and sufficient for the following three conditions to hold for all $a, b, c, d \in \mathcal{S}$, all sequences $a_0, \dots a_{n-1}, b_0, \dots, b_{n-1}$ (with $a_i, b_i \in \mathcal{S}$, $0 < n \leq |\mathcal{S}|$), and all permutations π, σ of $\{0, \dots, n-1\}$:

1. $(a, b) \succeq (c, d)$ or $(c, d) \succeq (a, b)$

2. $(a, b) \succeq (c, d) \Rightarrow (b, a) \succeq (d, c)$

3. If $(a_i, b_{\sigma(i)}) \succeq (b_i, a_{\pi(i)})$ for all i such that $0 < i < n$, then $(a_{\pi(0)}, b_0) \succeq (b_{\sigma(0)}, a_0)$

Letting $\succeq_{(x_n)}$ be the total ordering by tiers along $(x_n)_{n=1}^{\infty}$, all pairs of species are comparable, so condition 1 holds. The complexes (a, b) and (b, a) are equivalent ways to denote the complex containing species a and b, so we have the commutative property $(a, b) \sim_{(x_n)} (b, a)$ for all complexes (a, b), from which condition 2 follows directly.

It remains to show that condition 3 holds. To see this, consider any sequence of species $a_0, \dots a_{n-1}, b_0, \dots, b_{n-1}$ with $n > 0$, and any permutations π and σ of $\{0, \dots, n-1\}$, where

$$\lim_{n\to\infty} \frac{x_{a_i} x_{b_{\sigma(i)}}}{x_{a_{\pi(i)}} x_{b_i}} = k_i < \infty \quad \text{for all } i \text{ such that } 0 < i < n.$$

Suppose that condition 3 is false; that is, for some such construction, it is also the case that

$$\lim_{n\to\infty} \frac{x_{a_0} x_{b_{\sigma(0)}}}{x_{a_{\pi(0)}} x_{b_0}} = 0.$$

Then we can consider the product of all the terms:

$$\lim_{n\to\infty} \prod_{i=0}^{n-1} \frac{x_{a_{\pi(i)}} x_{b_i}}{x_{a_i} x_{b_{\sigma(i)}}} = 0 \cdot \prod_{i=1}^{n-1} k_i = 0$$

However, the product on the LHS of the above expression is simply a reordering of

$$\prod_{i=0}^{n-1} \frac{x_{a_i} x_{b_i}}{x_{a_i} x_{b_i}} = 1$$

which is a contradiction. So condition 3 also holds. □

We now know that $\succeq_{(x_n)}$ is realizable by a utility function $\phi : \mathcal{S} \to \mathbb{R}$. Intuitively, such a utility function gives us a way to translate the ordering over pairs of species into a concrete measure on each individual species. To finish the construction, let $m = \min_{s\in\mathcal{S}} \phi(s)$. Then define $\psi(s) = \phi(s) - m$, noting that subtracting a constant preserves the utility function condition:

$$\begin{aligned} &\psi(a) + \psi(b) \geq \psi(c) + \psi(d) \\ \Leftrightarrow\ &\phi(a) + \phi(b) - 2m \geq \phi(c) + \phi(d) - 2m \\ \Leftrightarrow\ &(a, b) \succeq_{(x_n)} (c, d) \end{aligned}$$

Thus we may assume without loss of generality that the utility function ψ guaranteed by Theorem 2.22 satisfies the additional property that $\psi(s) \geq 0$ for all $s \in \mathcal{S}$ with $\psi(a) + \psi(b) = 0 \Leftrightarrow \psi(a) = \psi(b) = 0$.

Corollary 2.23. *If $\exists b \in \mathcal{S}$ such that $\lim_{n\to\infty}(x_n)_b = 0$, then for all $a \in \mathcal{S}$,*

$$\psi(a) = 0 \Leftrightarrow \lim_{n\to\infty} (x_n)_a \neq 0$$

Proof. Let $z = \lim_{n\to\infty} x_n$ (recall that the limit exists because x_n is tier-preserving). For any $a, b \in \mathcal{S}$ such that $z_a, z_b > 0$, select some $c \in \mathcal{S}$ such that $z_c > 0$ (there is no requirement for a, b, or c to be distinct, so this is always possible – see [7] for a discussion of why $z = \vec{0}$ is not a limit point of any trajectory from an interior starting point).

Then $\frac{x_a(q)x_c(q)}{x_b(q)x_c(q)} \to \frac{z_a}{z_b} = k$ for some constant $0 < k < \infty$. Thus $(a, c) \sim_{(x_n)} (b, c)$, and by the utility function property, $\psi(a) + \psi(c) = \psi(b) + \psi(c)$, so $\psi(a) = \psi(b)$ for any a, b such that $z_a, z_b > 0$.

Further, for any a with $z_a > 0$, b with $z_b = 0$, and c with $z_c > 0$, $\lim_{n\to\infty} \frac{x_c x_b}{x_c x_a} = \frac{z_b}{z_a} = 0$, and thus $\psi(c) + \psi(b) > \psi(c) + \psi(a)$. So $\psi(b) > \psi(a)$ for any a, b with $z_a > 0, z_b = 0$.

Recalling that ψ is defined such that it achieves its minimum value of 0 (and that, by assumption, at least one $b \in \mathcal{S}$ has $\lim_{n\to\infty}(x_n)_b = 0$), it follows that $\psi(a) = 0 \Leftrightarrow z_a > 0$ (which is equivalent to $z_a \neq 0$ given that $z_i \geq 0\ \forall\ i$). □

Theorem 2.21 now follows immediately from Theorem 2.22, letting $w_i = \psi(i)$ as defined in the proof of Corollary 2.23.

Remark 2.24. A similar result to Theorem 2.21 was proved independently by a different method in [6], Lemma 4.4.

A Generalization of Tiers

Unlike prior methods, the utility function method as described above also allows for a generalization of Corollary 2.20. Other orderings over the set of complexes may also admit utility functions $\mathcal{S} \to \mathbb{R}_{\geq 0}$, and if those functions are zero for all species in the support of a limit point z yet not identically zero overall, the same argument holds to show that there is an edge between some complexes in different equivalence classes under that total order. We summarize this observation in the following theorem:

Theorem 2.25 (Realizable orderings)**.** *For any boundary limit point z of $x(t)$, and any total ordering $\succeq$ on $\mathcal{C}$ which is realizable by a utility function $\psi : \mathcal{S} \to \mathbb{R}_{\geq 0}$ with $\psi \neq \vec{0}$ and $\psi(a) = 0$ for all $a \in supp(z)$, there exists $w \in \mathbb{R}_{\geq 0}$ such that for all $y, y' \in \mathcal{C}$,*

$$w \cdot y \geq w \cdot y' \Leftrightarrow y \succeq y'$$

and there exists a reaction $(y, y') \in \mathcal{R}$ such that $y \nsim y'$.

Examples

In each example below, consider a subsequence $(x_n)_{n=0}^{\infty}$ of times such that $\lim_{n\to\infty} x_n^{y'-y}$ is well-defined for all $(y, y') \in \mathcal{R}$.

Example 2.26 (Log-tiers)**.** *Let*

$$y \succeq_L y' \Leftrightarrow \lim_{n\to\infty} \frac{\log(x_n^y)}{x_n^{y'}} = 0.$$

(That is, if y is in a higher equivalence class than y', x^y approaches zero exponentially faster than $x^{y'}$.) Then $\succeq_L$ is realizable by a utility function.

Example 2.27 (Approaching a constant above or below 1)**.** *Let $M_{y,y'} := \lim_{n\to\infty} x_n^{y'-y}$.*
Define $\succeq_A$ by: $\begin{cases} y \succ_A y' & \text{if } M_{y,y'} \in (1, \infty] \\ y \prec_A y' & \text{if } M_{y,y'} \in [0, 1) \\ y \sim_A y' & \text{if } M_{y,y'} = 1. \end{cases}$
Then $\succeq_A$ is realizable by a utility function.

Example 2.28 (Approaching from the positive or negative side)**.** *Take some subsequence $(x_m)_{m=0}^{\infty}$ of $(x_n)_{n=0}^{\infty}$ along which, if $\lim_{m\to\infty} x_m^{y'-y} = 1$, it approaches 1 from either only the positive side, only the negative side, or is exactly 1 on the entire subsequence. Denote $L_{y,y'} := 1^+, 1^-$, or 1° in those cases respectively, and otherwise let $L_{y,y'} := \lim_{m\to\infty} x_m^{y'-y}$.*
Define $\succeq_S$: $\begin{cases} (a, b) \succ_S (c, d) & \text{if } L_{ab,cd} \in (1, \infty] \cup \{1^+\} \\ (a, b) \prec_S (c, d) & \text{if } L_{ab,cd} \in [0, 1) \cup \{1^-\} \\ (a, b) \sim_S (c, d) & \text{if } L_{ab,cd} = 1^\circ \end{cases}$
Then $\succeq_S$ is realizable by a utility function.

In each of these examples, applying Theorem 2.25, if the subsequence in question approaches a boundary limit point, it is either the case that $\vec{0}$ is a utility function for $\succeq$ (in which case all pairs of species reside in the same equivalence class under $\succeq$), or that there exists a reaction (y, y') with $y \nsim y'$. While the following section will make use of the standard tier definition only, these examples provide a wider family of related results that could be applied in a similar way.

2.4.3 ELLT Networks are Persistent

The following approach unifies the previously known global attractor and persistence results for reversible single linkage and strongly endotactic networks as described in [50], [7], and [6], with a simpler proof that generalizes the result to the broader class of reversible "ELLT" networks. We first consolidate some language from the three prior papers.

Definition 2.23 (Transversal tier sequence)**.** A tier sequence is *transversal* if there exists at least one reaction $(y, y') \in \mathcal{R}$ such that $y \in T_i$, $y' \in T_j$, and $i \neq j$ under the tier ordering $\succeq_{(x_n)}$.

Definition 2.24 (Tier Descending)**.** Let T^* be the set of source complexes which are minimal among source complexes under the tier ordering $\succeq_{(x_n)}$. (Note that for reversible networks, all complexes are source complexes.) Then the tier sequence $(x_n)_{n=0}^{\infty}$ is *tier descending* if

1. For all $y \in T^*$ and all $(y, y') \in \mathcal{R}$ we have $y' \succeq_{(x_n)} y$; and

2. There exist $y \in T^*$ and $(y, y') \in \mathcal{R}$ with $y' \succ_{(x_n)} y$.

A reaction network is *tier descending* if every *transversal tier sequence* is *tier descending.*

With this terminology, Corollary 2.20 can be restated as saying that every proper tier sequence in a reaction network is transversal. (This is actually the form stated in [6].)

Definition 2.25 (Partition spanning)**.** A reaction network is *partition spanning* with respect to $T \subset \mathcal{C}$ if there exists an edge (y, y') such that $y \in T$, $y' \in \mathcal{C} \setminus T$.

With this language, we can now state a simpler sufficient condition for persistence, which as we will later see, using Theorem 2.25, encompasses both the strongly endotactic and single-linkage cases for reversible quadratic systems:

Definition 2.26 (ELLT)**.** Let T be the minimal nonempty tier under the tier ordering $\succeq_{(x_n)}$. A proper tier sequence is *ELLT* (*"Edge Leaving the Lowest Tier"*) if the reaction network is partition spanning with respect to T. A reaction network is ELLT if, for every convergent subsequence q of $x(t)$ that is is a proper tier sequence with limit point on the boundary of the state space, q is ELLT.

Note that the ELLT condition is *only* required along tier ordering(s) generated by convergent subsequences of $x(t)$, *not* along all vectors in $\mathbb{R}^N$. This is in contrast to the strongly endotactic condition described by Gopalkrishnan, Miller, and Shiu [50], which we show in Theorem 2.38 requires the ELLT condition on all vectors in $\mathbb{R}^N$.

With this definition, we can now state the main result of this section:

Theorem 2.29. *Every reversible, mass-preserving ELLT network is persistent; and the Global Attractor Conjecture holds for all reversible, mass-preserving ELLT networks.*

Proof. Fix a reversible reaction network under mass action kinetics β, with the ELLT property. Let $\tilde{R} = \{(y_i, y_j) \in \mathcal{R} \mid i < j\}$, the set of equivalence classes on reactions up to change of direction.

We analyze the behavior of $x(t)$ via the function $V_\mu(t) := \sum_{i=1^n} x_i \ln\left(\frac{x_i}{\mu_i}\right)$, with $\mu \in \mathbb{R}^N_{>0}$. When μ is complex balanced, this is the standard Lyapunov function often used in chemical reaction network theory; see, for example, [7]. However, in this case we proceed without any such restriction on the value of μ. Observing that

$$\frac{dV_\mu(x(t))}{dt} = \sum_{y,y' \in \tilde{R}} \beta_{y,y'} x^y(t) \left(1 - \frac{\beta_{y',y}}{\beta_{y,y'}} x^{y'-y}(t)\right) \left[\ln\left(x^{y'-y}(t)\right) - \ln\left(\mu^{y'-y}\right)\right],$$

we begin with a lemma:

Lemma 2.30. *Given a reversible ELLT network, suppose that $x(t)$ has no interior limit points. Then for every $\mu \in \mathbb{R}^N$, there exists t^* such that for all $t > t^*$, $\frac{dV_\mu(x(t))}{dt} < 0$.*

Proof. For a given proper tier sequence t_n, let $\bar{y}$ be some complex in the lowest tier.

Then, defining $m_{y,y'}$ to be the appropriate summand, we may write:

$$\frac{dV_\mu(x)}{dt} = x^{\bar{y}} \left[\sum_{y,y' \in \tilde{R}} \beta_{y,y'} x^{y-\bar{y}} \left(1 - \frac{\beta_{y',y}}{\beta_{y,y'}} x^{y'-y}\right) \left[\ln\left(x^{y'-y}\right) - \ln\left(\mu^{y'-y}\right)\right]\right] = x^{\bar{y}} \left[\sum_{y,y' \in \tilde{R}} m_{y,y'}\right]$$

Note that for any $a, c \in \mathbb{R}$ with $c > 0$,

$$\lim_{\theta \to 0} (1 - c\theta)(\ln(\theta) + a) = -\infty$$

$$\lim_{\theta \to \infty} (1 - c\theta)(\ln(\theta) + a) = -\infty$$

Let $T(y) : \mathcal{C} \to \{T_0, \cdots, T_k\}$ be the tier number of complex y. Then, analyzing each term of the summation, in the four possible cases:

1. $T(y) = 0 < T(y')$:
   $x^{y-\bar{y}} \to c \in \mathbb{R}_+$ as $n \to \infty$ and $x^{y'-y} \to 0$ as $n \to \infty$, so

$$\lim_{n \to \infty} m_{y,y'} = \beta_{y,y'} c \left(1 - \frac{\beta_{y',y}}{\beta_{y,y'}} x^{y'-y}\right) \lim_{n \to \infty} \left[\ln\left(x^{y'-y}\right) - \ln\left(\mu^{y'-y}\right)\right] = -\infty.$$

2. $T(y) = T(y') = 0$:
   $x^{y-\bar{y}} \to c \in \mathbb{R}_+$ as $n \to \infty$ and $x^{y'-y} \to c' \in \mathbb{R}_+$ as $n \to \infty$, so

$$\lim_{n \to \infty} m_{y,y'} = \beta_{y,y'} c \left(1 - \frac{\beta_{y',y}}{\beta_{y,y'}} c'\right) \left[\ln(c') - \ln\left(\mu^{y'-y}\right)\right] = c'' \in \mathbb{R}.$$

3. $T(y) = T(y') > 0$:
   $x^{y-\bar{y}} \to 0$ as $n \to \infty$ and $x^{y'-y} \to c \in \mathbb{R}_+$ as $n \to \infty$, so

$$\lim_{n\to\infty} m_{y,y'} = \beta_{y,y'} \left(1 - \frac{\beta_{y',y}}{\beta_{y,y'}} c\right) \left[\ln(c) - \ln\left(\mu^{y'-y}\right)\right] \lim_{n\to\infty} x^{y-\bar{y}} = 0.$$

4. $T(y) \neq T(y'), T(y) > 0, T(y') > 0$:
   $x^{y-\bar{y}} \to 0$ as $n \to \infty$ and $x^{y'-y} \to 0$ or $+\infty$ as $n \to \infty$, so

$$\lim_{n\to\infty} m_{y,y'} = \beta_{y,y'} \lim_{n\to\infty} (1 - cx^{y'-y}) \left[\ln(x^{y'-y}) + a\right] \cdot \lim_{n\to\infty} x^{y-\bar{y}} \leq 0.$$

Thus summing over all terms,

$$\lim_{n\to\infty} \sum_{y,y'\in\tilde{R}} m_{y,y'} = -\infty, \text{ and } \lim_{n\to\infty} x^{\bar{y}} = c_{\bar{y}} \in \mathbb{R}_{\geq 0}.$$

Thus there exists some n^* such that for all $n > n^*$, $\frac{dV_\mu(x(t_n))}{dt} < 0$.

Let us return now to the full sequence $x(t)$. For any subsequence $x(t_q)$ of $x(t)$, t_q has a convergent subsequence t_r which is a proper tier sequence, and along which we have proven that V_μ eventually decreases. Thus, there can be no infinite subsequence $x(t_q)$ of $x(t)$ such that $\frac{dV_\mu(x(t))}{dt} > 0$ for all $t \in t_q$, and so there is some time t^* such that for all $t > t^*$, $\frac{dV_\mu(x(t))}{dt} < 0$. □

Returning to the proof of Theorem 2.29, suppose for contradiction that $x(t)$ has no interior limit points. Then let $(x_n)_{n=0}^{\infty}$ be a convergent subsequence of $x(t)$, and let $z := \lim_{n\to\infty} x_n$. Following a proof strategy from [62], we will construct a vector μ such that, for any given t^*, some $x(t')$ with $t' > t^*$ has $x(t')$ and μ satisfying $\frac{dV_\mu(x(t'))}{dt} \geq 0$. In fact we find in Lemma 2.31 that this is the case for any trajectory with at least one boundary limit point.

Lemma 2.31. *Suppose a mass-preserving reaction network has a boundary limit point z. Then for any $t^* > 0$, there exists some $\mu \in \mathbb{R}^N$ and some time $t' > t^*$ such that $\frac{dV_\mu(x(t'))}{dt} \geq 0$.*

Proof. Let $S := supp(z)$ and $\beta := \min\{z_i | i \in S\}$. For a given x, let $T := \{i \in S | x_i > z_i\}$, and for any $\delta > 0$ define

$$(\mu_{x,\delta})_i = \begin{cases} z_i + \frac{2\delta}{|T|} - \frac{\delta}{|S|} & \text{for } i \in T \\ z_i - \frac{2\delta|}{|S-T|} - \frac{\delta}{|S|} & \text{for } i \in S \setminus T \\ \frac{\delta}{|\bar{S}|} & \text{for } i \in \bar{S}. \end{cases}$$

Note that this definition ensures that $\sum_{i\in\mathcal{S}}(\mu_{x,\delta})_i = 1$. (In fact, the values of $(\mu_{x,\delta})_i$ for $i \in \bar{S}$ need not be identical; the proof proceeds in the same way for any values of these entries subject to the constraint that $\sum_{i\in\bar{S}}(\mu_{x,\delta})_i = \delta$.)

For any $\gamma > 0$ and any $t^* > 0$, there is some $x(t) \in (x_n)_{n=0}^{\infty}$ such that $t > t^*$ and $|x_i(t) - z_i| < \gamma$ for all $i \in \mathcal{S}$. Let x^* be one such state, with $\gamma < \frac{\beta}{2}$. Set $\delta < \frac{1}{3}\beta$, so that $(\mu_{x^*,\delta})_i \geq \frac{\beta}{6}$ for all $i \in S$. We will refer to this $\mu_{x^*,\delta}$ by the name μ^* going forward.

Now define $f_i(x_i^*) = x_i^* \ln\left(\frac{x_i^*}{\mu_i^*}\right)$, with the continuous extension to $f_i(0) = 0$. For any $i \in S$ we have

$$\begin{aligned} f_i(x_i^*) - f_i(z_i) &= x_i \ln\left(\frac{x_i^*}{\mu_i^*}\right) - z_i \ln\left(\frac{x_i^*}{\mu_i^*}\right) \\ &= \left(1 + \ln\frac{z_i}{\mu_i^*}\right)(x_i^* - z_i) + \frac{2}{c}(x_i^* - z_i)^2 \end{aligned}$$

from the Taylor expansion, with $c \in [\min\{x_i^*, z_i\}, \max\{x_i^*, z_i\}]$. Note that this implies $c \leq z_i + \gamma < \beta + \frac{\beta}{2} = \frac{3\beta}{2}$.

$$f_i(x_i^*) - f_i(z_i) < (x_i^* - z_i) + \ln\left(\frac{z_i}{\mu_i^*}\right)(x_i^* - z_i) + 3\beta|x_i^* - z_i|^2$$

Noting that $(x_i^* - z_i)$ and $\ln\left(\frac{z_i}{\mu_i^*}\right)$ have opposite signs for $i \in S$, and letting $\alpha := \min\{\left|\ln\frac{z_i}{\mu_i^*}\right| \, | i \in S\}$, we have:

$$f_i(x_i^*) - f_i(z_i) \leq (x_i^* - z_i) - \underbrace{\alpha|x_i^* - z_i|}_{A} + \underbrace{3\beta|x_i^* - z_i|^2}_{B}$$

Note that the above holds for any $\gamma < \frac{\beta}{2}$, and no other quantities in the definition of μ^* depend on γ. In particular, for any γ sufficiently small, $|A| > |B|$ so that $f_i(x_i^*) - f_i(z_i) < (x_i^* - z_i)$.

Therefore we have

$$V_{\mu^*}(x^*) - V_{\mu^*}(z) = \sum_{i \in S}\left[f_i(x_i^*) - f_i(z_i)\right] + \sum_{i \in \bar{S}}\left[x_i^* \ln\frac{x_i^*}{\mu_i^*} - 0\right]$$

For $i \in \bar{S}$, $|x_i^* - z_i| = x_i^* < \gamma$; and for any $\gamma < \mu_i^*$, it is the case that $x_i^*/\mu_i^* < 1$, so for γ sufficiently small, we have

$$V_{\mu^*}(x^*) - V_{\mu^*}(z) < \sum_{i \in S}\left[f_i(x_i^*) - f_i(z_i)\right] \leq \sum_{i \in S}(x_i^* - z_i) < 0$$

with the final step following because $x_i^* > z_i^*$ for all $i \in \bar{S}$ and $\sum_{i \in \mathcal{S}} z_i = \sum_{i \in \mathcal{S}} x_i^* = 1$.

And so, for any $\gamma > 0$ sufficiently small, $V_{\mu^*}(x^*(t)) < V_{\mu^*}(z)$ for some $x^* = x(t)$ with $t > t^*$. Thus, by continuity of $V_{\mu^*}(x(t))$, for some $t' \geq t$ it must be the case that $\frac{dV_\mu^*(x(t'))}{dt} \geq 0$. $\square$

Observation 2.32. *The mass-preserving property is used only in the final step; without this property, it still holds that for γ sufficiently small,*

$$V_{\mu^*}(x^*) - V_{\mu^*}(z) < \sum_{i \in S} \left[f_i(x_i^*) - f_i(z_i) \right].$$

If $S = \emptyset$, then, $V_{\mu^*}(x^*) - V_{\mu^*}(z) < 0$, and the remainder of the proof using continuity of $V_{\mu^*}(x(t))$ holds. Therefore, Lemma 2.31 also holds in the setting where $S = \emptyset$, even if the reaction network is not *mass-preserving.*

Corollary 2.33. *If trajectory $x(t)$ of a mass-preserving reaction network has a single boundary limit point z with support S, then for any $x \in \mathbb{R}^N$ sufficiently close to z, there exists some $\mu \in \mathbb{R}^N$ such that $V_\mu(x) < V_\mu(z)$. Furthermore, there exists some $\epsilon > 0$ such that for any specified set of values $\{m_i : i \in \bar{S} \mid \sum_{i \in \bar{S}} m_i = \epsilon\}$, there is some such μ with $\mu_i = m_i$ for all $i \in \bar{S}$.*

For any reversible ELLT system with no interior limit point, Lemma 2.31 directly contradicts our finding from Lemma 2.30. Therefore it is the case that every reversible ELLT system has at least one interior limit point.

In the complex balanced setting, recalling Theorem 2.17, this is already sufficient to prove that persistence and the global attractor condition hold. Similarly, in the quadratic setting, Rabinovich, Sinclair, and Wigderson [77] showed that, given a trajectory $x(t)$ of a quadratic dynamical system for which some convergent subsequence has limit z with full support, then z is stationary and moreover, $x(t) \to z$ overall.

Outside of those settings, it is possible that a network could have some combination of interior and boundary limit points. Corollary 2.34 shows that this is not the case: in fact, a reversible ELLT system has only a single limit point.

Corollary 2.34. *For any reaction network, if there exists some t^* and some set $M := \{\mu_1, \ldots, \mu_n\} \subset \mathbb{R}^N$ whose differences span $\mathbb{R}^N$ such that $\frac{dV_\mu}{dt} < 0$ for all $t > t^*$ and all $\mu \in M$, then $x(t)$ converges to a single limit point.*

Proof. Suppose $x(t)$ has distinct limit points z_1, z_2. Using Lemma 2.30 and continuity of V_μ,

$$V_{\mu_1}(z_1) - V_{\mu_1}(z_2) = V_{\mu_2}(z_1) - V_{\mu_2}(z_2) = 0$$

for any $\mu_1, \mu_2 \in M$. This happens if and only if $(z_1 - z_2) \cdot (\ln(\mu_1) - \ln(\mu_2)) = 0$. Since $\{\mu_1 - \mu_2 : \mu_1, \mu_2 \in M\}$ spans $\mathbb{R}^N$, $z_1 = z_2$. □

Finally, we are able to complete the proof of Theorem 2.29. Noting that, in the reversible, mass-preserving ELLT setting, the conditions of Corollary 2.34 hold for all $\mu \in \mathbb{R}^N$, and letting M be the standard basis vectors in $\mathbb{R}^N$, then $x(t)$ does indeed have a single limit point z. Thus we conclude that all reversible ELLT networks are in fact persistent. □

Observation 2.35. *We further note that the above argument is not unique to autonomous kinetics; substituting $\beta_{y,y'}(t)$ for $\beta_{y,y'}$, the same argument used to prove Theorem 2.29 also applies also to (1) reversible, mass-preserving ELLT networks under bounded-rate non-autonomous kinetics, and (2) reversible, non-mass-preserving ELLT networks under bounded-rate non-autonomous kinetics for which some limit point z has $supp(z) = \emptyset$.*

We note that this observation will become important to our treatment of networks under projection in Section 2.4.5.

2.4.4 Applications of the ELLT Property

The following examples illustrate how Theorem 2.29 can simplify existing proofs of the persistence and global attractor conjectures for specific classes of networks. We then explore the case of a specific reaction network which could not be analyzed with these previously existing methods, and apply Theorem 2.29 to prove that the persistence and global attractor conditions hold.

Single Linkage Networks

Definition 2.27 (Single linkage)**.** A reversible reaction network is *single linkage* when the reaction graph contains exactly one connected component (in the non-reversible case, the network is single linkage exactly when the reaction graph is weakly connected).

It follows immediately that a single linkage network is partition spanning with respect to any $T \subset \mathcal{C}$; and in particular, every single linkage network is ELLT, rederiving the result from Anderson [7] in the reversible setting.

Quadratic Endotactic Networks

The strongly endotactic property weakens the above partition spanning condition by asking it to hold only on certain types of sets T. To form T, we must pick a vector w, and define T to be the set of source complexes whose inner products with w are maximal. (A source complex is $y \in \mathcal{C}$ such that there exists a reaction (y, y')). Then the corresponding property is the existence of a reaction (y, y') such that $y \in T$ and $y' \notin T$. The following two definitions formalize this idea.

Definition 2.28 (w-maximal [6])**.** For a vector $w \in \mathbb{R}^{|\mathcal{S}|}$, a complex $y \in \mathcal{C}$ is *w-maximal* if (1) there exists a reaction (y, y') and (2) for any complex v such that there exists a reaction (v, v'), then $w \cdot y \leq w \cdot v$.

Note that in the case of an undirected reaction graph, this simply means that $w \cdot y \leq w \cdot y'$ for all $y' \in \mathcal{C}$; in the directed case, we only stipulate that this is true for complexes that are a source (as opposed to a product) in some reaction.

Definition 2.29 (Strongly Endotactic [6])**.** A reaction network is *strongly endotactic* if every $w \in \mathbb{R}^N$ not orthogonal to its stoichiometric subspace has the following properties:

1. If y is a w-maximal complex, then for all reactions $y \to y'$, $w \cdot y' \leq w \cdot y$

2. There exists a w-maximal complex y and a reaction $y \to y'$ with $w \cdot y' < w \cdot y$.

Note that for a network to be strongly endotactic, this property must hold for *all* $w \in \mathbb{R}^N$, not just those arising from a tier ordering; so this condition is significantly stricter than the ELLT condition.

Theorem 2.36 (Anderson et al. [6])**.** *A reaction network is strongly endotactic if and only if it is tier descending*

Theorem 2.37 (Gopalkrishnan, Miller, and Shiu [50])**.** *If a reaction network is weakly reversible and single linkage, then it is strongly endotactic.*

Using the utility function for tiers from Section 2.4.2, we can now observe that, in the reversible quadratic case, ELLT networks generalize strongly endotactic ones.

Theorem 2.38. *Every strongly endotactic (i.e., tier descending), weakly reversible, quadratic reaction network is ELLT.*

Proof. Consider a proper tier sequence with limit point z on the boundary of the state space. Let $w \in \mathbb{R}^N_{\geq 0}$ such that $w \cdot y \geq w \cdot y' \Leftrightarrow y \succeq y'$ and $w_I > 0 \Leftrightarrow I \notin supp(z)$ as per Theorem 2.22. Define $L = \min_{y \in C} \langle w, y \rangle$ and let T be the minimal tier under $\succeq$. Then $\langle w, y \rangle = L, \langle w, y' \rangle > L$ iff y and y' are in different tiers and $y \in T$. There are at least 2 nonempty tiers, as per Corollary 2.20, so some such y' exists, and w is not orthogonal to the stoichiometric subspace. Using the strongly endotactic condition, there exists some edge (y, y') such that $\langle w, y \rangle = L$ and $\langle w, y' \rangle > L$, which completes the proof. □

With this, a direct application of Theorem 2.29 completes a significantly simpler proof of persistence for strongly endotactic, weakly reversible reaction networks in the quadratic case.

2.4.5 Example: A Persistent, Downward Closed Reaction Network

This section presents an algorithm applying Theorem 2.29 to check for sufficient conditions for a given network to satisfy the persistence and global attractor conditions. The algorithm first obtains a list of candidate stationary supports T containing at least one state compatible (with respect to linear invariants) with the interior of the state space. Then, for each such T, a series of deductions based on a projection technique determine whether the ELLT condition is necessarily satisfied given a limit point with support T. We illustrate the algorithm on the following example.

Example 2.39. *Define the set of elements* $E = \{a, b, c, d, e, f\}$.
Define set

$$A = \{\{a, b, c, f\}, \{b, e, f\}, \{a, c, e\}, \{a, b, d\}, \{b, c, e\}, \{d, e, f\}, \{a, b, e\}\},$$

and define the species set $\mathcal{S}$ *to be the downward closure of* A*, containing a total of 32 species, including the empty set. Define the reaction set* $\mathcal{R}$ *to be all quadratic reactions* (y, y') *with* $y, y' \in \mathcal{S} \times \mathcal{S}$ *such that* $\biguplus_{I \in y} I = \biguplus_{I \in y'} I$ *(where* $\biguplus$ *denotes multiset union). Let* $\mathcal{B} := \{\mathcal{S}, \mathcal{C}, \mathcal{R}\}$ *under autonomous mass-action kinetics.*

Observe that this example is not composed of a single linkage class, nor is it strongly endotactic (consider, for example, the vector w with $w_I = |I|$ for $|I| \in \{0, 1, 2\}$, and $w_I = 3$ otherwise); as a result, the prior work described in Section 2.4.1 is insufficient to prove whether or not this network is persistent.

Example 2.39 is also of combinatorial interest as an example of a *simplicial reaction network*, which will be the primary subject of study in Chapter 3. In particular, this means that the species set is *downward closed*: all species are subsets of a ground set E, and if $I \in \mathcal{S}$ and $J \subseteq I$, then $J \in \mathcal{S}$. The algorithm as described below uses this fact in order to take advantage of additional optimizations; however, the algorithm could easily be adapted to general reaction networks (with possible losses in computational tractability, depending on network size).

First Pass: Invariant Compatibility

The first pass of the algorithm eliminates all boundary stationary supports that are incompatible (with respect to invariants) with *all* interior points. That is, after this pass we are left with only those boundary supports which are stationary and which contain at least one point that is invariant-compatible with at least one interior point. One procedure to check this condition is summarized in Algorithm 1. In this algorithm, we use the fact that any set S with stationary support for a downward closed species set is necessarily upward closed on $\mathcal{S}$; this and other theorems relating to downward closed species sets can be found in Section 3.2.3.

Note in Algorithm 1 that for a fixed S, every stationary $S' \supset S$ with $(S' \cap A) = (S \cap A)$ is uniquely defined by $E^* \subseteq E$, such that $S' := S \cup \{I - J | I \in S, J \subset I, J \subset E^*\}$. For each candidate support S, this algorithm executes checks to rule out S if it is incompatible with invariants or lacks stationarity, and similarly for all $S' \subset S$ as described above.

The subroutine IS_STATIONARY$(S, \mathcal{B})$ simply checks whether, for each $(y, y') \in \mathcal{R}$ with $y \in S$, it is also the case that $y' \in S$. The subroutine FIND_INVARIANT uses a linear program to determine whether $\mathcal{B}$ gives rise to any linear invariant incompatible with support S, as detailed in Appendix A. We refer the interested reader to a detailed analysis in the appendix, with the conclusion that a single linear program suffices to determine whether each stationary support S contains any candidate limit points, and a related observation providing a matrix nullity condition sufficient to guarantee that a trajectory admits only a single limit point.

Algorithm 1 Find Compatible Supports for a Downward Closed Network

Require: Reaction network $\mathcal{B} := \{\mathcal{S}, \mathcal{C}, \mathcal{R}\}$ such that $\mathcal{S}$ is downward closed with element set E. $A \leftarrow$ set of maximal species in the system.

$\bar{A} \leftarrow$ all other species in the downward closure of A
$L \leftarrow \emptyset$, a list of candidate supports
for each nonempty subset $S \subseteq A$ **do**
    $\bar{S} \leftarrow (A \cup \bar{A}) \setminus S$
    if $\exists e \in E$ such that $e \notin I$ for all $I \in S$ **then** continue
    if not FIND_INVARIANT$(S, \mathcal{B})$ and IS_STATIONARY$(S, \mathcal{B})$ **then**
        $L \leftarrow L \cup S$
    for $E^* \subseteq E$ **do**
        $S' \leftarrow S \cup \{I - J | I \in S, J \subset I, J \subset E^*\}$
        if not FIND_INVARIANT$(S', \mathcal{B})$ and IS_STATIONARY$(S', \mathcal{B})$ **then**
            $L \leftarrow L \cup S'$
return L

For Example 2.39, Algorithm 1 performs 1403 stationarity checks and 3277 invariant checks (as performed by the linear program described in Appendix A). It returns a single boundary support, which happens to be exactly the set A. Note that this system contains 32 species total, and so naïvely checking all possible supports for stationarity and invariant compatibility would require at least 2^{32} such checks. With the algorithm as stated, we improve this to a (crude) upper bound of $\leq 2^{|A|} \cdot 2^{|E|} = 2^{13}$ checks. Preprocessing steps (as in, for example, the case when $\exists e \in E$ with $e \notin I$ for all $I \in S$), and the fact that we do not check stationarity if the invariant test does not pass, reduces this further to the observed number of calls.

Second Pass: Projection

We now know that there exists only a single boundary stationary support that could contain limit points of the trajectory. This enables the use of a projection argument, similar to that of [7]. Letting S represent the single candidate stationary support found in the previous pass, we *project* the network onto exactly those species in $\mathcal{S} \setminus S$ to obtain a network on a reduced species set, but with identical dynamics to the original reaction network.

Definition 2.30 (Reduced Network)**.** Given a reaction network $\mathcal{B} = \{\mathcal{S}, \mathcal{C}, \mathcal{R}, \beta\}$ and a support $S \subseteq \mathcal{S}$, we define the *reduced network* $\mathcal{B}' := \{\mathcal{S}', \mathcal{C}', \mathcal{R}', \beta'(t)\}$ obtained by projecting $\mathcal{B}$ onto $\mathcal{S} \setminus S$ as follows. Let $\mathcal{S}' := \mathcal{S} \setminus S$, and let

$$\mathcal{C}' := \{y' |\ y' = y \cap \mathcal{S}' \text{ for some } y \in \mathcal{C}\} \setminus \emptyset.$$

That is, any complex containing no species from $\mathcal{S}'$ is not included in $\mathcal{C}'$; all other complexes from $\mathcal{C}$ remain, but with any species not in $\mathcal{S}'$ removed. Note that there may be some pairs

of distinct complexes $y, y' \in \mathcal{C}$ such that $y \cap \mathcal{S}' = y' \cap \mathcal{S}'$; these become the same complex after projection, and their respective vertices in the reaction graph are accordingly merged. That is:

$$\mathcal{R}' := \{(y'_1, y'_2) \in \mathcal{C}' \times \mathcal{C}' \mid \exists (y_1, y_2) \in \mathcal{R} \text{ such that } y_1 \cap \mathcal{S}' = y'_1 \text{ and } y_2 \cap \mathcal{S}' = y'_2\}.$$

Finally, we define a specific set of *non-autonomous* rate functions $\beta'(t)$ to ensure that reduced network $\mathcal{B}'$ has identical dynamics to the initial network $\mathcal{B}$. To accomplish this, for any species $I \in S$, the behavior of $x_I(t)$ in $\mathcal{B}$ must be included as a factor in the rate function of any reaction (y, y') with $I \in y$. That is: for a given reaction $(y, y') \in \mathcal{R}'$, let

$$\beta'_{y,y'}(t) := \beta_{y,y'} \prod_{I \in (y \cap S)} x_I(t)$$

where $x_I(t)$ is the trajectory of species I in $\mathcal{B}$.

Because S contains only those species I such that x_I does not approach zero along any trajectory, incorporating any such $x_I(t)$ into a corresponding rate function $\beta_{y,y'}(t)$ preserves the property that $\beta_{y,y'}$ is positive and bounded away from zero. Furthermore, the reduced network remains reversible: a reaction $(y, y') \in \mathcal{R}$ remains in $\mathcal{R}'$ if and only if $y, y' \in \mathcal{C}'$; thus if $(y, y') \in \mathcal{R}'$, so is (y', y). While the reduced network is *not* mass-preserving, it admits only a single boundary limit point $z = \vec{0}$. Therefore, if the ELLT property holds, the network will satisfy the conditions of Observation 2.35 to apply Theorem 2.29 and ultimately conclude that the system is persistent.

Indeed, in Example 2.39, we observe that the reduced network always has an edge leaving the lowest tier. This is computed by constructing a graph G whose vertices are $\mathcal{C}'$ along with an additional vertex v^*, and placing an edge between two complexes if they are either both in the lowest tier, or both not in the lowest tier, under the assumption there is no edge leaving the lowest tier. We also add an edge between any complex which cannot reside in the lowest tier and v^*. The procedure to add these edges is summarized in Algorithm 2. If this procedure generates a graph with all complexes residing in a single connected component, then we conclude that either there is only a single tier, or the initial assumption was false and there is an edge leaving the lowest tier in $\mathcal{B}'$.

Recalling from Corollary 2.20 that all proper tier sequences are transversal, there must be an edge between complexes from two different tiers, so it is not the case that $\mathcal{B}'$ has only a single tier. Rather, if G has a single connected component, we conclude that the reduced network has an edge leaving the lowest tier along every convergent subsequence of $x(t)$, and so $\mathcal{B}'$ is a reversible ELLT network with bounded-rate non-autonomous mass-action kinetics. Thus by Theorem 2.29 and Observation 2.35, the system is persistent and, if complex-balanced, the unique interior equilibrium point for any given initial condition is a global attractor. In particular, as shown in Figure 2.4, the equivalence graph generated by Example 2.39 after projection onto $\mathcal{S} \setminus S$ has a single connected component, and we conclude that the reduced system is persistent.

Algorithm 2 Construct equivalence class over tiers

Require: Reaction network $\mathcal{B}' := \{\mathcal{S}', \mathcal{C}', \mathcal{R}'\}$.

$V \leftarrow \mathcal{C}' \cup v^*$
$E \leftarrow \emptyset$
for $(y, y') \in \mathcal{R}'$ **do**
    $E \leftarrow E \cup (y, y')$ ▷ No edge leaves the lowest tier, so y is in the lowest tier iff y' is.
for $y \in \mathcal{C}'$ **do**
    if $y' \subset y$ for some $y' \in \mathcal{C}'$ **then**
        $E \leftarrow E \cup (y, v^*)$ ▷ y' is in a lower tier than y, so y is not in the lowest tier.
return G

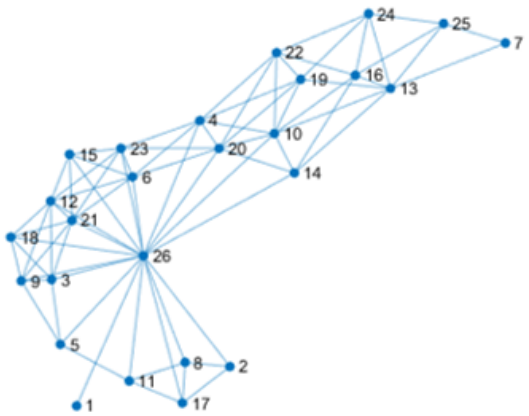


Figure 2.4: (Part of) the graph G generated by applying Algorithm 2 to Example 2.39 after projection. Vertex 26 represents v^*. Vertices 1 through 25 represent the single-species complexes in $\mathcal{C}'$, which include $\emptyset$ and $\{I\}$ for every species $I \in \mathcal{S}'$. All other complexes in $\mathcal{C}'$ are not pictured; we note that each has an edge to v^* from the second for-loop of Algorithm 2. Note that G has a single connected component.

Because the systems $\mathcal{B}$ and $\mathcal{B}'$ share identical dynamics, persistence of the reduced network $\mathcal{B}'$ implies that any limit point of $\mathcal{B}$ is also supported on $\mathcal{S}'$. Based on the deductions in the first pass of the algorithm, the only possible support satisfying this condition is $\mathcal{S}$; that is, the initial system $\mathcal{B}$ is also persistent and, if complex-balanced, satisfies the global attractor condition.

2.4.6 Edges within the Lowest Tier

Building further on the analysis in Section 2.4.3, we look beyond the class of networks for which an edge leaving the lowest tier can be guaranteed. In the absence of such an edge, we ask: Which other reactions dominate the eventual dynamics of the system?

Theorem 2.40. *Given a symmetric, quadratic reaction network, suppose that for any proper tier sequence such that no edge leaves the lowest tier, there instead exists at least one reaction (y, y') with $y \in T_0$ such that $\lim_{n\to\infty} x^{y'-y}(t_n) \neq 1$. Then the following all hold:*

1. *There exists some $\mu \in \mathbb{R}^N$ and some t^* such that $\frac{dV_\mu}{dt} < 0$ for all $t > t^*$.*
2. *There exists a set $\{\mu_1, \ldots, \mu_N\}$ spanning $\mathbb{R}^N$ such that each μ_i satisfies the above condition.*
3. *The trajectory $\{x(t)\}$ has a single limit point z.*

Proof. We begin with a similar analysis to that of Lemma 2.30. For a given proper tier sequence t_n, let $\bar{y}$ be some complex in tier T_0. Again defining $m_{y,y'}$ to be the appropriate summand, and using the fact that the reaction network is symmetric, we may write:

$$\frac{dV_\mu(x)}{dt} = x^{\bar{y}} \left[\sum_{y,y' \in \tilde{R}} \beta_{y,y'} x^{y-\bar{y}} \left(1 - x^{y'-y}\right) \left[\ln\left(x^{y'-y}\right) - \ln\left(\mu^{y'-y}\right) \right] \right] = x^{\bar{y}} \left[\sum_{y,y' \in \tilde{R}} m_{y,y'} \right].$$

As observed in the proof of Lemma 2.30, for any proper tier sequence t_n such that an edge leaves T_0, it holds that $\frac{dV_\mu(x)}{dt}$ is eventually decreasing for any choice of μ. It remains to consider those proper tier sequences without an edge leaving the lowest tier; that is, proper tier sequences such that all reactions (y, y') with $y \in T_0$ also have $y' \in T_0$. Consider some such sequence $\{t_n\}_{n=0}^{\infty}$. Let $T(y) : \mathcal{C} \to \{T_0, \cdots, T_k\}$ be the tier number of complex y. Then, analyzing each term of the summation in each of the three possible cases:

1. $T(y) = T(y') = 0$:
   Noting that $\lim_{n\to\infty} x^{y'-y}(t_n) = c \in \mathbb{R}_+$ as $n \to \infty$, if $c \neq 1$ we have
   $$\lim_{n\to\infty} \left(1 - x^{y'-y}(t_n)\right) \ln\left(x^{y'-y}(t_n)\right) < 0,$$
   and otherwise the limit approaches 0 from the negative side. Additionally, $x^{y-\bar{y}} \to c' \in \mathbb{R}_+$ as $n \to \infty$, so for some $c'' \leq 0$ we have
   $$\lim_{n\to\infty} m_{y,y'} = \beta_{y,y'} c' \left[c'' - 1 + \ln\left(\mu^{y'-y}\right) \right],$$
   with $c'' < 0$ strictly for at least one such term.

2. $T(y) = T(y') > 0$:
   $x^{y-\bar{y}} \to 0$ as $n \to \infty$ and $x^{y'-y} \to c \in \mathbb{R}_+$ as $n \to \infty$, so
   $$\lim_{n\to\infty} m_{y,y'} = \beta_{y,y'} (1 - c) \left[\ln(c) - \ln\left(\mu^{y'-y}\right) \right] \cdot \lim_{n\to\infty} x^{y-\bar{y}} = 0.$$

3. $T(y) \neq T(y'), T(y) > 0, T(y') > 0$:
   $x^{y-\bar{y}} \to 0$ as $n \to \infty$ and $x^{y'-y} \to 0$ or $+\infty$ as $n \to \infty$, so
   $$\lim_{n\to\infty} m_{y,y'} = \beta_{y,y'} \cdot \lim_{n\to\infty} (1 - x^{y'-y}) \left[\ln(x^{y'-y}) - \ln(\mu^{y'-y}) \right] \cdot \lim_{n\to\infty} x^{y-\bar{y}} \leq 0.$$

We now consider which $\mu \in \mathcal{R}^N$ satisfy Condition 1. Note that if $\mu^{y'} = \mu^y$ for a given $(y, y') \in \mathcal{R}$ with $T(y) = T(y') = 0$, the corresponding $m_{y,y'} = \beta_{y,y'} c' c'' \leq 0$, with strict inequality if $c \neq 1$. Choosing μ such that $\mu^{y'} = \mu^y$ for all $(y, y') \in \mathcal{R}$ is therefore sufficient to guarantee at least one term of $\sum_{y,y' \in \tilde{R}} m_{y,y'}$ satisfies $\lim_{n\to\infty} m_{y,y'} < 0$, and all other terms have $\lim_{n\to\infty} m_{y,y'} \leq 0$. Thus there exists some n^* such that for all $n > n^*$,

$$x^{\bar{y}}(t_n) \left[\sum_{y,y' \in \tilde{R}} m_{y,y'}(t_n) \right] < 0.$$

Because this argument holds for each proper tier sequence, and because there is some t^* such that for any $t > t^*$, $x(t)$ belongs to one of a finite number of proper tier sequences, it is the case that $\frac{dV_\mu}{dt} < 0$ for all $t > t^*$, and Condition 1 holds.

We additionally observe that the requirement that $\mu^{y'} = \mu^y$ can be relaxed: It is sufficient to have $|\ln(\mu^{y'-y}) - 1| < c$ for all $(y, y') \in \mathcal{R}$ to achieve the same result. As a result, there exists some $\epsilon > 0$ such that perturbing μ by up to ϵ in any index leaves $|\ln(\mu^{y'-y}) - 1| < c$, and so Condition 2 holds: there exists a set of vectors μ spanning $\mathcal{R}^N$ such that Condition 1 holds for each. Then by Corollary 2.34, the trajectory $\{x(t)\}$ has a single limit point overall. □

2.4.7 Excluded Orthants

We next explore an additional restriction arising from a similar argument to the preceding section, which constrains trajectories approaching boundary limit points for complex balanced systems. In particular, we find that for a given boundary limit point z, each positive equilibrium μ will rule out certain directions of approach for any trajectory approaching z.

Definition 2.31 (Orthant)**.** Define an *orthant* Q_z with respect to a point $z \in \mathbb{R}^{|\mathcal{S}|}$ as a partition of species set $\mathcal{S}$ into two groups, T and $\mathcal{S} \setminus T$, such that a point $x \in \mathbb{R}^{|\mathcal{S}|}$ is in the orthant $(T, \mathcal{S} \setminus T)$ with respect to z if for all $i \in T$, $x_i > z_i$, and for all $i \in \mathcal{S} \setminus T$, $x_i \leq z_i$.

Theorem 2.41. *For a reversible, complex balanced reaction network, an orthant Q_z with respect to a limit point z cannot contain a subsequence of trajectory points $x(t) \in Q_z$ if Q_z contains any positive equilibria.*

To see why this holds, we begin with a fundamental observation, inspired by the derivation in [44] of positive equilibria for complex balanced systems; here, we adapt it to describe even boundary equilibria for reversible reaction networks. In what follows, we define $\mathcal{R}_x^+$ to be the set $\{(y, y') \in \mathcal{R} \mid x^y > 0\}$; that is, reactions for which all reactants have positive concentration at x.

It is instructive to first understand the relationship between distinct equilibria of a given network. Under both detailed and complex balance, that relationship is characterized by a property known as *normality*.

Definition 2.32 (Normal point)**.** A vector $\pi \in \mathbb{R}^N_{>0}$ is *normal* with respect to a reaction network β with reaction set $\mathcal{R}$ if, for every reaction $(y, y') \in \mathcal{R}$, $\pi^y = \pi^{y'}$.

It is immediately evident that, for any two equilibria π and σ of a detailed balanced reaction network, $\frac{\pi^y}{\pi^{y'}} = \frac{\sigma^y}{\sigma^{y'}} = \frac{\beta_{y',y}}{\beta_{y,y'}}$, and so $\left(\frac{\pi}{\sigma}\right)^y = \left(\frac{\pi}{\sigma}\right)^{y'}$ for all $(y, y') \in \mathcal{R}$. It turns out that this condition is both necessary and sufficient for all pairs of equilibria, not only for detailed balanced reaction networks, but also for complex balanced ones, as shown in the following theorem.

Lemma 2.42. *For a reversible reaction network, $x \in \mathbb{R}^N$ with stationary support, and $\pi \in \mathbb{R}^N_{>0}$ a point of complex balance, the standard Lyapunov function V_π satisfies*

$$V'_\pi(x) \leq 0$$

with equality if and only if $\left(\frac{x}{\pi}\right)^y = \left(\frac{x}{\pi}\right)^{y'}$ for all $(y, y') \in \mathcal{R}^+_x$, and furthermore, x is an equilibrium if and only if equality holds.

Proof. Define $S := supp(x)$, and let $\ell := \begin{cases} \ln \frac{x_I}{\pi_I} & I \in S \\ 1 & I \notin S \end{cases}$.

Then we have:

$$V'_\pi(x) = \ell \cdot \dot{x} = \sum_{y \to y' \in \mathcal{R}} \beta_{y \to y'} x^y (y' - y) \cdot \ell.$$

Now, noting that all terms in $\mathcal{R}$ but not $\mathcal{R}^+$ have $x^y = 0$, this becomes

$$\begin{aligned}
&= \sum_{y \to y' \in \mathcal{R}^+} \beta_{y \to y'} \pi^y \frac{x^y}{\pi^y} (y' \cdot \ell - y \cdot \ell) \\
&= \sum_{y \to y' \in \mathcal{R}^+} \beta_{y \to y'} \pi^y e^{y \cdot \ln \frac{x}{\pi}} \left(y' \ln \frac{x}{\pi} - y \ln \frac{x}{\pi} \right) \\
&\leq \sum_{y \to y' \in \mathcal{R}^+} \beta_{y \to y'} \pi^y \left(e^{y' \ln \frac{x}{\pi}} - e^{y \ln \frac{x}{\pi}} \right) \\
&= \sum_{y \in \mathcal{C}^+} \left[\sum_{\mathcal{R}_{\to y}} \beta_{y' \to y} \pi^{y'} - \sum_{\mathcal{R}_{y \to}} \beta_{y \to y'} \pi^y \right] e^{y \ln \frac{x}{\pi}} = 0
\end{aligned}$$

where in the third line we used $e^p(p' - p) \leq e^{p'} - e^p$ for all p, with equality if and only if $p' = p$, and in the final line we used complex balance of π. Thus we have $V'_\pi(x) \leq 0$, with equality if and only if $y' \ln \frac{x}{\pi} = y \ln \frac{x}{\pi}$ for all reactions $y \to y' \in \mathcal{R}^+$; that is, if and only if $\left(\frac{x}{\pi}\right)^y = \left(\frac{x}{\pi}\right)^{y'}$ for all $y \to y' \in \mathcal{R}^+$. And furthermore, if x is an equilibrium, $V'_\pi(x) = 0$, and so this equality holds for all equilibria.

It remains to show that $\left(\frac{x}{\pi}\right)^y = \left(\frac{x}{\pi}\right)^{y'}$ for all $y \to y' \in \mathcal{R}^+$ implies that x is an equilibrium. Suppose that for all $(y, y') \in \mathcal{R}$,

$$\prod_{I \in y} \frac{x_I}{\pi_I} = \prod_{J \in y'} \frac{x_I}{\pi_J}.$$

Then we have, for any $y \in \mathcal{C}$ such that $x^y \neq 0$,

$$\begin{aligned}\sum_{\mathcal{R}_{\to y}} \beta_{y' \to y} x^{y'} - \sum_{\mathcal{R}_{y \to}} \beta_{y \to y'} x^y &= \sum_{\mathcal{R}_{\to y}} \beta_{y' \to y} \pi^{y'} \frac{x^{y'}}{\pi^{y'}} - \sum_{\mathcal{R}_{y \to}} \beta_{y \to y'} \pi^y \frac{x^y}{\pi^y} \\ &= \left(\frac{x}{\pi}\right)^y \left[\sum_{\mathcal{R}_{\to y}} \beta_{y' \to y} \pi^{y'} - \sum_{\mathcal{R}_{y \to}} \beta_{y \to y'} \pi^y\right] = 0\end{aligned}$$

and so x is complex balanced. In particular,

$$\dot{x} = \sum_{y \in \mathcal{C}} \left[\sum_{\mathcal{R}_{\to y}} \beta_{y' \to y} x^{y'} - \sum_{\mathcal{R}_{y \to}} \beta_{y \to y'} x^y\right] y = 0,$$

and x is an equilibrium. □

This in turn implies Theorem 2.11 for reversible reaction networks, as well as its extension to the boundary of the state space:

Corollary 2.43. *All equilibria of a complex balanced, reversible reaction network are complex balanced.*

Returning now to prove Theorem 2.41, we apply Lemma 2.42 and Lemma 2.31 to arrive at a contradiction.

Proof of Theorem 2.41. Suppose for contradiction that the trajectory $\{x(t)\}$ continues to pass through Q_z along some subsequence approaching limit point z; that is, for any $t > t_0$, $\exists t' > t$ such that $x_i(t') > z_i$ for all $I \in T$ and $x_i \leq z_i$ for all $I \in \mathcal{S} \setminus T$. Suppose in addition that there exists some positive equilibrium $\mu \in Q_z$ (note that this implies that for all $i \in \mathcal{S}$ not in the support of z, $i \in T$). Using Lemma 2.42, since μ is complex balanced, we should have $V'_\mu(x) \leq 0$ for all x with stationary support.

Fix some $\gamma > 0$ such that $\mu_i > \gamma$ for all $i \notin \mathit{supp}(z)$. Let x be a point on the trajectory $x(t)$ with $|z_i - x_i| < \gamma$ for all $i \in \mathcal{S}$. For γ sufficiently small, and using μ in the place of μ^*, it follows from an argument identical to Lemma 2.31 that there exists some such x with $V_\mu(x) < V_\mu(z)$. This contradicts the fact that $V_\mu(x)$ is decreasing for all x with stationary support (including for all $x \in \mathbb{R}^N_{>0}$). □

Theorem 2.41 puts strong restrictions on the directions of approach to any boundary limit point z for complex balanced reaction networks; and, if enough of the orthants Q_z

contain equilibria, may rule out z as a limit point entirely. A natural next question for a given candidate boundary limit point z is to ask is whether the existing normal points rule out so many angles of approach to z that it is altogether impossible to approach along any trajectory; or, if not, whether it meaningfully restricts the space through which the trajectory moves. An interesting first direction for future study would be an investigation of which combinations of recurring orthants are compatible with each other along a given trajectory, and how the positive equilibria of the corresponding system are distributed among those orthants.

Chapter 3

Simplicial Reaction Networks

> *Creativity [...] consists largely of rearranging what we know in order to find out what we do not know.*
>
> - George Kneller, *The Art and Science of Creativity*

In this chapter, we continue to focus on *structural* properties; that is, properties of the reaction graph which hold regardless of the values of the reaction constants on each directed edge. However, while past structural results have looked at properties which are generally expected to arise only in very small reaction networks, such as those which might be studied in a single chemical pathway, in this chapter we take a new approach which considers reaction networks over *combinatorial structures* that arise naturally from a given problem domain. The structure and properties of such networks scale to arbitrarily large problem sizes, and the problems which they address are of considerable interest in theoretical computer science, including the problem of sampling faces or facets of an abstract simplicial complex.

Considering first an analogy to real-world chemical reactions, the species of a chemical reaction network are not arbitrary abstract objects, but rather molecules such as $C_2H_3O_6$ or H_2O, each with a chemical formula describing which specific multiset of atoms comprise that species. A given set of molecules cannot simply react to produce any arbitrary set of products via a chemical reaction: Without resorting to nuclear fusion or fission, the number and type of atoms making up the molecules in a reaction must remain constant throughout. These rules give physically realistic chemical reaction networks additional structure not guaranteed by the most general reaction network model. Similarly, if a reaction network is used to model other transformations on sets of objects, it is natural to ask what constraints are imposed by that problem domain. Several combinatorially-inspired problems naturally give rise to analogous constraints, in which the reaction process redistributes conserved sub-units among different species.

Simplicial reaction networks make this idea concrete by stipulating that each species of a reaction network is composed of atomic *elements*, which can be neither created nor destroyed. For example, these elements may represent edges in a graph, basis vectors in a vector space,

or ground-set elements in a matroid. In each case, these elements are recombined by reactions into new combinatorial objects composed of the same base ingredients; and in particular, for simplicial reaction networks, a species can always be partitioned by a reaction into any two complementary subsets.

In the chapter that follows, we introduce simplicial reaction networks for the first time, and additionally define several specific families of reaction network structures which arise automatically from working with these combinatorial objects, each of which can be seen as an instance of a simplicial complex or its basis set. The combinatorial properties of these networks lead to strong guarantees about their behavior, including convergence to equilibrium via the persistence and global attractor conditions. These guarantees fulfill a critical prerequisite to using mass action kinetics in the design of new algorithms, and in particular, to build new algorithms for random sampling of combinatorial objects.

3.1 Simplicial Reaction Networks

To describe systems with a conserved sub-structure within each species, we coin the term *elemental*, referring to the idea that species are composed of elements drawn from a common ground set:

Definition 3.1 (Elemental)**.** A reaction network is *elemental* if every species $I \in \mathcal{S}$ represents a subset of elements from a finite ground set E, and the reactions satisfy $\beta_{y,y'} = 0$ if

$$\biguplus_{I \in y} I \neq \biguplus_{I \in y'} I$$

(where $\biguplus$ denotes multiset union).

We will choose special species sets for an elemental reaction network, such that useful combinatorial properties arise in the resulting systems. In particular, we focus on those which are downward closed under the subset operation, so that $\mathcal{S}$ represents the sets of an abstract simplicial complex. In order to make meaningful use of this condition, we also guarantee some minimal set of reactions between those species as follows:

Definition 3.2 (Simplicial Reaction Network)**.** An elemental reaction network is *simplicial* if, for every $I \in \mathcal{S}$ and for every $J \subset I$, $J \in \mathcal{S}$; and furthermore $\beta_{y,y'} > 0$ for a given $y, y' \in \mathcal{S} \times \mathcal{S}$ if and only if $\biguplus_{I \in y} I = \biguplus_{I \in y'} I$.

Note that this definition implies immediately that all simplicial reaction networks are reversible. Also note that this definition restricts attention to the quadratic case, defining the complex set $\mathcal{C} := \mathcal{S} \times \mathcal{S}$. In Sections 3.2 and 3.7, we will also see relaxations of the nonzero reaction constant requirement, demonstrating that in certain cases, only a small subset of these reactions are required in order to guarantee many of the same convergence properties enjoyed by various families of simplicial reaction networks.

We observe that, despite the fairly severe structural constraints placed on simplicial reaction networks, existing tools are apparently insufficient to analyze them in general. These tools, such as the single linkage [7] and strongly endotactic [50] theorems described in Chapter 2, tend to rely on strong connectivity properties in the reaction graph, properties which are not present in simplicial reaction networks. It is straightforward to observe that simplicial reaction networks have many linkage classes, as a separate linkage class exists for each unique multiset $\biguplus_{I \in y} I$ for $y \in \mathcal{C}$. Furthermore, to see that non-trivial simplicial reaction networks are not strongly endotactic, consider the vector w with $w_I = |I|$ for $|I| \in \{0, 1, 2\}$, and $w_I = 3$ otherwise. If there exist any species $I \in \mathcal{S}$ with $|I| \geq 2$, then there exists at least one reaction of the form $\{a\} + \{b\} \leftrightarrow \{a, b\} + \emptyset$ for some $a, b \in E$, and $\min_{y \in \mathcal{C}} w \cdot y = 2$. Note also that there can exist no reaction (y, y') with $w \cdot y = 2$ and $w \cdot y' > 2$. So for any such simplicial reaction network, w provides a simple counterexample to strong entotacticity. The more general tools described in Section 2.4 *do* provide convergence results for some such systems, but stop short of a full proof. In particular, an empirical search finds that while the linear programming method used to analyze Example 2.39 works in many cases, there exist examples of simplicial reaction networks on which this method, too, is insufficient to show persistence.

Instead, we take a new approach, and first derive basic properties of all simplicial reaction networks in Section 3.2, including a full characterization of the invariants of these systems. We then see a few examples of simplicial reaction networks in Section 3.3, narrowing our focus to the class of *matroid* reaction networks in Section 3.4. In the main result of Section 3.4, we prove that the basis exchange property of matroids together with the invariants of the network are sufficient to conclude that such networks satisfy persistence and the global attractor property. Section 3.5 subsequently provides a geometric interpretation of this result, leading to an alternate proof strategy based on the matroid polytope. We further detail how to translate between these two views using the example of matchings reaction networks in Section 3.6. Finally, Section 3.7 investigates a related class to simplicial reaction networks, in which only the maximal species (or *facets*) of a simplicial complex are included in $\mathcal{S}$, and provides a persistence and global attractor result for such networks when the underlying simplicial complex represents forests in a graph (so that the species in the reaction network are spanning trees).

3.1.1 Related Work

Examples of simplicial reaction networks can be found in earlier works, particularly in the symmetric quadratic operators over matchings in graphs as described by Rabinovich, Sinclair, and Wigderson [77]. Simplicial reaction networks generalize the matchings reaction network described in [77], removing the symmetry constraint and abstracting from matchings to other classes of simplicial complexes. As observed in that paper, this paradigm can also describe genetic algorithms with binary recombination (and without mutation) over any such combinatorial objects.

We also note here other definitions, similar to simplicial reaction networks, which can be found in prior literature; and note the subtle, yet important, properties which differentiate them from the simplicial reaction networks studied here. Most similar to the elemental condition described earlier is the primitive atomic condition defined in Doty and Zhu [36]. A reaction is considered *primitive atomic* if every species $I \in \mathcal{S}$ represents a nonempty multiset of elements from a finite ground set E, not necessarily unique, such that $\beta_{y,y'} = 0$ if $\biguplus_{I \in y} I \neq \biguplus_{I \in y'} I$. However, in contrast to definition Definition 3.1: (1) the empty set is not a valid element in primitive atomic networks, but is required in simplicial reaction networks; and (2) species in a primitive atomic network may be multisets over the elements of E, not just subsets.

Two other, more restrictive definitions from [36] are also comparable to simplicial reaction networks: *subset atomic* and *reachably atomic* reaction networks. *Subset atomic* reaction networks are primitive atomic with the additional requirement that for every $e \in E$, the set $\{e\}$ is represented by at least one $I \in \mathcal{S}$. More restrictively, *reachably atomic* requires that a reaction network be subset atomic and additionally, for any $I \in \mathcal{S}$, there must be a sequence of reactions whose input consists of the species I alone, and whose output is the complex $\{\{e\}$ for each $e \in I\}$.

In the primitive and subset atomic settings, whether persistence and other convergence properties hold remains an open question. In contrast, in the case of reachably atomic networks, persistence and the global attractor condition are both known to hold, as shown in [37]. However, the proof of this fact rests heavily on the assumption that every species can "spontaneously" (without requiring any other species to be present in the complex) decompose into precisely its constituent elements, which is unrealistic in the combinatorial setting we explore here. For example, note that no quadratic reaction network can be reachably atomic.

3.2 Properties of Simplicial Networks

In this section, we derive some foundational properties of simplicial reaction networks in their full generality, with the goal of fully characterizing the invariants of these systems, and deriving conditions to narrow down the space of possible limit points. These results are interesting in their own right, but will also be central to the proof of Theorem 3.14 in Section 3.4, which shows that the global attractor conjecture holds for the class of *matroid reaction networks*.

3.2.1 Normal Points

We begin by characterizing the *normal* states of a network, which we recall from Definition 2.32 are those states in $\pi \in \mathbb{R}^N_{>0}$ such that $\pi^y = \pi^{y'}$ for all reactions $(y, y') \in \mathcal{R}$. In the following theorem, we find a form that all normal points π of a simplicial reaction network must satisfy, and subsequently use it to construct a convenient basis for the space of invariants. In the

process, we also prove that the form holds for a larger class of networks which relax some requirements from the definition of simplicial reaction networks.

Theorem 3.1. *For a simplicial reaction network β, a point π with full support is normal with respect to β if and only if there exist constants λ_e for each $e \in E$ such that $\pi_I = \pi_\emptyset \prod_{e\in I} \lambda_e$.*

Proof. It is easy to check that any point with the specified form is normal. For the other direction, let π be a normal point on species set $\mathcal{S}$. For each $e \in \bigcup_{i\in\mathcal{S}} i$, define $\lambda_e = \pi_{\{e\}}/\pi_\emptyset$; note that each such λ_e is positive, finite, and well-defined when π has full support.

Now we proceed inductively to show that for every $I \in \mathcal{S}$, $\pi_I = \pi_\emptyset \prod_{e\in i} \lambda_e$. The claim is trivially true for all I such that $|I| = 0$ or $|I| = 1$. Suppose the claim is true for all $I \in \mathcal{S}$ of size $|I| < |J|$ for a given J with $|J| \geq 2$. Then, there exist $I, K, L \in \mathcal{S}$ of size $< |J|$ such that $\beta_{(I,J),(K,L)} > 0$ and $I \uplus J = K \uplus L$. In particular, this condition is satisfied by $I = \emptyset$, $K = J \setminus j$ for some $j \in J$, and $L = \{j\}$. By normality of π and the inductive hypothesis applied to I, K, and L,

$$\pi_I \pi_J = \pi_K \pi_L = \pi_\emptyset^2 \prod_{e\in K\uplus L} \lambda_e = \pi_\emptyset^2 \prod_{e\in I\uplus J} \lambda_e = \pi_I \left(\pi_\emptyset \prod_{e\in J} \lambda_e \right)$$

Since π has full support, $\pi_I \neq 0$, and so the claim holds for J, completing the induction. □

We see next how to extend Theorem 3.1 to points π on the boundary of the state space. Noting that any normal π must have stationary support, consider a fixed stationary support S. This support induces a partition of E in the following way:

Definition 3.3. An element $e \in E$ is *removable* if, for every $I \in S$ such that $e \in I$, $(I - e) \in S$. Conversely, an element $e \in E$ is *non-removable* if, for every $I \in S$ such that $e \in I$, $(I - e) \notin S$.

We confirm that these terms describe a partition of E with the following lemma:

Lemma 3.2. *In a simplicial reaction network, every element $e \in E$ is either removable or non-removable.*

Proof. Consider $I \in S$, with $e \in I$. Suppose first that $I - e \in S$. Then for any $J \in S$ with $e \in J$, $J + (I - e) \leftrightarrow (J - e) + I$ is a reaction; and because $I, J, (I - e) \in S$, we have $(J - e) \in S$ as well. Similarly, suppose $I - e \notin S$. Then for any $J \in S$ with $e \in J$, $J + (I - e) \leftrightarrow (J - e) + I$ has species$(I - e) \notin S$ on the left side of the reaction, and at least one species on the right side must not be in S; we know this is not the case for I, so it must be that $(J - e) \notin S$. □

Remark 3.3. As a result of Lemma 3.2, in a simplicial network we may partition any set of elements $I \subseteq E$ into its removable elements, denoted R_I, and its non-removable elements, denoted N_I.

We are now ready to state the analog of Theorem 3.1 for normal points on the boundary of the state space.

Theorem 3.4. *For any simplicial reaction network, a state π, possibly without full support, is normal if and only if the following three conditions hold:*

1. *π has stationary support S.*

2. *For all $(y, y') \in \mathcal{R}$ such that y and y' contain only species which are minimal with respect to set inclusion in S, $\pi^y = \pi^{y'}$.*

3. *There exist λ_e for all removable $e \in E$, such that for all $I \in S$,*

$$\pi_I = \pi_{N_I} \prod_{e \in R_I} \lambda_e.$$

Equivalently, for any $I \in S$ such that $I - e \in S$, $\pi_I = \lambda_e \pi_{I-e}$.

Proof. Suppose 1, 2, and 3 hold. Then for any reaction $(y, y') \in \mathcal{R}$,

$$\pi^y = \left[\prod_{e \in R_I \mid I \in y} \lambda_e \right] \prod_{I \in y} \pi_{N_I} = \left[\prod_{e \in R_J \mid J \in y'} \lambda_e \right] \prod_{J \in y'} \pi_{N_J} = \pi^{y'},$$

with the second-to-last equality following because y and y' contain the same)multi-)sets of removable elements; and because removing those elements from every species containing them yields a new reaction between minimal species from S.

For the other direction, it is immediate that 1 and 2 are necessary for normality. Suppose π is normal, and let $\lambda_e = \frac{\pi_e}{\pi_\emptyset}$ for each removable element $e \in E$. For any species I with some removable element e, consider the reaction

$$(I - e) + e \leftrightarrow I + \emptyset.$$

For π to be normal, we must have $\pi_{I-e}\pi_e = \pi_I \pi_\emptyset$ and so $\pi_I = \frac{\pi_{I-e}\pi_e}{\pi_\emptyset} = \lambda_e \pi_{I-e}$. □

Remark 3.5. In fact, only a subset of the reactions guaranteed by the simplicial property are required in the proof of Theorem 3.4; specifically, the reactions that represent removing a single element e from a species and adding it to the empty set.

3.2.2 Invariants

Using the observations about normal points, we will see that the invariants $q_e(x) = \sum_{I \in \mathcal{S} | e \in I} x_I$ form a basis for the space of linear invariants of any simplicial reaction network.

Theorem 3.6. *For a mass-preserving simplicial reaction network β, the invariants*

$$q_e(x) = \sum_{I \in \mathcal{S} | e \in I} x_I$$

together with the trivial invariant $\sum_{I \in \mathcal{S}} x_I$ form a basis for the space of linear invariants of β.

Proof. From Lemma 2.2, we know that the space of linear invariants is exactly

$$\left\{ \sum_{I \in \mathcal{S}} x_I \alpha_I \mid (y - y') \cdot \alpha = 0 \ \forall \ (y, y') \in \mathcal{R} \right\}.$$

For a given invariant $q(x) = \sum_{I \in \mathcal{S}} x_I \alpha_I$, and letting $\pi = e^{\alpha}$, we have $q(x) = \sum_{I \in \mathcal{S}} x_I \ln \pi_I$ with $(y - y') \cdot \ln \pi = 0$ for all $(y, y') \in \mathcal{R}$. That is, $\ln(\pi^y) = \ln(\pi^{y'})$ for all $(y, y') \in \mathcal{R}$, and so $q(x) = \sum_{I \in \mathcal{S}} x_I \ln \pi_I$ for a normal point π with full support. Applying Theorem 3.1, we have

$$\begin{aligned} q(x) &= \sum_{I \in \mathcal{S}} x_I \ln \left(\frac{1}{Z} \prod_{e \in I} \lambda_e \right) \\ &= \sum_{I \in \mathcal{S}} \left(x_I \ln \frac{1}{Z} + x_I \sum_{e \in I} \ln \lambda_e \right) \\ &= \ln \frac{1}{Z} \sum_{I \in \mathcal{S}} x_I + \sum_{I \in \mathcal{S}} \left(\sum_{e \in I} x_I \ln \lambda_e \right) \\ &= \ln \frac{1}{Z} \sum_{I \in \mathcal{S}} x_I + \sum_{e \in E} \ln \lambda_e \sum_{I \in \mathcal{S} | e \in I} x_I \\ &= \ln \frac{1}{Z} \sum_{I \in \mathcal{S}} x_I + \sum_{e \in E} \lambda_e q_e(x) \end{aligned}$$

Thus any linear invariant $q(x)$ can be written as a linear combination of $\{q_e(x)\}_{e \in E}$ together with the trivial invariant $\sum_I x_I = 1$. □

Because $q_e(x)$ is invariant along each trajectory $x(t)$, when the choice of trajectory is clear from context, we will refer to this invariant as q_e for simplicity.

3.2.3 Stationary Supports and Limit Points

Under the simplicial rule, stationary support sets take a very restricted form, limiting the space of locations where boundary limit points might be found.

Theorem 3.7 (Stationary Supports for Simplicial Networks)**.** *For a simplicial reaction network, let S be a stationary support. Suppose that, for all $e \in E$, there is some species $K \in S$ with $e \in K$. Then S is upward closed and $\overline{S}$ is downward closed.*

Proof. For any $I \in S$ and any $J \supset I$, let $b \in (J \setminus I)$. Then there is some $K \in S$ with $b \in K$. Now we consider the reaction

$$I + K \leftrightarrow (I + b) + (K - b)$$

By stationary support of S, $I, K \in S$ implies that $(I + b), (K - b) \in S$. Repeating with successive $b_1, \ldots, b_n \in (J \setminus I)$ yields $(I + b_1 + \ldots + b_n) = J \in S$. Thus S is upward closed, as desired. Further, suppose $I \notin S$, and $J \subset I$. Then $J \notin S$. So equivalently, $\overline{S}$ is downward closed. □

Remark 3.8. Note that for a limit point z of a trajectory with full-support initial condition, $supp(z)$ satisfies the condition in the above lemma: That is, z has stationary support and for any $e \in E$, $e \in I$ for some $I \in supp(z)$. So $supp(z)$ is upward closed for all such limit points.

Remark 3.9. For any stationary support $S \subset \mathcal{S}$, then for any $I, J \in S$ with $I \cap J = \emptyset$, there is no species $I \cup J$. To see why, suppose some such species did exist, and consider the reaction $I + J \leftrightarrow (I \cup J) + \emptyset$. Note that $\emptyset \notin S$, since upward closure of S would then imply $S = \mathcal{S}$. Furthermore, because $\emptyset \in \bar{S}$ and S is a stationary support, then either I or J (or both) are in $\bar{S}$, a contradiction.

We next apply Theorem 3.6 to weighted sums over the elements in a species to rule out certain boundary stationary supports S, as follows:

Theorem 3.10. *For a mass-preserving, simplicial mass action system, and a given support S, suppose $\exists w \in \mathbb{R}^{|E|}$ such that*

1. *$\sum_{e \in I} w_e = m$ for all $I \in S$ and for some constant m;*
2. *$\sum_{e \in J} w_e \leq m$ for all $I \in \bar{S}$; and*
3. *$\exists\ J' \in \bar{S}$ such that $\sum_{e \in J'} w_e < m$.*

Then S is not the support of a limit point for any trajectory with full-support initial condition.

Proof. Let π be a limit point with support S for the trajectory with initial condition $x(t_0)$.

$$\sum_{e \in E} w_e q_e(x(t_0)) = \sum_{I \in \mathcal{S}} \left[\sum_{e \in I} w_e x_I(t_0) \right] = \sum_{I \in \mathcal{S}} \left[x_I(t_0) \sum_{e \in I} w_e \right] = m \sum_{I \in \mathcal{S}} x_I(t_0) = m$$

However, at limit point π we also have

$$\sum_{e \in E} w_e q_e(\pi) = \sum_{I \in \mathcal{S}} \left[\pi_I \sum_{e \in I} w_e \right] = \sum_{I \in S} \left[\pi_I \sum_{e \in I} w_e \right] + \sum_{I \notin S} \left[\pi_I \sum_{e \in I} w_e \right] < m \sum_{I \in \mathcal{S}} \pi_I = m$$

A contradiction; so S is not the support of any limit point π. □

In Appendix A, we apply Theorem 3.10 to create a linear program for testing whether there exist invariants to rule out various limit point supports. The following example further demonstrates how invariants can be used to prove that a specific simplicial reaction network is persistent.

Example 3.11. *Suppose the two reactions over species set* $\{\emptyset, a, b, c, d, ab, cd\}$ *are:*

$$b + d \leftrightarrow bd + \emptyset$$

$$a + c \leftrightarrow ac + \emptyset$$

For a given trajectory $x(t)$, noting that $\sum_{i \in \mathcal{S}} x_i(t)$ is invariant,

$$\sum_{i \in \mathcal{S}} x_i(t) - q_c - q_d = x_a(t) + x_b(t) + x_\emptyset \text{ is invariant}$$
$$\sum_{i \in \mathcal{S}} x_i(t) - q_a - q_d = x_c(t) + x_b(t) + x_\emptyset \text{ is invariant}$$
$$\sum_{i \in \mathcal{S}} x_i(t) - q_c - q_b = x_a(t) + x_d(t) + x_\emptyset \text{ is invariant}$$
$$\sum_{i \in \mathcal{S}} x_i(t) - q_a - q_b = x_c(t) + x_d(t) + x_\emptyset \text{ is invariant}$$

Now let us examine a candidate limit point support S, for a trajectory with full-support initial condition. If $\emptyset \notin S$, then within each set $\{a, b\}, \{a, d\}, \{c, b\}, \{c, d\}$, at least one of the one-element species must be in S in order to maintain the above invariants. It is straightforward to check that this implies that either $a, c \in S$, or $b, d \in S$. But if either of these is the case, then stationarity implies that $\emptyset \in S$. We have already seen that if $\emptyset \in S$, S has full support. So this reaction network is persistent.

3.3 Classes of Simplicial Reaction Networks

We first define the class of *matroid reaction networks* $\beta_{\mathcal{M}}$, acting on the independent sets of a matroid $\mathcal{M}$. We begin by recalling the definition of a matroid in terms of its independent sets $\mathcal{S}$ [75].

Definition 3.4. A *matroid* $\mathcal{M}$ is an ordered pair $(E, \mathcal{S})$ consisting of a finite set E and a collection $\mathcal{S}$ of subsets of E having the following properties:

1. $\emptyset \in \mathcal{S}$

2. $\mathcal{S}$ is downward closed; so if $I \in \mathcal{S}$ and $I' \subseteq I$, then $I' \in \mathcal{S}$

3. $\mathcal{S}$ has the *matroid exchange property*: If I_1 and I_2 are in $\mathcal{S}$ and $|I_1| < |I_2|$, then there is an element e of $I_2 - I_1$ such that $I_1 \cup e \in \mathcal{S}$.

It follows that a matroid can be defined in terms of its maximal elements, called *bases*. In particular, we recall that for any bases B_1 and B_2 of matroid $\mathcal{M}$, $|B_1| = |B_2| = r$, the *rank* of the matroid. Furthermore, all matroids have the *basis exchange property*: For any bases B_1 and B_2 of $\mathcal{M}$ and any $x \in B_1 \setminus B_2$, there is an element $y \in B_2 \setminus B_1$ such that $(B_1 \setminus x) \cup y$ and $(B_2 \setminus y) \cup x$ are both bases of $\mathcal{M}$. For other equivalent definitions of matroids and useful theorems from matroid theory, see Oxley [75] and Schrijver [83].

Since matroids are simplicial complexes, it is straightforward to define a simplicial reaction network based on a matroid, as follows:

Definition 3.5 (Matroid Reaction Network)**.** Let $\mathcal{M}$ be a matroid with ground set E and independent sets $\mathcal{I}$, with $|\mathcal{I}| = N$. Assume without loss of generality that $\mathcal{M}$ contains no self-loops[1]. Let $\mathcal{S} := \mathcal{I}$, $\mathcal{C} := \mathcal{S} \times \mathcal{S}$, and for $(y, y') \in \mathcal{C}$, let $(y, y') \in \mathcal{R}$ and $\beta_{y,y'} > 0$ if and only if $\biguplus_{I \in y} I = \biguplus_{J \in y'} J$. Then the reaction network defined by $\{\mathcal{S}, \mathcal{C}, \mathcal{R}, \beta\}$ specifies the *matroid reaction network* $\beta_{\mathcal{M}}$ on $\mathcal{M}$.

Remark 3.12. Matroid reaction networks are mass-preserving, with $|y| = 2$ for all $y \in \mathcal{C}$; and so the trajectory $x(t)$ of $\beta_{\mathcal{M}}$ is contained in the simplex

$$\Delta_N = \left\{ x \in \mathbb{R}^n_{\geq 0} | \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i(t_0) \right\}.$$

We will typically consider matroid reaction networks to be normalized with $\sum_{i=1}^{N} x_i(t_0) = 1$.

Remark 3.13. Note that the exchange properties of matroids create a significant degree of combinatorial structure in matroid reaction networks. Consider, for example, a matroid reaction network which contains independent sets $\{a, b, c\}$ and $\{c, d, e\} \in \mathcal{S}$. The existence of these species implies (by downward closure) independence of $\{c\}, \{a, b\}$, and $\{d, e\}$, among other species, and thus we have, e.g., the resulting reaction $\{a, b\} + \{c, d, e\} \leftrightarrow \{a, b, c\} + \{d, e\}$ in $\mathcal{R}$.

The above would be true for any simplicial reaction network; but in the matroid setting, the matroid exchange property also guarantees that, given independent sets $A = \{a, b, c\}$ and $B = \{d, e\} \in \mathcal{S}$, we have $B \cup \{a\} = \{d, e, a\} \in \mathcal{S}$. This observation implies the existence of additional species and reactions in the system, such as $\{a, b, c\} + \{d, e\} \leftrightarrow \{b, c\} + \{d, e, a\}$.

We can also form simplicial reaction networks from other natural simplicial complexes in a similar way. For a concrete example, we take the case of matchings in a graph:

Definition 3.6 (Matchings Reaction Network)**.** A simplicial reaction network is a *matchings* reaction network if the ground set E is the edge set of some graph $G = (V, E)$, the species $\mathcal{S}$ are exactly the matchings in G, and the complex set $\mathcal{C} := \mathcal{S} \times \mathcal{S}$.

[1] A *self-loop* in a matroid is a dependent set consisting of a single ground set element. If a self-loop is present in $\mathcal{M}$, define matroid $\mathcal{M}'$ with ground set $E' := \{e \in E \mid e \text{ does not participate in a self-loop in } \mathcal{M}\}$, and independent sets $\mathcal{I}$. The dynamical systems $\beta_{\mathcal{M}}$ and $\beta_{\mathcal{M}'}$ are identical.

For many instances of simplicial complexes, there is significant computational relevance to developing sampling schemes, understanding the mixing of such processes, and characterizing their limiting distributions. Proving persistence is an important step towards a new approach that builds sampling algorithms based on reaction network dynamics, as we will discuss in Section 3.5.2.

We will see in Section 3.4 how persistence can be resolved in the case of matroid reaction networks, even under a relaxation of the circumstances under which $\beta_{y,y'} > 0$ is required; persistence is also known for matchings reaction networks, as discussed in Section 3.6. We conjecture that an analogous result can be derived for the full class of simplicial reaction networks; resolving the conjecture remains a major goal for future work in this area.

3.4 A Proof of the Global Attractor Conjecture for Matroid Reaction Networks

In this section, we prove the following theorem:

Theorem 3.14. *Let $\beta_{\mathcal{M}}$ be a matroid reaction network, with $x(t_0) \in \mathbb{R}^N_{>0}$. Then for any limit point π of $x(t)$ such that π has stationary support, $\pi \in \mathbb{R}^N_{>0}$.*

That is, under $\beta_{\mathcal{M}}$, any limit point with stationary support also has full support. Recalling from Section 2.2.5 that all limit points of a reaction network have stationary support, and applying Theorem 2.17, Theorem 3.14 also implies the following result.

Theorem 3.15. *Any matroid reaction network $\beta_{\mathcal{M}}$ with $x(t_0) \in \mathbb{R}^N_{>0}$ is persistent, and, if complex balanced, converges to the unique complex balanced equilibrium in its positive stoichiometric compatibility class.*

In other words, the global attractor and persistence conjectures hold for $\beta_{\mathcal{M}}$.

Remark 3.16. We will actually prove something stronger. For the purposes of proving Theorems 3.14 and 3.15, a relaxation of the quadratic completeness condition will be sufficient; we only need those reactions which exchange a single element. Furthermore, we don't need the reaction network to be quadratic; we just need it to be mass-preserving (i.e.: $\sum_{I \in \mathcal{S}} x_I(t) = 1$ for all times t). The quadratic condition is one way to achieve this; more generally, it is also true if all reactions $(y, y') \in \mathcal{R}$ have $|y| = |y'|$.

For the remainder of this section, let $x(t_0) \in \mathbb{R}^N_{>0}$, and let $\pi \in \mathbb{R}^N$ be a limit point of $x(t)$ with stationary support. Define $S := supp(\pi) \subseteq \mathcal{S}$.

Our primary tool will be a (partial) list of the invariants of $x(t)$. For $e \in E$, let $\mathcal{S}_e = \{I \in \mathcal{S} | e \in I\}$ and recall the marginal $q_e(x) = \sum_{I \in \mathcal{S}_e} x_I(t)$. Because $q_e(t)$ is invariant for all t, we denote it by the constant q_e (which depends on $x(t_0)$). As shown in Theorem 3.6, $\{q_e\}_{e \in E}$ together with the trivial invariant $\sum_i x_I = 1$ form a basis for the space of invariants of $\beta_{\mathcal{M}}$.

3.4.1 Stationary Supports Not Containing $\emptyset$

We first turn our attention to the possibility that $\emptyset \notin S$. Through Lemmas 3.17 and 3.18, we identify an invariant that is only consistent with initial conditions on the boundary of the state space if $\emptyset \notin S$; in particular, we show that $\emptyset \notin S$ implies that $x_\emptyset(t_0) = 0$.

Lemma 3.17. *Suppose $\emptyset \notin S$, and let $m = \min_{I \in S} |I|$. Then there exists some $U \subseteq E$ such that $rank(U) = |U \cap I| = m$ for all $I \in S$.*

Proof. Let

$$U := \bigcup_{J \in S \ s.t. \ |J|=m} J.$$

It is clear from the definition of U that $rank(U) \geq m$. We will now see that equality holds.

Let $A = \{a_1 \dots a_m\} \in S$ with $|A| = m$, and suppose there exists some independent set $B = \{b_1 \dots b_n\} \subseteq U$ with $|B| = n > m$. In particular, for each $b_i \in B$ there exists at least one $J_i \in S$ such that $b_i \in J_i$ and $|J_i| = m$; denote these as $J_1 \dots J_n$ respectively (not necessarily unique).

Then, by basis exchange, we have for some $b_i \in B$ a reaction

$$A + B \leftrightarrow (A + b_i) + (B - b_i).$$

From this we can deduce the existence of a second reaction

$$(A + b_i) + (J_i - b_i) \leftrightarrow A + J_i.$$

Note that $A, J_i \in S$, and S has stationary support, so $(A + b_i) \in S$ and $(J_i - b_i) \in S$. But $|J_i - b_i| = m - 1$, contradicting the assumption that $m = \min_{I \in S} |I|$. Thus, no independent set $B \subseteq U$ has $|B| > m$, and so $rank(U) = m$.

Now we consider $|U \cap I|$. For any $I \in S$ with $|I| > m$, pick an arbitrary $K \in S$ with $|K| = m$. By basis exchange, some reaction exists of the form

$$I + K \leftrightarrow (I - i) + (K + i)$$

where i is an element of I. Let $I' := (I - i)$.

Since $I \in S$ and $K \in S$, $I' \in S$. Thus, if there exists any $I \in S$ with $|I| > m$, then there exists some $I' \subset I$ with $I' \in S$ and $|I'| = |I| - 1$, and proceeding inductively, there exists some $I^* \subset I$ with $I^* \in S$ and $|I^*| = m$. It follows that $I^* \subseteq U$, and so $|I \cap U| \geq |I^*| = m$.

Furthermore, I is independent, so $|I \cap U| = rank(I \cap U) \leq rank(U) = m$. Thus $|I \cap U| = m$. □

Lemma 3.18. *Suppose $\emptyset \notin S$. Then $x_\emptyset(t_0) = 0$.*

Proof. We observe the following expression, invariant for all times t:

$$\sum_{u \in U} q_u = \sum_{I \in \mathcal{S}} |U \cap I| \cdot x_I(t). \tag{3.1}$$

Furthermore, note that $\sum_{u \in U} q_u$ counts the population of each species $I \in S$ exactly m times, since by Lemma 3.17 each $I \in S$ has exactly m elements in common with U. Hence we have

$$\sum_{u \in U} q_u = \sum_{I \in S} m \cdot x_I(t) + \sum_{J \in \mathcal{S}, J \notin S} |U \cap J| \cdot x_J(t) \tag{3.2}$$

for all times $t \geq t_0$. Note that each coefficient $|U \cap J| \leq m$ since $rank(U) = m$. Now at $t = t_0$, Equation (3.2) gives

$$\sum_{u \in U} q_u \leq m \cdot \sum_{I \in S} x_I(t_0) + m \cdot \left(1 - \sum_{I \in S} x_I(t_0)\right) = m \tag{3.3}$$

with strict inequality unless $x_I(t_0) = 0$ for all $I \in \mathcal{S}$ such that $|U \cap I| < m$.

Also, since π is a limit point, by continuity it must have the same value of the invariant $\sum_{u \in U} q_u$, and thus Equation (3.2) gives $\sum_{u \in U} q_u = m$. This contradicts Equation (3.3) unless $x_I(t_0) = 0$ for all $I \in \mathcal{S}$ with $|U \cap I| < m$.

□

3.4.2 Main Convergence Result

Now, we apply the previous result to show that the only stationary support S compatible with a fully-supported initial condition $x(t_0) \in \mathbb{R}^N_{>0}$ is $S = \mathcal{S}$.

Theorem 3.19. *Exactly one of the following holds:*

1. $S = \mathcal{S}$.
2. $q_e = 0$ *for some* $e \in E$.
3. $\emptyset \notin S$, $q_e \neq 0$ *for all* $e \in E$, *and* $x_\emptyset(t_0) = 0$.

Proof. It is straightforward to see that (1) and (2) are mutually exclusive, since for any $e \in E$, $\{e\} \in \mathcal{S}$ and $q_e = 0$ implies that $\{e\} \notin S$. It is also clear that (2) and (3) are mutually exclusive, and (1) and (3) are mutually exclusive (since $\emptyset \in \mathcal{S}$).
We next show that the conditions are exhaustive:

1. If $q_e = 0$ for some $e \in E$, then (2) holds.

2. If $q_e \neq 0 \ \forall \ e \in E$ and additionally $\emptyset \in S$, (1) holds.
   To see this: If there is any $e \in E$ such that $\{e\} \notin S$, then using downward closure of independent sets of a matroid, there exists a reaction $\emptyset + I \leftrightarrow (I \setminus \{e\}) + \{e\}$ for any $I \in \mathcal{S}$ containing e. Recalling that S has stationary support, $I \notin S$ for any $I \in \mathcal{S}$ containing e. This contradicts the assumption that $q_e \neq 0$, so $\{e\} \in S \ \forall \ e \in E$.
   Note that for any $I \in \mathcal{S}$ such that $I \setminus \{e\} \in S$, there exists reaction $\emptyset + I \leftrightarrow \{e\} + (I \setminus \{e\})$. Since $\emptyset, \{e\}, (I \setminus \{e\}) \in S$, and S has stationary support, $I \in S$.

   Proceeding inductively, we find that $I \in S$ for all $I \in \mathcal{S}$.

3. If $q_e \neq 0 \ \forall \ e \in E$ and additionally $\emptyset \notin S$, (3) holds.
   This is proven in Lemma 3.18: for any $I \in \mathcal{S}$ such that $|U \cap I| < m$, $x_I(t_0) = 0$. In particular, $x_\emptyset(t_0) = 0$.

This completes the proof. □

We now prove Theorem 3.14, which states that, for a matroid reaction network, if $x(t_0) \in \mathbb{R}^N_{>0}$, then any limit point π with stationary support satisfies $\pi \in \mathbb{R}^N_{>0}$.

Proof. Let π be a limit point of $x(t)$ with stationary support S, and suppose the initial condition $x(t_0)$ has full support. Note that $q_e = 0$ implies that the independent set $\{e\}$ has concentration 0 for all times t, including t_0; so this case is not possible given a full-support initial condition. Similarly, if $x_\emptyset(t_0) = 0$, then $x(t_0)$ is not fully supported. This rules out cases 2 and 3 in Theorem 3.19, so we have $S = \mathcal{S}$. □

Our main result, as stated in Theorem 3.15, then follows as a direct application of Theorem 3.14. Applying Theorem 3.14 to Theorem 2.8, all limit points of a matroid reaction network with initial condition $x(t_0) \in \mathbb{R}^N_{>0}$ must have support $\mathcal{S}$; and applying the same result to Theorem 2.17 in the complex balanced setting implies that the global attractor condition holds for all matroid reaction networks.

Remark 3.20 (Quadratic Exchange Reactions)**.** We note that the results in this section can be generalized in the following ways. First, it is immediate that the same proof holds for reaction networks containing complexes of size other than 2, as long as $\biguplus_{I \in y} I = \biguplus_{J \in y'} J$ (the elemental condition) and $|y| = |y'|$ (the mass-preserving condition). Moreover, we observe that the proofs above only made use of reactions of the form

$$A + B \leftrightarrow (A \setminus \{a\}) + (B \cup \{a\})$$

such that $a \in A$, $|A| > |B|$, and $A, B, (A \setminus \{a\}), (B \cup \{a\}) \in \mathcal{S}$. Defining the set of all such reactions as the *quadratic exchange reactions*, we therefore conclude that the same global attractor and persistence results hold as long as all quadratic exchange reactions are present in $\mathcal{R}$. Any other reactions satisfying the elemental condition, which may have any number of complexes and may or may not be reversible, can be added or removed without affecting the persistence result.

Remark 3.21 (Non-Autonomous Networks)**.** Similarly, we note that Theorem 3.15 still holds if the rate constants are allowed to vary over time, as long as they are bounded away from zero and $+\infty$. To see this, we observe that the same invariants q_e for all $e \in E$ and $\sum_{I \in \mathcal{S}} x_I = 1$ hold for non-autonomous matroid reaction networks. Therefore, Theorem 3.19 and Theorem 3.14 still hold: Any limit point with stationary support must have full support. Next, we observe that $\beta_{y,y'} x^y$ satisfies Assumption 2.4 from Anderson [9]: it is monotone increasing in x_I for each $I \in y$, equals zero if any such $x_I = 0$, and depends only on species $I \in y$. Thus, the non-autonomous version of Theorem 2.8 from that paper applies: all limit points of such networks have stationary support, and so we conclude that the all limit points of the non-autonomous matroid reaction network have full support.

This generalization highlights an advantage of the method used in the proof of Theorem 3.15: the only properties required are the invariants and the fact that all limit points have stationary support, eliminating any dependence on the specifics of the reaction rates or any other information about the equilibrium distribution which might be sensitive to changes in those rates.

3.5 Connections with the Matroid Polytope

The following section presents an alternate proof path for Theorem 3.15, which provides an additional layer of geometric intuition based on the matroid polytope. We will see that this proof enables a natural generalization of the persistence and convergence results to initial conditions without full support, given a straightforward condition on the system's invariants. We are also optimistic that this proof idea may be applicable to a broader class of combinatorial objects for which a corresponding polytope is well understood; for these reasons, we present the technique in its entirety despite having already proved the matroid result.

Consider a trajectory $x(t)$ of any elemental reaction network under which the invariants

$$\sum_{I \in \mathcal{S}:\ e \in I} x_I = q_e \text{ and } \sum_I x_I = 1$$

hold. We define a mapping ι from the simplex of probability distributions on $\mathcal{S} \subseteq 2^{|E|}$ to $\mathbb{R}^{|E|}_{\geq 0}$ such that every point on the trajectory $x(t)$ maps to a single point q, which is a convex combination of the species incidence vectors s_I, defined as follows:

$$q = \sum_{I \in \mathcal{S}} x_I s_I.$$

The invariant vector q specifies the values of invariants q_e for the entire trajectory $x(t)$.

Define the invariants polytope $conv(\mathcal{S})$ as the convex hull of the set of species incidence vectors $s_I \in \mathbb{R}^{|E|}$, $I \in \mathcal{S}$. Then every trajectory $x(t)$ maps to a single point $q \in conv(\mathcal{S})$. Suppose there exists some $f : 2^E \to \mathbb{R}_{\geq 0}$, such that for all $q \in conv(\mathcal{S})$ and all $U \subseteq E$,

$$\sum_{e \in U} q_e \leq f(U).$$

Then for any point $x \in \mathbb{R}^N_{\geq 0}$ and its corresponding invariant vector $q = \iota(x)$,

$$\sum_{e \in U} q_e = \sum_{I \in \mathcal{S}} |U \cap I| x_I,$$

and we have

$$\sum_{I \in \mathcal{S}} |U \cap I| x_I \leq f(U) = \sum_{I \in \mathcal{S}} x_I f(U).$$

Now, suppose additionally that there exists some $U^* \neq \emptyset$, such that for any I with $x_I > 0$, $|U^* \cap I| = f(U^*)$. Then for any x' with full support, such that $\iota(x) = \iota(x') = q_e$,

$$f(U^*) = \sum_{I \in \text{supp}(x)} |U^* \cap I| x_I = \sum_{I \in \text{supp}(x')} |U^* \cap I| x'_I \leq \sum_{I \in \text{supp}(x')} f(U^*) x'_I = f(U^*).$$

And so, given f, U^*, x, and x' as specified, equality holds in the above expression: for all $I \in \mathcal{S}$, $|U^* \cap I| = f(U^*)$. Put another way, if $|U^* \cap I| < f(U^*)$ for some $I \in \mathcal{S}$, then x is not on the same trajectory as any x' of full support.

3.5.1 Persistence of Matroid Reaction Networks

Suppose we have a matroid $\mathcal{M}$, and let $\mathcal{S}$ be the independent sets of $\mathcal{M}$. The matroid polytope[2] P is defined as the convex hull of the set of incidence vectors of $\mathcal{S}$, identical to our definition of the invariants polytope. We also define the following notation. For set $F \subseteq E$ and vector $q \in \mathbb{R}^{|E|}$, let $q(F)$ denote $\sum_{e \in F} q_f$.

Theorem 3.22 (Edmonds [39])**.** *The matroid polytope P is equivalent to*

$$P = \left\{ q \in \mathbb{R}^{|E|} : \begin{array}{ll} q_e \geq 0 & \forall\ e \in E \\ q(U) \leq rank(U) & \forall\ U \subseteq E \end{array} \right\}.$$

Now, we fix some x on the boundary of the state space. Let f be the rank function of $\mathcal{M}$. Then

$$|I \cap U| \leq rank(U)\ \forall U \subseteq E, I \in \mathcal{S}.$$

Theorem 3.23. *If x is stationary and does not have full support, then x maps to a point in a facet of the matroid polytope P.*

We will prove the theorem below; first, let us see its immediate consequences. If x maps to a facet of P, then at least one of the inequalities defining P is tight. If $q_e = 0$ for some

[2]Also known as the Independent Set Polytope

$e \in E$, we have already seen that the invariant q_e is incompatible with any fully-supported starting state $x(t_0)$. Alternatively, if $q(U^*) = rank(U^*)$ for some $U^* \subseteq E$ (with $U^* \neq \emptyset$), then

$$\sum_{I \in \mathcal{S}} |U^* \cap I| x_I = rank(U^*) = \sum_{I \in \mathcal{S}} rank(U^*) x_I$$
$$\Rightarrow |U^* \cap I| = rank(U^*) \ \forall \ I \in \mathcal{S}.$$

Yet

$$|\emptyset \cap U| = 0 < rank(U) \text{ for } \emptyset \subsetneq U \subseteq E.$$

And so in this case, too, x is not on the same trajectory as any fully-supported $x(t_0)$. Thus, we can conclude that for any $x(t_0)$ with full support, no points with the same invariants as $x(t_0)$, including any limit points of the trajectory, map to a facet of the matroid polytope.

Proof of Theorem 3.23

We first refer to a more precise characterization of the matroid polytope and its facets. If P is full-dimensional (which is the case if and only if $\mathcal{M}$ has no self-loops), there is a unique minimal collection of linear inequalities defining P up to scalar multiplication, and which correspond precisely to the facets of P.

We define a *flat* to be a subset $U \subseteq E$ such that $U = span(U)$, and we say a flat U is *inseparable* if there exist no flats U_1, U_2 partitioning U such that $rank(U_1) + rank(U_2) = rank(U)$.

Theorem 3.24 (Edmonds [40])**.** *If G has no self-loops, the following is a minimal system for the matroid polytope of G:*

1. *$q_e \geq 0$ for all $e \in E$*

2. *$q(U) \leq rank(U)$ for $U \subseteq E$ a nonempty inseparable flat.*

Accordingly, for P full-dimensional, $q \in \mathbb{R}^{|E|}$ is in a facet of P if and only if there is equality in one or more of the inequalities in this minimal system:

Corollary 3.25. *$q \in \mathbb{R}^{|E|}$ is in a facet of P iff at least one of the following is true:*

1. *$q_e = 0$ for some $e \in E$*

2. *$q(U) = rank(U)$ for some $U \subseteq E$ a nonempty inseparable flat.*

We restate Theorem 3.23 using this characterization of facets:

Theorem 3.26. *Let $S \subseteq I$ be all the independent sets in the support of some stationary distribution $z \in \mathbb{R}^{|I|}$ of $\beta^{\mathcal{M}}$. Let $q \in \mathbb{R}^{|E|}$ be the edge-distribution of z. Then one or more of the following hold:*

1. $S = I$

2. $q_e = 0$; *that is, some ground set element* $e \in E$ *is not included in any independent set in* S.

3. $q(U) = rank(U)$ *for some* $U \subseteq E$ *such that* U *is a nonempty inseparable flat.*

   Equivalently, for some $U \subseteq E$ *such that* U *is a nonempty inseparable flat, every independent set* $F \in S$ *has* $rank(F \cap U) = rank(U)$.

Proof. Fist we see that the two formulations of Condition 3 are equivalent. Every point in P is the convex combination of the edge-incidence vectors v_F of some independent sets $F \in \mathcal{S}$. Restricting these v_F to v'_F containing only those indices corresponding to $U \cap F$, if $\sum_F a_F v'_F(F) = rank(U)$ with $\sum_F a_F = 1$ (and a_F non-negative, and $v'_F(F) \leq rank(U)$), then $v'_F(F) = rank(F \cap U) = rank(U)$ for each F.

We next prove that at least one of the conditions must hold. Assume Conditions 1-3 are all false. We also assume without loss of generality that $\mathcal{M}$ is connected. (If $\mathcal{M}$ is not connected, it can be written as the direct sum of some nonempty connected matroids $\mathcal{M}_1, \ldots \mathcal{M}_m$; in that case, we can apply the following analysis independently to each $\mathcal{M}_i$.) We begin with two basic observations, stated in two lemmas.

Lemma 3.27. *There exists some* F *in* S *and some* $e \in E$ *such that* $e \notin span(F)$.

Proof. Note that E is an inseparable flat. So there exists some $F \in S$ such that $rank(F) < rank(E)$. That is, there is some independent set $F' \subseteq E$ such that $|F| = rank(F) < rank(E) = |F'|$. Thus by the independent set exchange property, there is some $e \in F' \setminus F$ such that $(F \cup \{e\})$ is independent. □

Lemma 3.28. *Let* F, e *be defined as in Lemma 3.27. Then for all* $G \in S$ *such that* $e \in G$, $(G \setminus \{e\}), (F \cup \{e\}) \in S$. *Furthermore, some such* G *exists.*

Proof. By assumption that Condition 2 is false, there is some $G \in S$ such that $e \in S$, and $G \neq F$. By the hereditary property, $(G \setminus \{e\}) \in \mathcal{S}$. Furthermore, $rank(F \cup \{e\}) = rank(F) + 1 = |F| + 1 = |F \cup \{e\}|$, since $e \notin span(F)$. So $(F \cup \{e\}) \in \mathcal{S}$. Thus $\beta^{\mathcal{M}}$ contains reaction $F + G \leftrightarrow F' + G'$, where $F' := (F \cup \{e\})$ and $G' := (G \setminus \{e\})$; and by stationarity of S, $F, G \in S$ implies that $F', G' \in S$. □

Returning now to the main proof of Theorem 3.26, for any edge e, let $S_e^+ \subseteq$ denote those sets in S that contain e, and $S_e^- := S \setminus S_e^+$. Note that stationarity of S implies that S_e^- is stationary. We will now show that S_e^- does not satisfy any of Conditions 1-3 on edge set $E \setminus \{e\}$. Suppose that it does:

Case 1: $S_e^- = \mathcal{S}_e^-$, all independent sets on $E \setminus \{e\}$.

If $S_e^+ = \emptyset$ or $\mathcal{S} \setminus S = \emptyset$, then $S = \mathcal{S}$ trivially (contradicting the assumption that Condition 1 does not hold for S); so assume neither is nonempty.

Let $G \in \mathcal{S} \setminus S$ (so $e \in G$), and $h \in S_e^+$. Then we have the reaction $(G \setminus \{e\}) + h \leftrightarrow G + (h \setminus \{e\})$. Noting that $(G \setminus \{e\}) \in S$ and $h \in S$, by stationarity of S, we have $G \in S$. Contradiction.

Case 2: For some $d \in (E \setminus e)$, there is no $h \in S_e^-$ such that $d \in h$.

By assumption that S does not satisfy Condition 2, there is some $G \in S_e^+$ which does contain d. Then for F as defined earlier, we have reaction $(F \cup \{e\}) \in S$ and $(G \setminus \{e\}) \in S$. But $d \in (G \setminus \{e\}) \in S_e^-$. Contradiction.

Case 3: There exists $U \subseteq (E \setminus \{e\})$ an inseparable flat such that for all $h \in S_e^-$, $rank(F \cap U) = rank(U)$.

Then consider arbitrary $G \in S_e^+$. By Lemma 3.28, $(G \setminus \{e\}) \in S_e^-$. Then since $e \notin U$, $rank((G \setminus \{e\}) \cap U) = rank(G \cap U) = rank(U)$.

Let $U' = span(U)$. Then $rank(U') = rank(U) = rank(G \cap U)$ for all $G \in S_e^+$.

If $U' = U$, U' is inseparable. Else, $U' = U \cup \{e\}$. In this case, suppose

$$rank(U_1) + rank(U_2) = rank(U') = rank(U)$$

for some partition (U_1, U_2) of U'. Then (assuming WLOG that $e \in U_1$)

$$rank(U_1 \setminus \{e\}) + rank(U_2) \leq rank(U)$$

and since $rank(A) + rank(B) \geq rank(A \cup B)$ for any A, B, we have equality in the above expression, and so $(U_1 \setminus \{e\}, U_2)$ is a separation for U; contradiction. So U' is an inseparable flat such that for all $F \in S$, $rank(F \cap U) = rank(U')$.

But we assumed that Condition 3 is false for S; contradiction.

We have seen that for any stationary set $S \subseteq \mathcal{S}$, if S satisfies none of Conditions 1-3, then neither does stationary set S_e^- for some $e \in E$. We can repeat the same process until we have a matroid reaction network defined over the ground set $E = \{d\}$ for a single $d \in E$. Now any stationary set S of this reaction network either satisfies Condition 1 (it contains both $\{d\}$ and $\emptyset$, so $S = \mathcal{S}$), Condition 3 (d appears in no set in S), or Condition 2 (S contains $\{d\}$ only, and for inseparable flat $U = \{d\}$ and for every $F \in S$, $rank(F \cap U) = 1 = rank(U)$). Contradiction.

So at least one of Conditions 1-3 is satisfied for any stationary S, concluding the proof of Theorem 3.26. □

Remark 3.29. Since Theorem 3.26 is just a restatement of Theorem 3.23, we have also proved the latter.

Corollary 3.30. *A stationary population p of $\beta^{\mathcal{M}}$ has full support if and only if $\iota(p)$ lies in the interior of P.*

Proof. Noting that conditions 2 and 3 of Theorem 3.26 are identical to the definition of a facet of P, Theorem 3.26 immediately implies that for any stationary p without full support, p lies in a facet of P.

It does not take much extra work to show the converse is also true. First note that if p has full support, then in particular each independent set consisting of a single element has nonzero concentration; so, using q_e as defined previously, $q_e > 0$ for all $e \in E$. So Condition 1 of Corollary 3.25 is not satisfied. Similarly, when p has full support, Lemma 3.28 showed that there is some $e \in E$ such that for any $G \in S_e^+$, $G' = (G \setminus \{e\}) \in S$. Consider the set $U = span(e)$. It is simple to check that U is an inseparable flat. But $rank(G \cap U) > 0 = rank(G' \cap U)$. So Condition 2 of Corollary 3.25 does not hold either, and so for any stationary population p with full support, $iota(p)$ does not lie in a facet of P. □

Theorem 3.31. *The matroid reaction network $\beta^{\mathcal{M}}$ with any (full-support) initial condition is persistent.*

Proof. First note that for any $x \in \mathbb{R}^{|\mathcal{S}|}$ with full support (whether or not it is stationary), Conditions 1 and 2 of Corollary 3.25 cannot hold. So in particular, $\iota(x(t_0))$ is not in a facet of P. Furthermore, the entire trajectory of x is mapped to a single point in P, and this point is not in a facet of P if and only if the stationary population has full support. So if $x(t_0)$ has full support, the stationary population reached from this initial point has full support. □

3.5.2 Generating Full-Support Distributions

Traditionally, reaction network theory has been primarily concerned with proving persistence and convergence results for systems with an initial condition in the interior of the state space, as we have done in the preceding sections. We ask an additional question here. Suppose a system, such as a matroid reaction network, has an initial condition $x(t_0)$ that does not have full support, i.e., $x_I(t_0) = 0$ for some $I \in \mathcal{S}$. Under what conditions does this system still converge to an equilibrium distribution with full support? This question is of particular interest in applications to sampling combinatorial objects, such as independent sets or bases of a matroid, or matchings in a graph, a question studied in much recent work, including Anari, Oveis Gharan, and Vinzant [5] in the case of matroid bases. In this setting, it is desirable to create a process which does not require an initial distribution supported on all the target structures, but rather uses the reactions to generate them from a small but sufficiently rich initial support. Knowing that such a system converges to a full-support distribution – and often, as in the case of detailed balanced reaction networks, a distribution which is easily determined from the initial conditions – provides a mechanism to sample from that distribution.

The polytope method for proving the results on matroid reaction networks neatly provides a result in this vein. That is, it provides a condition under which initial distributions without full support are guaranteed to have full-support limit points and, if complex balanced, are guaranteed to converge to the unique full-support complex balanced equilibrium consistent

with that initial condition. Because all stationary points without full support map to points in the facets of the matroid polytope P, it is sufficient to choose an initial condition which is not in a facet of P.

Corollary 3.32. *For a matroid reaction network $\beta_{\mathcal{M}}$ with any initial condition $x(t_0)$, the image of which under the mapping ι does not lie in a facet of the matroid polytope P, $\beta_{\mathcal{M}}$ is persistent and, if the system is complex balanced, $x(t)$ converges to the unique positive complex balanced equilibrium in its stoichiometric compatibility class.*

Recalling that $x(t_0)$ is in a facet of P exactly when either $q_e = 0$ for some $e \in E$ or $q(U) = rank(U)$ for some $U \subseteq E$ a nonempty inseparable flat of $\mathcal{M}$, it is now possible to check in polynomial time whether any given initial condition necessarily generates a full-support equilibrium.

We note that, while the flats of a matroid can be listed in time polynomial in the output size, the total number of flats may not be polynomial in $|E|$ [70]. However, assuming the existence of an oracle for matroid independence or rank, there are nonetheless polynomial time algorithms for solving polytope membership and most violated inequality problems on submodular functions, of which the matroid rank function is an example; see, for example, [34] and [72].

For additional intuition, we note that certain simpler conditions on the invariant vector q are also sufficient to guarantee that $x(t_0)$ is on the interior of P. For example, it is sufficient for $q = \iota(x(t_0))$ to satisfy both $q_e \neq 0$ for all $e \in E$ and $x_\emptyset(t_0) + \sum_{e\in E} q_e = \sum_{I\in\mathcal{S}} x_I(t_0) = 1$.

To see this, suppose $x(t_0)$ has support $S := \{I \in \mathcal{S} \mid |I| \leq 1\}$. It is immediate that this state satisfies both of the stated conditions. Furthermore, for any $I = \{e_1, \ldots, e_k\} \in \mathcal{S} \setminus S$, there exists a series of reactions generating I starting from only species in S, as follows.

$$\begin{aligned} \{e_1\} + \{e_2\} &\to \{e_1, e_2\} + \emptyset \\ \{e_1, e_2\} + \{e_3\} &\to \{e_1, e_2, e_3\} + \emptyset \\ &\vdots \\ \{e_1, \ldots, e_{k-1}\} + \{e_k\} &\to I + \emptyset \end{aligned}$$

With an appropriate choice of sufficiently small coefficients $\alpha_{y,y'}$ for each successive reaction, there exists a linear combination of reaction vectors $r := \sum_{(y,y')\in\mathcal{R}} \alpha_{y,y'}(y' - y)$ such that $x(t_0) + r \in \mathbb{R}^N_{>0}$. This tells us that there exists some $x' := x(t_0) + r$ in the simplex with full support and with invariant vector $\iota(x') = q$; so by Corollary 3.25, q is not in a facet of P. Thus $\iota(x(t_0)) = q$ is likewise not in a facet of P.

3.5.3 Other Polytopes

We note here other combinatorial reaction networks on which the invariants polytope is well-defined, which suggest future work to which the polytope method could apply.

Matroid Intersection Reaction Networks

For two matroids $\mathcal{M}_1, \mathcal{M}_2$ on the same set E, with independent sets $\mathcal{S}_1, \mathcal{S}_2$ and rank functions r_1, r_2 respectively, let $\mathcal{S}$ be the independent sets in $\mathcal{S}_1 \cap \mathcal{S}_2$. The invariants polytope is then equivalent to the matroid intersection polytope, defined as

$$P = \left\{q \in \mathbb{R}^{|E|} \mid q(U) \leq \min(r_1(U), r_2(U)) \ \forall\ U \subseteq E \text{ and } q_e \geq 0 \ \forall\ e \in E\right\} = conv(\mathcal{S})$$

Matroid Bases Reaction Networks

For a matroid $\mathcal{M}$ on E, let $\mathcal{S}$ be the set of bases of $\mathcal{M}$. The invariants polytope is given by the matroid basis polytope P_B, defined as

$$P_B = \left\{q \in \mathbb{R}^{|E|} \mid q(U) \leq rank(U) \ \forall\ U \subseteq E,\ q(E) = rank(\mathcal{M}), \text{ and } q_e \geq 0 \ \forall e \in E\right\} = conv(\mathcal{S})$$

Thus, for any x in a facet of P_B, either $q_e = 0$ for some $e \in E$ (in which case x is trivially incompatible with any full-support initial state), or $q(U^*) = rank(U^*)$ for some $U^* \subseteq E$. Assuming the latter, we know that for any basis $I \in \mathcal{S}$ such that $x_I > 0$, $|I \cap U^*| = rank(U^*)$. In order to get persistence for the matroid basis reaction network, then, it would be sufficient to show that, for some other $I' \in \mathcal{S}$ (with I' not in the support of x), $|I' \cap U^*| < f(U^*)$.

Polymatroid Reaction Networks

Given a submodular $f : 2^E \to \mathbb{R}_+$, the associated polytope is known as a polymatroid, and is defined as

$$P_f = \left\{q \in \mathbb{R}_+^E \mid q(U) \leq f(U) \ \forall\ U \subseteq E\right\}$$

The polymatroid is a generalized permutahedron, translated to have a vertex at the origin. Note that a permutahedron is the convex hull of all permutations of the coordinates of a single vector in $\mathbb{R}^E$; the generalized permutahedron can have some deformations applied to these vertices, while preserving edge direction and orientation. If P_f is the convex hull of some vectors $\mathcal{S}$, we can define a corresponding simplicial reaction network on $\mathcal{S} \times \mathcal{S}$.

3.6 Persistence of Matchings using Invariants

One natural and important example of matroid intersection is bipartite matching. The persistence of the matchings reaction network was already shown via a polytope method in an unpublished manuscript of Rabinovich, Sinclair, and Wigderson [76]; this section shows how to construct a set of invariants which witness that fact, using our new understanding of the two complementary proof paths.

In [76] it is shown that for any stationary support S corresponding to a limit point $\vec{z}$ of β_{match}, at least one of the following is the case:

1. For some $e \in E$, $e \notin I$ for all $I \in S$

2. For some vertex v, v is in some edge of every matching I in S.

3. There exists some odd subset $U \subseteq V$ such that every matching in S has $\frac{|U|-1}{2}$ edges with both endpoints in U

4. $S = \mathcal{S}$

For each of conditions 1-3, we can provide an invariant showing that for any fully-supported initial condition, S does not satisfy the condition.

1. q_e where $e \notin I$ for all $I \in S$

2. $\sum_{e \in A} q_e$ where A is the set of edges adjacent to v

3. $\sum_{e \in E_U} q_e$ where E_U is the set of edges with both vertices in U

In case 1, for any $I \in \mathcal{S}$ containing e, $x_I(t_0) > 0$, and so $q_e(x(t_0)) > 0$. Yet $q_e(z) = 0$, a contradiction.

Note that cases 2 and 3 both correspond to finding a nonempty set U such that for all $J \in S$, $|U \cap J| = \max_{I \in \mathcal{S}} |U \cap I|$. Then we have

$$\sum_{u \in U} q_u(z) = \sum_{I \in \mathcal{S}} x_I(z) \cdot |U \cap I| = \max_{I \in \mathcal{S}} |U \cap I|$$

while

$$\sum_{u \in U} q_u(x(t_0)) = \sum_{I \in \mathcal{S}} x_I(t_0) \cdot |U \cap I| < \max_{I \in \mathcal{S}} |U \cap I|$$

This again yields a contradiction, since $\sum_{u \in U} q_u$ is invariant. So case 4 must hold: $S = \mathcal{S}$. Given that there is always some limit point of β_{match} with stationary support, then β_{match} is persistent. □

3.7 Persistence of Spanning Tree Reaction Networks

A related class of networks modifies a matroid reaction network to only include those species which are maximal with respect to set inclusion. The species set, then, is exactly the set of *bases* of the matroid. Despite the lack of downward closure on the species set, we find that such networks are persistent in the case of graphic matroids, in which the matroid bases are spanning trees in a graph G.

Definition 3.7 (Spanning tree reaction network)**.** Define a *spanning tree reaction network* for a graph G to be the network with species set $\mathcal{S} := \{\text{all spanning trees of } G\}$, complex set $\mathcal{C} := \mathcal{S} \times \mathcal{S}$, and with $\beta_{y,y'} > 0$ for $y, y' \in \mathcal{C}$ iff $\biguplus_{I \in y} I = \biguplus_{J \in y'} J$.

Theorem 3.33. *Every spanning tree reaction network is persistent and, if complex balanced, satisfies the global attractor condition.*

Remark 3.34. It is sufficient to prove Theorem 3.33 in the setting where $\mathcal{M}$ is a connected matroid; that is, not the direct sum of any other two matroids. (If $\mathcal{M}$ is in fact the direct sum of two matroids $\mathcal{M}_1$ and $\mathcal{M}_2$, it can be analyzed as two separate matroid reaction networks $\beta_{\mathcal{M}_1}$ and $\beta_{\mathcal{M}_2}$ instead.) For a graphic matroid, this is equivalent to the underlying graph G being 2-connected; that is, having at least two distinct paths between any pair of vertices.

To prove Theorem 3.33, we consider the possible stationary supports T for limit points of a spanning tree reaction network with basis set $\mathcal{B}$, generated by graph $G = \{V, E\}$. We consider three conditions which this support T might satisfy:

1. First facet condition: Some edge $e \in E$ satisfies $e \notin I$ for all $I \in T$

2. Second facet condition: There exists some $U \subset E$ such that $|I \cap U| = |V(U)| - 1$ for all $I \in T$. We say that U is *saturated* for all $I \in T$ when this condition holds.

3. Full support: $T = \mathcal{B}$

Our goal will be to show that these three conditions are exhaustive; that is, if the facet conditions (1) and (2) are not met for a given stationary support T, then T must have full support.

We begin with a few lemmas.

Lemma 3.35. *Suppose stationary support T has neither of the above two facet conditions, and fix $g = (g_1, g_2) \in E$. Then there exists a tree $L_g \in T$ with g_1 as a leaf.*

Proof. Note that because T has no facet condition, some tree $I \in T$ exists with $g \in I$. Removing edge g from I generates two connected components; let U_I be the component containing vertex g_1. Select $I \in T$ to be a spanning tree with $g \in I$ which minimizes $|V(U_I)|$.

If $|V(U_I)| = 1$ then g_1 is a leaf in I and we are done. So suppose $|V(U_I)| \geq 2$. Because T has no facet condition, there exists some $J \in T$ with $V(U_I)$ not saturated in J. Then the following sequence of reactions (which exist by the basis exchange property) reduce the

difference between $I \cap U_I$ and $J \cap U_I$ until U_I is no longer saturated in I:

$$I + J \leftrightarrow \underbrace{(I \setminus i_1 \cup j_1)}_{I_2} + \underbrace{(J \setminus j_1 \cup i_1)} \quad (\text{with } i_1 \in E(U_I) \cap (I \setminus J))$$

$$I_2 + J \leftrightarrow \underbrace{(I_2 \setminus i_2 \cup j_2)}_{I_2} + \underbrace{(J \setminus j_2 \cup i_2)} \quad (\text{with } i_2 \in E(U_I) \cap (I_2 \setminus J))$$

$$\vdots$$

Continuing analogously creates $I_2, \ldots, I_k$ (each in T by stationarity of the support) until the first reaction for which $j_k \notin E(U_I)$; since each reaction strictly decreases $|(U_I \cap I_k) \setminus (U_I \cap J)|$, and $|U_I \cap J| < |U_I \cap I|$, this must happen at some step k. Removing edge g from I_k yields two connected components, one of which (call it U_{I_k}) contains g_1 and has vertex set $V(U_{I_k}) \subsetneq V(U_I)$. This contradicts the earlier assumption that $|V(U_I|$ was minimal among all $I \in T$ with $g \in I$.

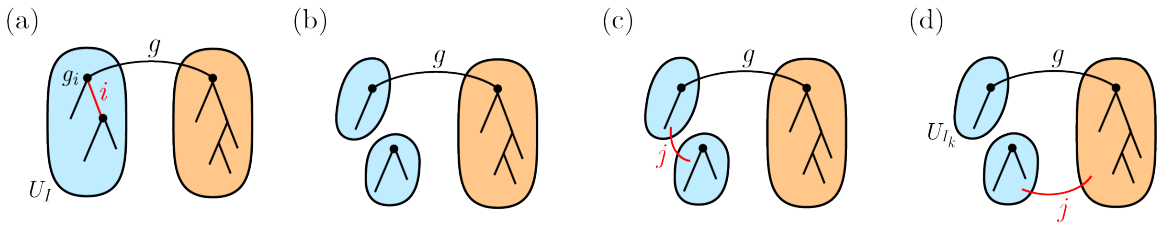


Figure 3.1: Illustration of the procedure removing an edge i from U_I and replacing it with an edge j from J in the proof of Lemma 3.35. (a) shows the selection of edge i from U_I, (b) shows the removal of that edge creating two connected components, and (c) and (d) depict the two possible cases for the new edge j: either $j \in E(U_I)$, and the procedure continues to the next iteration, or $j \notin E(U_I)$, leaving $V(U_{I_k}) \subsetneq V(U_I)$

We conclude that $|V(U_I)| = 1$; in other words, $V(U_I) = \{g_1\}$, and g_1 is a leaf in I. □

We see in the next lemma that, in fact, a stationary support T with no facet conditions must contain an even larger family of trees.

Lemma 3.36. *Suppose stationary support T has neither of the two facet conditions. For any path $P = \{v_1 \ldots v_n\}$ in $I \in T$, if there exists an edge $(v_1, v_n) \in E(G)$ completing the cycle, then for any $v_i, v_{i+1} \in P$, we also have $(I \cup (v_1, v_n) \setminus (v_i, v_{i+1})) \in T$.*

Proof. We prove this via repeated reactions modifying the initial tree I, by adding leaf edges from trees guaranteed to exist by Lemma 3.35. For any edge $(u, v) \in E(G)$, let $L_{u,v}$ represent a tree in T containing this edge, such that u is a leaf. The following sequence of reactions exchanges edges around the cycle $P \cup (v_1, v_n)$:

$$I + L_{v_1,v_n} \leftrightarrow \underbrace{[I \cup (v_1, v_n) \setminus (v_1, v_2)]}_{I_2} + [L_{v_1,v_n} \setminus (v_1, v_n) \cup (v_1, v_2)]$$
$$I_2 + L_{v_2,v_1} \leftrightarrow \underbrace{[I_2 \cup (v_1, v_2) \setminus (v_2, v_3)]}_{I_3} + [L_{v_2,v_1} \setminus (v_1, v_2) \cup (v_2, v_3)]$$
$$\vdots$$

Observing that, since I and all the L_{v_i,v_j} are in T, the same holds for all the I_j, so it follows in particular that

$$I_{i+1} = (I \cup (v_1, v_n) \setminus (v_i, v_{i+1})) \in T.$$

□

Lemma 3.36 tells us that, when T has no facet condition, for any cycle created by adding a single edge d to $I \in T$, removing any other edge e in that cycle leaves $(I \setminus d \cup e) \in T$. We finish the proof of Theorem 3.33 by observing that this property is strong enough to conclude that T in fact contains all spanning trees of G.

Proof of Theorem 3.33. We first observe that facet conditions (1) and (2) are incompatible with any full-support initial condition by a straightforward invariant argument, analogous to those in the proof of Theorem 3.15.

For the first facet condition, $q_e(x) := \sum_{I \ni e} x_I$ is invariant for each $e \in E$, and $q_e(x) > 0$ for all $e \in E$ for any initial condition with full support. Similarly, $\sum_{u \in U} q_U(x) = \sum_{I \in S} |U \cap I| \cdot x_I(t)$ is invariant, and if the second facet condition held, we would have $\sum_{u \in U} q_U(x) = |V(U)| - 1$. Given an initial condition with full support, this can only occur if $|U \cap I| = |V(U)| - 1$ for every $I \in \mathcal{B}$. However, because $\mathcal{B}$ is the basis set of a connected matroid, there exists no such $U \subset E$.

Suppose next that $T \neq \mathcal{B}$ is a stationary support with neither of the two facet conditions. Consider any pair of spanning trees with $J \notin T$, $I \in T$. For any $j \in J \setminus I$, $(I \cup j)$ contains some cycle C, and furthermore there is some $i \in C$ with $i \notin J$. Then $I_2 := (I \cup j_1 \setminus i) \in T$ by Lemma 3.36, and $|I_2 \triangle J| < |I \triangle J|$. Repeating this procedure generates $I_3, \ldots, I_k \in T$ until, for some k, $|I_k \triangle J| = 0$ and $I_k = J$. Thus $I_k = J \in T$ for some k, contradicting the initial assumption that there exists some $J \in \mathcal{B} \setminus T$.

Therefore we conclude that the only stationary support admitted by a full-support starting condition is $\mathcal{B}$. □

Convergence properties are not yet known for the more general class of matroid basis reaction networks, but we conjecture that the persistence and global attractor properties hold in this additional generality as well. In particular, networks based on *binary* matroids provide a natural next step for future work, as the circuits and co-circuits in these networks provide a natural analog to the partitions in the graphic matroid used in the above proof.

Chapter 4

Hardness of Minimum Rank 3

The most natural thing in the world is complexity.

- Hank Green, *Vlogbrothers*

This chapter is based on the paper in preparation "*Minimum Rank 3 of Graphs is Complete for the Existential Theory of the Reals*", which is joint work with Kevin Grace, H. Tracy Hall, and Alathea Jensen.

4.1 Introduction

Consider the problem of recovering the adjacency matrix of a graph with a known set of edges, but unknown edge weights. There are infinitely many symmetric matrices which could fit the required sparsity structure – yet some may provide a "simpler" explanation for the observed edge set than others. Suppose, for example, the graph describes a social network, with edges representing an acquaintance between two individuals. If we were to observe an edge (u, v), with individuals u and v sharing a large fraction of their neighbors, a relatively high edge weight (indicating a close relationship) could be considered a simpler explanation for the correlation than if the two were distant acquaintances.

This problem of finding a set of edge weights which best explain an observed sparsity structure is a common goal of matrix completion problems, including the well-studied "Netflix problem" [21, 61]. In particular, when the true adjacency matrix is known to be sufficiently low-rank, it is often possible to reconstruct a matrix exactly from a sparse set of observations [26]. In this setting, the rank of the output matrix can be seen as a measure of the complexity of a given data set – the lower the rank, the "simpler" the explanation for the observed structure. Yet the general problem of finding the lowest rank matrix fitting a given sparsity structure is known to be NP-hard, and all known algorithms require exponential time [29, 78]. We consider, instead, the following sub-problem: without necessarily finding the matrix, determine the minimum achievable rank for a given sparsity structure. This is the objective of the *minimum rank problem*, and the subject of the present chapter.

4.1.1 Overview

Concretely, the minimum rank problem for a graph takes as input a simple graph $G = (V, E)$, whose vertices are identified with the index set $\{1, \ldots, |V|\}$, and asks for the lowest possible rank among a class of $|V| \times |V|$ matrices related to the graph by their pattern of nonzero off-diagonal entries. Minimum rank is a subproblem of the more difficult *Inverse Eigenvalue Problem for a Graph* (*IEP-G*), which asks for complete knowledge of spectral restrictions arising from the sparsity pattern of G, but already the subproblem exhibits a richness of difficulty that has led to an extensive literature, as surveyed, for example, in the recent monograph by Hogben, Lin, and Shader [56]. In the most-studied case, which will be assumed unless otherwise specified, the matrix is taken to be real symmetric, but the question can be asked for matrices over any field $\mathbb{F}$. When $\mathbb{F}$ is a subfield of the complex numbers $\mathbb{C}$, then the matrix is taken to be Hermitian (which includes real symmetric matrices as a special case). Otherwise the matrix is taken to be symmetric over $\mathbb{F}$. Precise definitions are given in Section 4.2.

The purpose of this chapter is to precisely quantify the difficulty of the minimum rank problem, showing in particular that over real symmetric matrices it is equivalent to a complexity class known as the *existential theory of the reals*, abbreviated $\exists\mathbb{R}$. We will consider two classes of problems. On the one hand, in the context of graph theory (or more particularly combinatorial matrix theory), we are given an arbitrary simple graph G and the problem asks whether the minimum rank of G over some infinite field $\mathbb{F}$ is equal to 3. On the other hand, in the context of algebraic geometry (and especially real algebraic geometry, when $\mathbb{F} = \mathbb{R}$), we are given an arbitrary system of multivariable polynomial equations with integer coefficients, and the problem asks whether the equations have a simultaneous solution over $\mathbb{F}$.

We will show that these two problem classes are equivalent in difficulty. More specifically, we show how to transform an arbitrary instance of either problem to an instance of the other with only a polynomial overhead. The second problem class, with $\mathbb{F}$ taken to be $\mathbb{R}$, is known to be complete for $\exists\mathbb{R}$, as shown in [82], and so the equivalence of the two sides over $\mathbb{R}$ gives the result claimed in the title. The minimum rank problem thus joins the ranks of problems which are complete for $\exists\mathbb{R}$, which is known to be in PSPACE and NP-hard, and for which other known hard problems include training neural networks [1], recognizing intersection graphs of convex sets in the plane [17], and the algorithmic Steinitz problem [81]. We additionally show in Section 4.6 that, while a finite list of forbidden induced subgraphs is sufficient to characterize those graphs with minimum rank less than or equal to 2, and similarly some such list exists for any minium rank d over finite fields, no such finite list is sufficient to characterize graphs of minimum rank 3 over the reals or over any infinite field.

4.2 Preliminaries

For a Hermitian $n \times n$ matrix A over $\mathbb{F}$, let $\mathcal{G}(A)$ be the graph with vertex set $\{1, 2, \ldots, n\}$ and edge set $\{\{i, j\} : i \neq j \text{ and } a_{ij} \neq 0\}$. For a graph G, let $\mathcal{S}(G)$ consist of all real symmetric matrices A such that $G = \mathcal{G}(A)$.

Definition 4.1 (Minimum rank)**.** The *minimum rank* of a graph G, denoted by $\mathrm{mr}(G)$, is the smallest integer k such that there is a matrix of rank k in $\mathcal{S}(G)$. The minimum rank of G over $\mathbb{F}$ for a field other than the reals is defined analogously.

In order to situate these problems within complexity theory, we formulate them as decision problems as detailed below, letting d represent the target rank, or dimension of the column space of the matrix.

Decision problem: RANK

- Input: a target rank d, given as a binary-encoded nonnegative integer, together with a simple graph $G = (V, E)$, given as the $|V|(|V| - 1)/2$ bits that specify its adjacency matrix.

- Output: YES if and only if there exists a real symmetric matrix of size $|V| \times |V|$, the positions of whose nonzero off-diagonal entries correspond exactly to the edges of G, and whose rank is at most d.

Decision problem: $\mathbb{F}$-RANK (for a particular field $\mathbb{F}$)

- Input: a target rank d together with the adjacency matrix of a simple graph $G = (V, E)$.

- Output: YES if and only if there exists a conforming matrix whose rank is at most d. Regardless of $\mathbb{F}$, a conforming matrix must have entries in $\mathbb{F}$ and must have its nonzero off-diagonal entries given exactly by the edges of G. In the case that $\mathbb{F}$ is a subfield of the complex numbers, a conforming matrix must be Hermitian. (In particular, when $\mathbb{F}$ is a subfield of the reals, a conforming matrix must be symmetric.) For all other fields, a conforming matrix must be symmetric.

Decision problem: d-RANK (for a particular non-negative dimension d)

- Input: a simple graph $G = (V, E)$, given as an adjacency matrix.

- Output: YES if and only if the answer to RANK is YES for graph G and target rank d.

Decision problem: $\mathbb{F}$-d-RANK (for a particular field and dimension)

- Input: a simple graph $G = (V, E)$, given as an adjacency matrix.

- Output: YES if and only if the answer to $\mathbb{F}$-RANK is YES for graph G and target rank d.

It is common in graph theory to denote the number of vertices $|V|$ by n, but here we instead follow the convention of letting n denote size of the problem—concretely, the number of bits required to specify an instance of a particular decision problem. Since an adjacency matrix requiring $|V|(|V|-1)/2$ bits makes up the bulk of the input to each of the decision problems defined above, $|V|$ scales as $\sqrt{n}$, meaning for example that an algorithm running in time $O(|V|^4)$ would be a solution of time complexity $O(n^2)$.

4.3 Outline of Main Results

Over any infinite field $\mathbb{F}$, the decision problems $\mathbb{F}$-0-RANK, $\mathbb{F}$-1-RANK, and $\mathbb{F}$-2-RANK can each be solved in time that is polynomial in the size of the input, because for each of those cases there is a known finite list of forbidden induced subgraphs [14]. In particular, where $\mathbb{F}$ is $\mathbb{R}$, the decision problems 0-RANK, 1-RANK, and 2-RANK have only polynomial complexity. For 3-RANK, however, and more generally for $\mathbb{F}$-3-RANK over any infinite field $\mathbb{F}$, we show in Section 4.6 that no such list of forbidden induced subgraphs exists.

Theorem 4.1. *There is no finite list of forbidden induced subgraphs for minimum rank 3.*

Theorem 4.2. *Given two distinct number fields K and L, there exists a graph G such that the Hermitian minimum rank of G over K is 3 but the Hermitian minimum rank of G over L is strictly greater than 3.*

This result hints that something may be fundamentally more difficult about d-RANK for $d \geq 3$ (and therefore more difficult for RANK overall). We make this observation concrete in Section 4.5 by showing that, over the reals, the complexity class of 3-RANK is equivalent to the existential theory of the reals, or $\exists\,\mathbb{R}$. $\exists\,\mathbb{R}$ is the problem of deciding whether a given polynomial system (in multiple variables with integer coefficients) has a common solution over the reals; that is, the set of all true sentences of the form

$$\exists X_1 \cdots \exists X_n F(X_1, \cdots, X_n)$$

where the X_i are variables with real number values, and F is a quantifier-free polynomial system. The corresponding decision problem is known to be NP-hard and in PSPACE [27].

The primary result of this paper provides a polynomial time reduction from $\exists\,\mathbb{R}$ to 3-RANK which, along with the (trivial) reverse direction, yields the following theorems (where $\equiv$ denotes polynomial time equivalence).

Theorem 4.3. 3-RANK $\equiv \exists\,\mathbb{R}$

Corollary 4.4. RANK $\equiv \exists\,\mathbb{R}$

Furthermore, we show that the same results extend analogously for $\mathbb{F}$-3-RANK over any infinite field $\mathbb{F}$.

4.3.1 Proof Roadmap

This section provides an overview of the proof of our main result, Theorem 4.3. One direction of the equivalence is straightforward to show. A symmetric matrix A of rank at most d has a rank factorization $A = M^T DM$ for which D is a diagonal matrix of size $d \times d$, and by letting $\mathbf{x}$ be a collection of indeterminates covering the diagonal entries of D and all entries of M, the existence of such a matrix A is reduced to a system of integer-coefficient equations $p_i(\mathbf{x}) = 0$ (coming from the non-edges of G) and non-equations $p_i(\mathbf{x}) \neq 0$ (coming from the edges of G). The existential theory of the reals is defined in terms only of solving equations, but it is well understood how to transform a question involving also non-equations and inequalities into a system of equations; for example, each non-equation $p_i(\mathbf{x}) \neq 0$ can be handled by introducing a new, otherwise unused variable y_i and requiring $y_i p_i(\mathbf{x}) = 1$.

The more difficult direction, which requires the construction of some machinery, is reducing an arbitrary system of polynomial equations S into a particular minimum rank 3 instance. The geometric idea employed is an equivalence between minimum rank 3 for a certain class of graphs (complements of bipartite graphs, as described in Section 4.4) and the existence of a certain point-line incidence structure within the projective plane over $\mathbb{F}$ (or within an affine representation of it, which over $\mathbb{R}$ is the familiar Cartesian plane).

We make the simplifying assumption that the polynomials in the system of equations S consist only of elementary additions of the form $q_i = q_j + q_k$ and elementary multiplications of the form $q_i = q_j \cdot q_k$. For systems which do not have this form, a pre-processing step to transform an arbitrary system of equations S into a compliant system T over an expanded set of variables $Q = \{q_i\}$ is provided in the appendix (Section 4.7.1). This step decomposes each of the full polynomial equations into a collection of elementary equations, each one involving a single binary operation of addition or multiplication, including building up constants for each positive or negative integer coefficient according to the binary representation of its absolute value, starting from the single constant 1. From here, the complete set of elementary addition and multiplication equations must be encoded into projective geometry, as described in Section 4.5. The gadgets employed for this reduction include:

- a small base structure of points and lines;
- for each q_i, a set of four associated lines and three associated points;
- for each elementary addition $q_i = q_j + q_k$, a single point that is incident to three particular lines associated respectively to q_i, q_j, and q_k; and similarly
- for each elementary multiplication $q_i = q_j \cdot q_k$, a single point that is incident to three particular lines associated respectively to q_i, q_j, and q_k.

The original basis for the geometric technique is attributed to Marshall Hall [53, 54].

These collections of point-line incidence relations in the projective plane in turn become the edges of a bipartite graph, one side of the partition representing points and the other side representing lines. A rank-3 realization of the complement of this bipartite graph will have an entry equal to zero corresponding to each such point-line incidence, implying that the

individual additions and multiplications are consistent on the expanded set of variables, which implies further that the original system of polynomial equations has a simultaneous solution. That is, the procedure as described above produces a graph G from a set of equations S with the following properties:

1. As explained in Section 4.5.5, every rank-3 representation of G over $\mathbb{F}$ gives a concrete set of numbers in $\mathbb{F}$ that satisfies the equations in S.
2. As explained in Section 4.5.4, every solution to S over $\mathbb{F}$ produces a concrete set of points and lines that are incident in $\mathbb{FP}^2$ wherever required by the construction. The matrix of inner products between points and lines thus has a zero entry wherever the lack of an edge in G requires it.

However, this result does not yet suffice to construct the desired graph G_S. The implication that every rank-3 realization of the complement of the bipartite graph yields a solution to the system S is only one direction of the desired equivalence. The reverse implication will also hold at least in part: Any solution to the system S will indeed produce a symmetric matrix of rank 3 with a zero entry for each of the prescribed point-line incidences. The difficulty is that minimum rank 3 not only requires certain equations to be satisfied (where the lack of an edge in the graph produces a 0 in the matrix), but also requires many equations not to be satisfied (where an edge in the graph must correspond to a nonzero entry in the matrix).

To address this possibility, we add a second pre-processing pass, described in detail in Section 4.7.2, to guarantee that for all indices $i \neq j$, we have $q_i \neq 0$, $q_i \neq q_j$, and $q_i + q_j \neq 1$. We show that these additional conditions ensure that no solutions to the polynomials produce, when translated to the projective geometry, additional point-line incidences that are *not* required by the construction, and therefore the overall construction results in a matrix no sparser than the pattern required by G.

The result is a complete, non-provisional equivalence between solutions to the original polynomial equations and rank-3 representations of an exactly specified larger graph G'. In the other direction, any rank-3 representation of G' yields a concrete set of values both for the variables in the original equations and for all of the slack variables.

4.4 Rectangular Minimum Rank and Bipartite Graph Complements

The decision problems of Section 4.2 concern the minimum rank of symmetric (or Hermitian) matrices with a pattern governed by a simple graph. The graphs that will be produced from any system of equations will all take the special form of the complement of a bipartite graph, for which the minimum rank problem reduces to the simpler problem of rectangular minimum rank.

Definition 4.2. A *rectangular pattern* $Y = [y_{ij}]$ is a matrix with entries from the set $\{0, *\}$.

Definition 4.3. A pattern Y is said to be *reduced* if every row of Y contains at least one $*$ entry, and every column of Y contains at least one $*$ entry.

Definition 4.4. For a rectangular pattern Y and a given field $\mathbb{F}$, the *minimum rank* of Y over $\mathbb{F}$, denoted $\mathrm{mr}_{\mathbb{F}}(Y)$, or $\mathrm{mr}(Y)$ in the case where $\mathbb{F} = \mathbb{R}$, is the smallest possible rank of a matrix $A = [a_{ij}]$ over $\mathbb{F}$ with the same dimensions as Y such that $a_{ij} = 0$ if and only if $y_{ij} = 0$.

Note that any rectangular minimum rank problem can be reduced to $\mathrm{mr}_{\mathbb{F}}(Y)$ for Y a reduced pattern, because deleting a row or column of all zeros does not change the rank of a matrix.

Suppose that G is a simple graph whose complement is bipartite with bipartition $V(G^c) = R \sqcup C = V(G)$, with R called the *row vertices* and C called the *column vertices* of G. Order the vertices of G so that row vertices precede column vertices, let $\mathbb{F}$ be an infinite field, and let A be a matrix over $\mathbb{F}$ that is symmetric (or Hermitian when $\mathbb{F}$ is a subfield of $\mathbb{C}$), whose pattern of off-diagonal nonzero entries is given exactly by G and whose diagonal entries are nonzero. Then A is partitioned naturally as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix} \text{ with pattern } \begin{bmatrix} * & Y \\ Y^T & * \end{bmatrix}$$

for some rectangular pattern Y. We call the pattern Y the *rectangular part* of G. The rectangular part of G is reduced if and only if every row vertex of G is adjacent to at least one column vertex of G and every column vertex of G is adjacent to at least one row vertex of G. Equivalently, the rectangular part of G is reduced if and only if, in the bipartite graph G^c, no row vertex dominates the set of column vertices and no column vertex dominates the set of row vertices.

Remark 4.5. A non-zero entry is generic, but each zero in the rectangular part of G imposes a constraint. These constraints come from the non-edges of G, or from the edges of the bipartite complement of G.[1]

Theorem 4.6 (Theorem 3.1 from [11])**.** *Let G be a simple graph whose complement is bipartite and whose rectangular part Y is reduced, and let $\mathbb{F}$ be an infinite field. Then the minimum rank of G over $\mathbb{F}$ is equal to the rectangular minimum rank of Y over $\mathbb{F}$.*

The case of particular interest, $\mathrm{mr}(G)$ (i.e., $\mathrm{mr}_{\mathbb{F}}(G)$ for $\mathbb{F} = \mathbb{R}$), is a case in which $\mathbb{F}$ is a subfield of the complex numbers and A is taken to be Hermitian. Over Hermitian matrices, questions about rank can be refined to questions about *inertia*, and there is a stronger version of Theorem 4.6 that takes this into account.

[1] Whether it is edges or non-edges that impose constraints can sometimes be a source of confusion. We find, for example, a typographical error in the publication [64], where Proposition 2.1 (or Proposition 1 in the eprint, version 2) contains the following expression twice: $v\langle i\rangle\, w\langle j\rangle^T = 0$. This should be replaced by " $\neq 0$" (or omitted entirely, in the second instance) to match the statement actually proven and later used.

Definition 4.5. The triple of integers expressing the number of positive, negative, and zero eigenvalues of a complex Hermitian matrix A is called the *inertia* of A.

Definition 4.6. Let G be a simple graph on n vertices, let $\mathbb{F}$ be a subfield of the complex numbers, and let k be the minimum rank of G taken over Hermitian matrices with entries in $\mathbb{F}$. Then there are $k+1$ possible inertias of rank k, from $(0, k, n-k)$ to $(k, 0, n-k)$ inclusive. In the case that every one of these inertias is realized by some Hermitian matrix over $\mathbb{F}$ with pattern G, we say that G is *inertially arbitrary* over $\mathbb{F}$.

The stronger version of Theorem 4.6 that we will prove is the following:

Theorem 4.7. *Let G be a simple graph whose complement is bipartite and whose rectangular part Y is reduced, and let $\mathbb{F}$ be a subfield of $\mathbb{C}$. Then G is inertially arbitrary over $\mathbb{F}$, with minimum rank equal to the rectangular minimum rank of Y over $\mathbb{F}$.*

Proof. We follow the original proof of Theorem 4.6 given in [11, Theorem 3.1], but making allowance for arbitrary inertia. Let $\mathbb{F}$ be a subfield of the complex numbers (hence infinite with $\mathbb{Q} \subseteq \mathbb{F} \subseteq \mathbb{C}$), let d be the minimum rectangular rank of Y, and let $(d-\nu, \nu, |V|-d)$ be the desired inertia.

In the easy direction, the rank of A is at least d because it has A_{12} as a submatrix, whose pattern is Y. Suppose then that a rectangular matrix A_{12} is given that has reduced pattern Y, rank d, and entries in $\mathbb{F}$. It suffices to show that A_{12} (or some nonzero multiple of A_{12}) can be extended to a Hermitian matrix A that

- has the correct pattern G;
- has inertia $(d-\nu, \nu, |V|-d)$; and
- has all entries in $\mathbb{F}$.

There exists a rank decomposition expressing A_{12} as a product of a matrix R with d columns and a matrix C with d rows,

$$A_{12} = RC,$$

where the rows of R are row vectors $r_i \in \mathbb{F}^d$ and the columns of C are column vectors $c_i \in \mathbb{F}^d$. Let D be a $d \times d$ diagonal matrix with $d-\nu$ diagonal entries equal to 1 and ν diagonal entries equal to -1, satisfying $D^2 = I_d$. Our goal is to find a matrix M of size $d \times |V|$ such that

$$A = M^* D M$$

has all the desired properties. We have a candidate $M_0 = \begin{bmatrix} DR^* & C \end{bmatrix}$ that produces a matrix

$$A_0 = M_0^* D M_0 = \begin{bmatrix} RD \\ C^* \end{bmatrix} D \begin{bmatrix} DR^* & C \end{bmatrix} = \begin{bmatrix} RDR^* & A_{12} \\ A_{12}^* & C^*DC \end{bmatrix}$$

that has the correct inertia (by Sylvester's Law of Inertia [90]) and that has the correct pattern Y in the block A_{12}. But A_0 may have unwanted zeros in the off-diagonal entries of its diagonal blocks, whereas the desired pattern G for all of A is the complement of a bipartite graph. Fortunately, enough degrees of freedom are available to make the diagonal blocks of A generically nonzero. For any invertible $d \times d$ matrix Q, we have another rank decomposition

$$A_{12} = (RQ)(Q^{-1}C),$$

that gives us another candidate for M, namely

$$\begin{bmatrix} D(RQ)^* & Q^{-1}C \end{bmatrix} = \begin{bmatrix} DQ^*R & Q^{-1}C \end{bmatrix}.$$

Let the $d \times d$ matrix $Q = [q_{ij}]$ consist entirely of rational-valued indeterminates $\mathbf{q} = (q_{11}, q_{12}, \ldots, q_{dd}) \in \mathbb{Q}^{d^2}$. Since Q is real, $Q^* = Q^T$. The invertibility of Q is equivalent to the nonvanishing of a polynomial

$$p(\mathbf{q}) := \det(Q).$$

Rather than using the inverse Q^{-1} to define the candidate matrix M, we use the adjugate matrix $P = \operatorname{adj}(Q)$, which has entries that are also polynomials in the indeterminates $\mathbf{q}$, and which satisfies

$$PQ = QP = p(\mathbf{q})I_d.$$

The adjusted rank decomposition becomes

$$(RQ)(PC) = p(\mathbf{q})A_{12},$$

giving us the candidate for M:

$$M_{\mathbf{q}} = \begin{bmatrix} D(RQ)^* & PC \end{bmatrix} = \begin{bmatrix} DQ^TR^* & PC \end{bmatrix};$$

and the candidate for A:

$$A_{\mathbf{q}} = M_{\mathbf{q}}^* D M_{\mathbf{q}} = \begin{bmatrix} RQD \\ C^*P^T \end{bmatrix} D \begin{bmatrix} DQ^TR^* & PC \end{bmatrix} = \begin{bmatrix} RQDQ^TR^* & p(\mathbf{q})A_{12} \\ p(\mathbf{q})A_{12}^* & C^*P^TDPC \end{bmatrix},$$

all of whose entries are polynomials in the indeterminates q_{ij}. Whenever $p(\mathbf{q}) \neq 0$, $A_{\mathbf{q}}$ has the correct pattern Y in the off-diagonal block, the correct inertia, and all entries are in $\mathbb{F}$. It remains only to take care of the diagonal blocks—i.e., to establish that for some assignment of the indeterminates q_{ij}, every entry of the matrices

$$A_{11} = RQDQ^TR^* \quad \text{and} \quad A_{22} = C^*P^TDPC$$

is nonzero while Q is nonsingular. Every entry of A_{11} is a polynomial $a_{ij}(\mathbf{q})$ with rational indeterminates but possibly complex coefficients, namely

$$a_{ij}(\mathbf{q}) = r_i Q D Q^T r_j^*$$

for some pair (not necessarily distinct) of row vectors $r_i, r_j \in \mathbb{F}^d$. Similarly, every entry of A_{22} is a polynomial

$$b_{ij}(\mathbf{q}) = c_i^* P^T D P c_j$$

for a pair of column vectors $c_i, c_j \in \mathbb{F}^d$. We now recall the fact that the pattern Y is reduced, which implies that every row vector r_i participates in a nonzero dot product and is therefore not the zero vector, and similarly that no column vector c_i is the zero vector. It follows that none of the polynomials $a_{ij}(\mathbf{q})$ or $b_{ij}(\mathbf{q})$ is identically the zero polynomial[2], as is also the case for the polynomial $p(\mathbf{q})$, which implies that there exists a choice of $\mathbf{q}$ for which all the polynomials evaluate simultaneously to nonzero numbers in $\mathbb{F}$. □

4.5 Construction: Graphs from Equations

In this section, we describe a process for taking any system S of polynomial equations with integer coefficients and constructing a simple graph G_S on which 3-RANK answers YES if and only if the system S has a simultaneous solution over the real numbers. This process runs in polynomial time in the size of S.

Throughout this section, we assume that the polynomial system is *triangulated* in the sense of Section 4.7.1; that is, all equations are either of the form $q_i = q_j + q_k$ or the form $q_i = q_j q_k$ with i, j, k not necessarily distinct. Section 4.7.1 contains a procedure to transform the set of polynomial equations S on variables $\mathbf{x}$ into a set of equations T on variables $\mathbf{q}$ which are triangulated.

4.5.1 Base Graph and Base Matrix

Rather than constructing G_S directly, we instead construct its complement, which we will call G for simplicity's sake. We first form the base of G, as shown in Figure 4.1. The base graph is bipartite, and we use the convention that one side of the bipartition has vertices with upper-case names and the other side has vertices with lower-case names.

We also construct a matrix M that represents the base graph. The rows correspond to upper-case vertices and the columns to lower-case vertices, as shown below. A zero entry appears in row r and column c if and only if the corresponding vertices are adjacent in the graph. M can be factorized into $M = RC$, where R is a 3-column matrix where each row corresponds to an upper-case vertex, and C is a 3-row matrix where each column corresponds

[2]This is trivial for $d = 1$. For $d \geq 2$, it suffices to consider one-parameter families $Q = I_d + tE_{k\ell}$ (giving $P = I_d - tE_{k\ell}$) where $k \neq \ell$ are chosen appropriately depending on where the nonzero vectors r_i and r_j (or c_i and c_j) have nonzero entries.

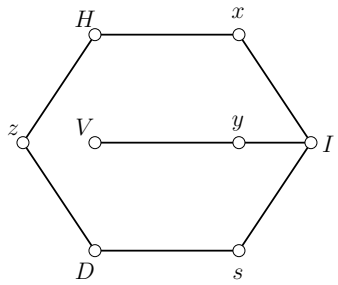


Figure 4.1: The eight vertices of the base graph.

to a lower-case vertex.

$$M = \begin{array}{c} \\ V \\ H \\ I \\ D \end{array}\!\!\begin{array}{c} \begin{array}{cccc} x & y & z & s \end{array} \\ \begin{bmatrix} 1 & 0 & -1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 \end{bmatrix} \end{array} = \begin{array}{c} \\ V \\ H \\ I \\ D \end{array}\!\!\begin{array}{c} \begin{array}{ccc} x & y & z \end{array} \\ \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \end{array} \begin{array}{c} \begin{array}{cccc} x & y & z & s \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

Working over the reals, we may think of each row in R as representing a plane in $\mathbb{R}^3$ or a line in the projective plane. The projective plane interpretation is useful in this context because it only matters whether the rows of R and columns of C have a zero or non-zero dot product, so the scaling of each row or column is unimportant.

More specifically, given a nonzero row $\begin{bmatrix} a & b & c \end{bmatrix}$ in R, we can think of it as the plane $\{(x, y, z) \in \mathbb{R}^3 : ax + by + cz = 0\}$ or as the corresponding line in the projective plane. Likewise, we may think of each column in C as representing the span of a vector in $\mathbb{R}^3$ or a point in the projective plane. More specifically, we think of a nonzero column vector $\begin{bmatrix} a & b & c \end{bmatrix}^\top$ in C as its span, the line $\{(at, bt, ct) \in \mathbb{R}^3 : t \in \mathbb{R}\}$, or as the corresponding point in the projective plane. In this manner, we may think of the graph as representing line-plane incidences in $\mathbb{R}^3$ or point-line incidences in the projective plane.

Figure 4.2 illustrates the eight vertices of the base graph as planes and lines in $\mathbb{R}^3$, which intersect an affine plane P (a copy of the Cartesian plane given by $z = 1$) to give, respectively, lines and points in P or its extension to a projective plane, shown in Figure 4.3. The four planes are as follows:

1. V ("Vertical"): The plane $1x + 0y - 1z = 0$ intersects P in the line $x = 1$.
2. H ("Horizontal"): The plane $0x + 1y + 0z = 0$ intersects P in the line $y = 0$.
3. I ("Infinity"): The plane $0x + 0y + 1z = 0$ is parallel to P and does not intersect it except in the "line at infinity".

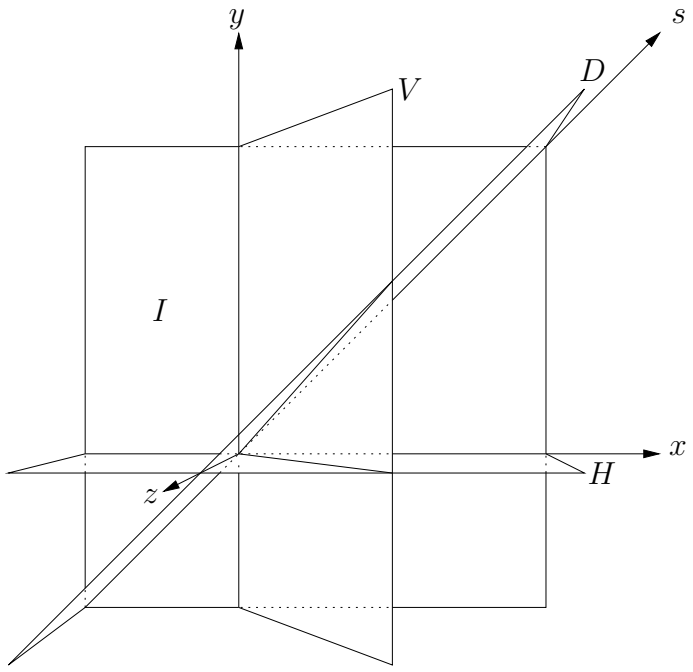


Figure 4.2: The eight vertices of the base graph represented as lines and planes in $\mathbb{R}^3$, also illustrating where they intersect the affine plane $z = 1$ or its extension to a projective plane.

4. D ("Diagonal"): The plane $-1x + 1y + 0z = 0$ intersects P in the line $y = x$.

The four points are as follows:

1. x: The line spanned by $\begin{bmatrix}1 & 0 & 0\end{bmatrix}^T$ does not intersect P except as a "point at infinity" in the x-direction $(1, 0)$.
2. y: The line spanned by $\begin{bmatrix}0 & 1 & 0\end{bmatrix}^T$ does not intersect P except as a "point at infinity" in the y-direction $(0, 1)$.
3. z: The line spanned by $\begin{bmatrix}0 & 0 & 1\end{bmatrix}^T$ intersects P at the origin of P, the point $(0, 0)$.
4. s ("slant"): The line spanned by $\begin{bmatrix}1 & 1 & 0\end{bmatrix}$ does not intersect P except as a "point at infinity" in the slanted direction $(1, 1)$.

4.5.2 Adding Variables

The next step in constructing G involves the equations T on variables $\mathbf{q}$. For each variable q_i in $\mathbf{q}$, we add seven new vertices to graph G, named $Z_i, v_i, X_i, d_i, Y_i, h_i, S_i$. They should be connected to the base graph and to each other in the manner shown in Figure 4.4.

Correspondingly, we also add rows for Z_i, X_i, Y_i, S_i to matrix R and columns for v_i, d_i, h_i to matrix C, in the manner shown below. This will cause the correct zero pattern (matching

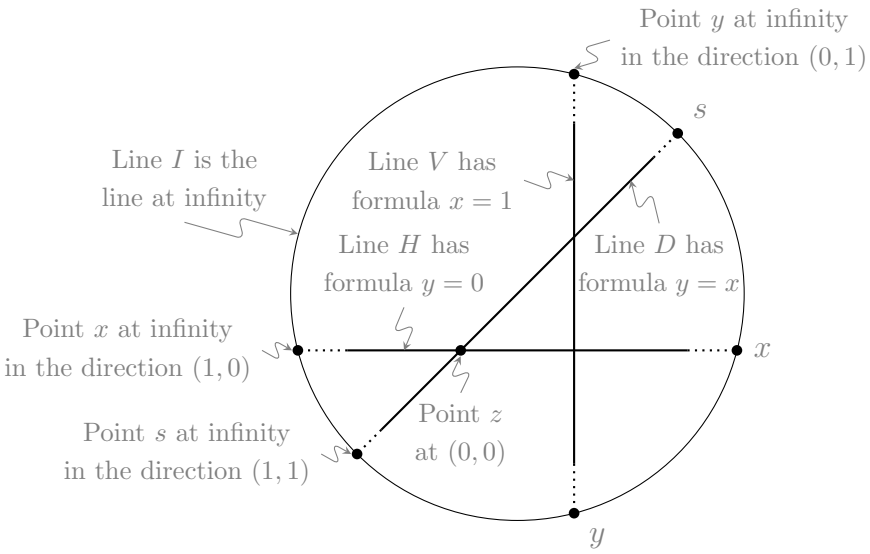


Figure 4.3: The eight vertices of the base graph represented as points and lines in the affine plane $z = 1$ and its extension to the projective plane.

the adjacency structure of G) to appear in matrix M. Recall that a row and column whose dot product is zero correspond to a pair of vertices that are adjacent in G.

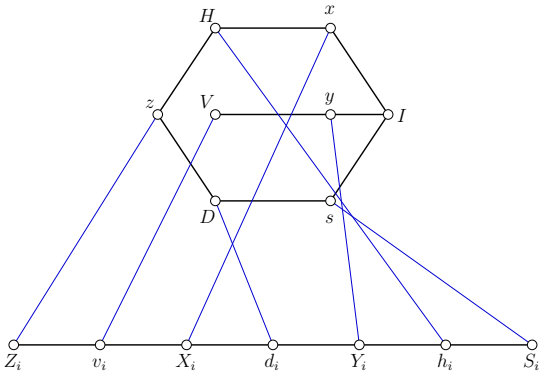


Figure 4.4: The base graph with the seven new vertices representing the variable q_i.

$$M = RC = \begin{array}{c} \\ V \\ H \\ I \\ D \\ Z_i \\ X_i \\ Y_i \\ S_i \end{array} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \\ q_i & -1 & 0 \\ 0 & 1 & -q_i \\ 1 & 0 & -q_i \\ -1 & 1 & q_i \end{bmatrix} \begin{array}{c} \begin{array}{ccccccc} x & y & z & s & v_i & d_i & h_i \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & q_i & q_i \\ 0 & 1 & 0 & 1 & q_i & q_i & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

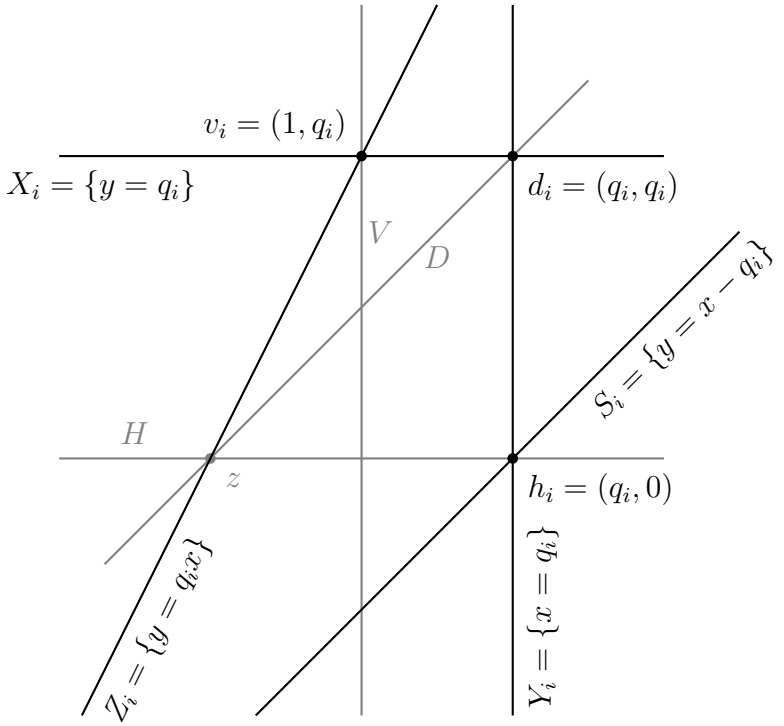


Figure 4.5: The eight starting vertices are depicted in gray as points and lines in the affine plane $z = 1$. Lines and points at infinity are not pictured. The seven points (lowercase) and lines (uppercase) associated with the named variable q_i are depicted in black, along with their coordinates or equations.

Just as with the base matrix and base graph, we may also think of each row in R as representing a plane in $\mathbb{R}^3$ and each column in C as representing a line in $\mathbb{R}^3$. Figure 4.5 depicts the intersection of these planes and lines with the affine plane P ($z = 1$), which can be extended to the projective plane. In this manner, we may again think of the graph G as representing point-line incidences in the projective plane. For example, the line X_i is incident

with the point v_i; this corresponds to the dot product of row X_i and column v_i being zero in the matrix, and the vertices X_i and v_i being adjacent in the graph.

The careful reader may have noticed a problem that can arise at this point in the construction. Our description of the graph makes it clear that none of the vertices $Z_a, v_a, X_a, d_a, Y_a, h_a, S_a$ should be adjacent to any of the vertices $Z_b, v_b, X_b, d_b, Y_b, h_b, S_b$ when $a \neq b$. However, if we are not careful, it is possible that some of the rows Z_a, X_a, Y_a, S_a may have a dot product of zero with the columns v_b, d_b, h_b for $a \neq b$. Equivalently, it is possible that some of the lines and points associated with q_a may be incident on those associated with q_b. We will list below a set of additional constraints on the values of the variables q_a and q_b which are sufficient to ensure that this does not happen, and call a set of values *sufficiently generic* when these constraints are satisfied for every pair q_a and q_b. These constraints are not satisfied in general for all solutions to a set S of equations, but a second pass at the construction, described in Section 4.7, will produce a larger graph and new variables for which the constraints are all satisfied. The twelve possible interactions, between the four lines of one variable q_a and the three points of another variable q_b, produce one constraint each (with some constraints repeated) as follows:

- The line Z_a does not contain
  - the point v_b from subspace Z_b if and only if $q_a \neq q_b$;
  - the point d_b from subspace D, if and only if $q_a \neq 1$;
  - or the point h_b from subspace H, if and only if $q_a \neq 0$.
- The horizontal line X_a does not contain
  - the point v_b from parallel X_b, if and only if $q_a \neq q_b$;
  - the point d_b from parallel X_b, if and only if $q_a \neq q_b$;
  - or the point h_b from parallel H, if and only if $q_a \neq 0$.
- The vertical line Y_a does not contain
  - the point v_b from parallel V, if and only if $q_a \neq 1$;
  - the point d_b from parallel Y_b, if and only if $q_a \neq q_b$;
  - or the point h_b from parallel Y_b, if and only if $q_a \neq q_b$.
- The line S_a, of slope 1 and equation $x = a + y$, does not contain
  - the point $v_b = (1, b)$ if and only if $q_a + q_b \neq 1$;
  - the point d_b from parallel D, if and only if $q_a \neq 0$;
  - or the point h_b from parallel S_b, if and only if $q_a \neq q_b$.

These twelve interactions produce only the following four constraints: $q_a \neq 1$, $q_a \neq q_b$, $q_a \neq 0$, and $q_a + q_b \neq 1$.

When working with a specific set of equations S and aiming to generate a small and explicitly described graph G_S, it may be preferable to allow pairs q_a and q_b that explicitly sum to 1, and then analyze the cases where these four constraints must be violated and add edges to the bipartite graph accordingly. However, for the general case, we instead add the additional pre-processing step described in Section 4.7, which constructs a larger graph using slack variables, and in doing so guarantees the satisfaction of all four constraints without requiring case-by-case analysis.

4.5.3 Adding Enforcing Equations

Now, to enforce the equations in T described at the beginning of Section 4.5, we need only add one new trivalent vertex per equation to the graph G, as detailed below.

For each equation of the type $q_i = q_j q_k$, we add a vertex e to the graph, with edges from e to Y_i, X_j, and Z_k, and correspondingly we add the column $\begin{bmatrix}1 & q_k & 1/q_j\end{bmatrix}^\top$ to the matrix C. The product of the new column with the rows X_i, Y_j, and Z_k in R then produces entries in M equal, respectively, to $q_k - q_i/q_j$, 0, and 0. For any assignment of q_i, q_j, q_k satisfying $q_i = q_j q_k$, the entry $q_k - q_i/q_j$ will equal 0 as well. These three zero entries in M correspond to the three new graph edges.

Likewise, for each equation of the type $q_i = q_j + q_k$, we add a vertex e to the graph, with edges from e to Y_i, X_j, and S_k, and correspondingly we add the column $\begin{bmatrix}q_i & q_j & 1\end{bmatrix}^\top$ to the matrix C. The product of the new column with the rows Y_i, X_j, and S_k in R then produces entries in M equal, respectively, to 0, 0, and $-q_i + q_j + q_k$. For any assignment of q_i, q_j, q_k satisfying $q_i = q_j + q_k$, the entry $-q_i + q_j + q_k$ will equal 0 as well. These three zero entries in M correspond to the three new graph edges.

Furthermore, for a new column of either type, the product with all other rows of R will be nonzero if and only if the constraints on variables listed in the last section are fulfilled (that is, for all indices $a \neq b$, we must have $q_a \neq q_b$, $q_a \neq 0$, and $1 \neq q_a + q_b$).

Figure 4.6 shows the subgraph of G corresponding to variables q_i, q_j, q_k, with vertices e_1 and e_2 added. Vertex e_1 corresponds to the equation $q_i = q_j q_k$, and vertex e_2 corresponds to the equation $q_i = q_j + q_k$.

Hence, if we are able to find a simultaneous solution to all the equations T, then we can construct R and C so that $M = RC$, and for all entries M_{ij} of M, $M_{ij} = 0$ if and only if vertices i and j are adjacent in the graph.

4.5.4 Final Construction

In this section, we define G_S in terms of G and show that if S (or, equivalently, T) has a solution, then G_S has minimum rank 3.

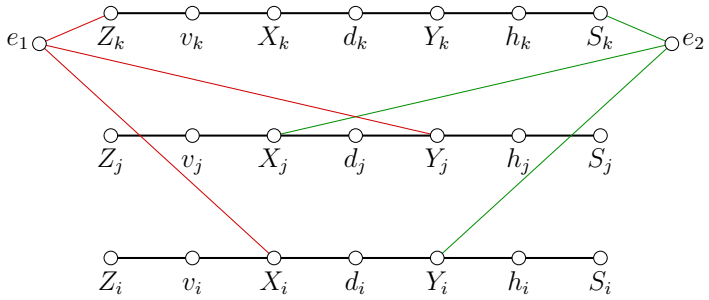


Figure 4.6: The equation vertex e_1 enforces $q_i = q_j q_k$; vertex e_2 enforces $q_i = q_j + q_k$.

First, we establish that the matrix M as constructed in the previous sections has rank 3. It should be clear from the first three rows and columns of R and C, as given in Section 4.5.1, that R and C are both of rank 3. To conclude that $M = RC$ is also of rank 3, we have the following lemma.

Lemma 4.8. *For an $n \times m$ matrix M, $\text{rank}(M) = 3$ if and only if M can be factored into RC where R is $n \times 3$, C is $3 \times m$, and $\text{rank}(R) = \text{rank}(C) = 3$.*

Proof. Let M be a matrix of rank 3 over a field $\mathbb{F}$. Consider the 3-dimensional vector space over $\mathbb{F}$ spanned by the rows of M. Let v_1, v_2, and v_3 be a basis for this space, and let C be the $3 \times m$ matrix whose rows are v_1, v_2, and v_3.

Let $w_1, w_2, \ldots, w_n$ be the row vectors of M. Then there exist $\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}$ such that $w_i = \alpha_{i,1}v_1 + \alpha_{i,2}v_2 + \alpha_{i,3}v_3$. Let R be the $n \times 3$ matrix such that $R_{i,j} = \alpha_{i,j}$. Then $M = RC$. It is well-known that the rank of a product of matrices is at most the rank of each factor. Therefore, $\text{rank}(R) \geq \text{rank}(M) = 3$. Since R has only three columns, $\text{rank}(R) = 3$. Similarly, $\text{rank}(C) = 3$.

Conversely, suppose there are rank-3 matrices R and C such that R is $n \times 3$ and C is $3 \times m$. We will show that $M = RC$ has rank 3. We know that $\text{rank}(M) \leq \min\{\text{rank}(R), \text{rank}(C)\} = 3$. Thus, we need only show that $\text{rank}(M) \geq 3$. Since R and C both have rank 3, they have 3×3 nonsingular submatrices (up to reordering of rows and columns). The product of these matrices is a 3×3 nonsingular submatrix of M, implying that M also has rank 3. □

We take the matrix M and expand it into A in the following manner, where $*$ indicates block matrices of the appropriate size with all nonzero entries.

$$A = \begin{bmatrix} * & M \\ M^T & * \end{bmatrix}$$

Now, we let $G_S = \overline{G}$, that is, G_S is defined as the complement of the graph G that we have constructed in the previous sections.

From the discussion in Section 4.4, we observe that the zero-nonzero pattern of M is the rectangular part of G_S. Then, from Theorem 4.6, we know that the minimum rank of G_S is equal to the minimum rank of the pattern of M, and thus, $\text{mr}(G_S) \leq \text{rank}(M)$. Therefore, if we are able to find a simultaneous solution to the system of triangulated equations, T, we can construct M of rank 3, and we have $\text{mr}(G_S) \leq 3$.

Now it only remains to show that $\text{mr}(G_S) > 2$. From [14], we know that if G_S has a path on 4 vertices as an induced subgraph, then $\text{mr}(G_S) > 2$. Since a path on 4 vertices is self-complementary, this condition is equivalent to G having a path on 4 vertices as an induced subgraph. G has multiple induced subgraphs of this type in the base graph alone, as can be seen in Figure 4.1.

In conclusion, if we are able to find a simultaneous solution to the system of triangulated equations T, then $\text{mr}(G_S) = 3$.

4.5.5 Minimum Rank 3 Implies a Solution to the Polynomials

For a given set of polynomial equations $S \subseteq \mathbb{Z}[x_1, x_2, ..., x_k]$, G_S is formed in the fashion described in the previous sections. To summarize the process, first, we triangulate and pre-process S to obtain T, a set of triangulated polynomial equations in $\mathbb{Z}[q_1, q_2, \cdots, q_m]$, whose solutions are in one-to-one correspondence with those of S. Then we build the graph G beginning with the core of eight vertices, add seven vertices for each variable q_i, and finally add a trivalent vertex for every equation in T. G_S is then defined as $\overline{G}$. We have seen that, if T has a solution, then $\text{mr}(G_S) = 3$.

It remains to show that if G_S has min rank 3, then S has a solution. Suppose G_S does have minimum rank 3. Then there is some matrix A with rank 3 such that $A_{i,j} = 0$ if and only if G contains the edge $\{i, j\}$. Since G is bipartite, its vertices can be partitioned into two sets V_R and V_C. Without loss of generality, we will designate the vertices corresponding to lines in the projective plane as V_R and the vertices corresponding to points in the projective plane as V_C, and by appropriate rearrangement of the rows and columns of A, we can decompose A as

$$A = \begin{bmatrix} * & M \\ M^T & * \end{bmatrix},$$

where the rows of M correspond to V_R and the columns of M correspond to V_C, and $*$ indicates a submatrix whose entries are all nonzero.

By Theorem 4.7, we know that the minimum rank of the pattern of A is equal to the minimum rank of the pattern of M. Furthermore, since M is rank 3, there exists a $|V_R| \times 3$ matrix R and a $3 \times |V_C|$ matrix C such that $M = RC$. Then we have that a particular row in R and column in C have a zero product if and only if their corresponding vertices are connected in G. In this sense, there is a one-to-one correspondence between vertices of V_R and the rows of R, and between the vertices of V_C and the columns of C.

By a suitable rearrangement of rows and columns of A, we can furthermore ensure that the first four rows of R correspond to the vertices V, H, I, D, and the first four columns of C

correspond to the vertices x, y, z, s. Furthermore, we can ensure that after the first four rows, each successive block of four rows in R corresponds to Z_i, X_i, Y_i, S_i, and each block of three columns in C corresponds to v_i, d_i, h_i, for $i = 1, \ldots, m$ (the indices of the variables q_i in T). This characterizes all the rows in R. As for C, we will rearrange columns suitably so that after the columns corresponding to the variables of T, we have the columns corresponding to the additive and multiplicative equations of T.

With this ordering, we have that the top left corner of M is

$$\begin{array}{c} \\ V \\ H \\ I \\ D \end{array} \begin{array}{c} \begin{array}{cccc} x & y & z & s \end{array} \\ \begin{bmatrix} * & 0 & * & * \\ 0 & * & 0 & * \\ 0 & 0 & * & 0 \\ * & * & 0 & 0 \end{bmatrix} \end{array}.$$

Since $M = RC$, any row operations we perform on R are equivalent to performing the same row operations on M, and any column operations we perform on C are equivalent to performing the same column operations on M. So long as our row and column operations preserve the rank and the zero structure of M, we will obtain another matrix of rank 3 which still corresponds to the graph G_S.

We begin by scaling the first two rows of M so that the top two entries in column s are 1. Then we scale the first three columns of M so as to make the first nonzero entry in each column become 1, 1, and -1 respectively. We then scale the third row of M so that its third entry is 1. The result is

$$\begin{array}{c} \\ V \\ H \\ I \\ D \end{array} \begin{array}{c} \begin{array}{cccc} x & y & z & s \end{array} \\ \begin{bmatrix} 1 & 0 & -1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ * & * & 0 & 0 \end{bmatrix} \end{array}.$$

Note that since the top left 3×3 submatrix of M is non-singular, this tells us that the top three rows of R and the left three columns of C are also non-singular submatrices. Now, since $M = RC$, we also have that for any invertible 3×3 matrix P, $M = (RP^{-1})(PC)$. Thus we can choose P to be the inverse of the first three columns of C, and then replace R with RP^{-1} and C with PC, so that the first three columns of C are the identity matrix.

This implies that R is identical to the first three columns of M. Furthermore, we can scale each row of M after the first three by any non-zero value, so as to control the value of one non-zero entry per row. Hence, R has the form below, where zeros correspond to edges in the graph G between the two vertices corresponding to the column and row of the entry in R. The reason for the particular entries chosen to be 1 or -1 will be clear later.

Note that in the representation of R below, we use the notation $M_{\sim,\sim}$ to stand for non-zero entries from the M matrix. Furthermore, it should be understood that the index i is the

index of the variables q_i in T, so that R actually contains rows Z_1, X_1, Y_1, S_1 followed by Z_2, X_2, Y_2, S_2, and so on. We may therefore write

$$R = \begin{array}{c} V \\ H \\ I \\ D \\ Z_i \\ X_i \\ Y_i \\ S_i \end{array} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & M_{D,y} & 0 \\ M_{Z_i,x} & -1 & 0 \\ 0 & 1 & M_{X_i,z} \\ 1 & 0 & M_{Y_i,z} \\ -1 & M_{S_i,y} & M_{S_i,z} \end{bmatrix}.$$

Now, turning our attention back to M, we can scale every column after the first three by any non-zero scalar, which allows us to choose one nonzero entry per column to be 1. Thus we force the first three rows of M to be as shown below, where $e_{i,j,k}$ are the vertices corresponding to multiplication equations in T and $f_{i,j,k}$ are the vertices corresponding to addition equations in T. The reason for choosing these particular entries to be 1 will be clear later. As in the case of R above, it should be understood that by columns v_i, d_i, h_i we really mean repeated blocks of three columns beginning with v_1, d_1, h_1, then v_2, d_2, h_2, and so on. Similarly, we use $e_{i,j,k}$ to stand for any column corresponding to a multiplicative equation $q_i = q_j q_k$ in T and $f_{i,j,k}$ to stand for any column corresponding to an additive equation $q_i = q_j + q_k$ in T.

$$M = \begin{array}{c} \\ V \\ H \\ I \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccccccc} x & y & z & s & v_i & d_i & h_i & e_{i,j,k} & f_{i,j,k} \end{array} \\ \begin{bmatrix} 1 & 0 & -1 & 1 & 0 & * & * & 1 & * \\ 0 & 1 & 0 & 1 & * & * & 0 & * & * \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & * & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \end{array}$$

Just as we determined R from the first three columns of C, we can also determine C from the first three rows of R. The first three rows of R are

$$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

which tells us that the second and third rows of C will be identical to those of M, whereas the first row of C will be the sum of the first and third rows of M. Hence, we can write C in the following form, where $M_{\sim,\sim}$ again denotes nonzero entries from the M matrix.

$$C = \begin{array}{c} \begin{array}{ccccccccc} x & y & z & s & v_i & d_i & h_i & e_{i,j,k} & f_{i,j,k} \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & M_{V,d_i}+1 & M_{V,h_i}+1 & M_{I,e_{i,j,k}}+1 & M_{V,f_{i,j,k}}+1 \\ 0 & 1 & 0 & 1 & M_{H,v_i} & M_{H,d_i} & 0 & M_{H,e_{i,j,k}} & M_{H,f_{i,j,k}} \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & M_{I,e_{i,j,k}} & 1 \end{bmatrix} \end{array}$$

We will use the notation $R_\sim$ for the row labeled $\sim$ in R and the notation $C_\sim$ for the column labeled $\sim$ in C. Recall that when vertices A, B are connected in the graph G, $R_A C_B = 0$. Hence, consulting Figure 4.4, we have the following. Below, i ranges over the indices of the variables q_i of the polynomials in T.

$$R_D C_s = 0 \Rightarrow M_{D,y} = 1$$

$$R_D C_{d_i} = 0 \Rightarrow M_{V,d_i} = M_{D,y} M_{H,d_i} - 1$$

$$R_{S_i} C_s = 0 \Rightarrow M_{S_i,y} = 1$$

$$R_{Z_i} C_{v_i} = 0 \Rightarrow M_{Z_i,x} = M_{H,v_i}$$

$$R_{X_i} C_{v_i} = 0 \Rightarrow M_{H,v_i} = -M_{X_i,z}$$

$$R_{X_i} C_{d_i} = 0 \Rightarrow M_{H,d_i} = -M_{X_i,z}$$

$$R_{Y_i} C_{d_i} = 0 \Rightarrow M_{V,d_i} = -M_{Y_i,z} - 1$$

$$R_{Y_i} C_{h_i} = 0 \Rightarrow M_{V,h_i} = -M_{Y_i,z} - 1$$

$$R_{S_i} C_{h_i} = 0 \Rightarrow M_{S_i,z} = M_{V,h_i} + 1$$

If we rename the entry M_{H,d_i} by the name q_i, the above equations yield:

$$\begin{array}{llll} M_{D,y} = 1 & M_{Z_i,x} = q_i & M_{X_i,z} = -q_i & M_{Y_i,z} = -q_i \\ M_{S_i,y} = 1 & M_{S_i,z} = q_i & M_{V,s} = 1 & M_{H,v_i} = q_i \\ M_{V,d_i} = q_i - 1 & M_{H,d_i} = q_i & M_{V,h_i} = q_i - 1 & \end{array}$$

Hence, our R and C matrices are as shown below:

$$R = \begin{array}{c} \\ V \\ H \\ I \\ D \\ Z_i \\ X_i \\ Y_i \\ S_i \end{array} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \\ q_i & -1 & 0 \\ 0 & 1 & -q_i \\ 1 & 0 & -q_i \\ -1 & 1 & q_i \end{bmatrix}$$

$$C = \begin{array}{c} \begin{matrix} x & y & z & s & v_i & d_i & h_i & e_{i,j,k} & f_{i,j,k} \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & q_i & q_i & M_{I,e_{i,j,k}} + 1 & M_{V,f_{i,j,k}} + 1 \\ 0 & 1 & 0 & 1 & q_i & q_i & 0 & M_{H,e_{i,j,k}} & M_{H,f_{i,j,k}} \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & M_{I,e_{i,j,k}} & 1 \end{bmatrix} \end{array}$$

Now we turn our attention to the columns $e_{i,j,k}$ and $f_{i,j,k}$ in C, corresponding respectively to the multiplication equation $q_i = q_j q_k$ and the addition equation $q_i = q_j + q_k$ in T. As shown in Figure 4.6, the vertex corresponding to a multiplication equation $q_i = q_j q_k$ is connected to exactly three vertices, X_i, Y_j, and Z_k. Hence, we have the following. For every multiplication equation of the form $q_i = q_j q_k$ in T,

$$R_{X_i} C_{e_{i,j,k}} = 0 \Rightarrow M_{H,e_{i,j,k}} = q_i M_{I,e_{i,j,k}}$$
$$R_{Y_j} C_{e_{i,j,k}} = 0 \Rightarrow M_{I,e_{i,j,k}} + 1 = q_j M_{I,e_{i,j,k}}$$
$$R_{Z_k} C_{e_{i,j,k}} = 0 \Rightarrow q_k (M_{I,e_{i,j,k}} + 1) = M_{H,e_{i,j,k}}$$

After substitution, these equations yield $q_i M_{I,e_{i,j,k}} = q_j q_k M_{I,e_{i,j,k}}$, and since $M_{I,e_{i,j,k}}$ is a nonzero entry of M, we recover the equation $q_i = q_j q_k$.

Likewise, the vertex corresponding to an addition equation $q_i = q_j + q_k$ is connected to exactly three vertices, X_j, Y_i, and S_k. Hence, for every addition equation of the form $q_i = q_j + q_k$ in T,

$$R_{X_j} C_{f_{i,j,k}} = 0 \Rightarrow M_{H,f_{i,j,k}} = q_j$$
$$R_{Y_i} C_{f_{i,j,k}} = 0 \Rightarrow M_{V,f_{i,j,k}} + 1 = q_i$$
$$R_{S_k} C_{f_{i,j,k}} = 0 \Rightarrow M_{H,f_{i,j,k}} + q_k = M_{V,f_{i,j,k}} + 1$$

After substitution, these yield the equation $q_i = q_j + q_k$.

Thus, we have shown that if G_S has minimum rank 3, then T has a solution, which implies that S has a solution, proving Theorem 4.3.

4.6 Examples and Minimal Obstructions

In this section we construct some examples to illustrate the technique, including, as promised by Theorem 4.1, an explicit infinite list of minimal obstructions to 3-RANK and to $\mathbb{F}$-3-RANK for any infinite field $\mathbb{F}$. For context, we start with a summary of known results for lower ranks, or for finite fields.

4.6.1 Ranks 0, 1, and 2

The only matrix of rank 0 is the zero matrix, corresponding to a graph with no edges, from which it follows that, over any field, the unique obstruction to rank 0 is the induced subgraph K_2.

Every matrix of rank 1 can be factored as an outer product $\mathbf{x}^T\mathbf{x}$, and so the pattern of such a matrix is a complete graph of some size (corresponding to the non-zero entries of $\mathbf{x}$) together with isolated vertices (corresponding to the zero entries of $\mathbf{x}$). Such graphs are characterized by having at most one connected component with any edges, within which component the maximum distance is 1. From this it follows that the only obstructions to rank 1, over any field, are the disjoint union of two edges $2K_2$ and the path on three vertices P_3.

A complete description of graphs with minimal rank 2 over any infinite field $\mathbb{F}$ is given in [14], in terms of a finite list of forbidden induced subgraphs. The same is done for finite fields in [15]. It follows, for any given field $\mathbb{F}$, that $\mathbb{F}$-2-RANK belongs to P, the class of decision problems solvable in time that is polynomial in the size of the input. This includes $\mathbb{R}$-2-RANK, which is to say 2-RANK. For example, because the largest obstruction to rank 2 for a real symmetric matrix is the 9-vertex graph $K_{3,3,3}$, a brute-force search for obstructions gives an algorithm of complexity $O(n^{9/2})$ for resolving 2-RANK (and a faster search for $K_{3,3,3}$ in particular would improve the running time). The infinite families that we are about to construct show that minimum rank 3 has no such finite list of obstructions.

4.6.2 Finite Fields

For finite fields more generally, the rank problem is known to be fixed-parameter tractable: If $\mathbb{F}$ is finite and d is specified, then $\mathbb{F}$-d-RANK belongs to P. This follows from the work of [35] that gives explicit bounds, in terms of $|\mathbb{F}|$ and d, on the number and size of a sufficient set of matrices, whose ranks can be individually checked to produce a complete finite list of obstructions. These explicit bounds are not practical, even for very small fields and small ranks, for finding all the obstructions for $\mathbb{F}$-d-RANK. The special case of $|\mathbb{F}| = 2$ and rank 3 is covered by [13], which gives an explicit list of all 62 obstructions, involving at most 8 vertices, whereas using only the bounds from [35] would require checking the ranks of all binary symmetric matrices of size up to 25×25—more than 10^{65} matrices, even up to symmetry. A brute-force search for all 62 obstructions would give an algorithm of running time $O(n^4)$ for $GF(2)$-3-RANK, which is probably far from optimal. The field $GF(2)$ is particularly friendly to graph-theoretic questions: Since there is only one non-zero element, the pattern alone tells you every non-diagonal entry. The remaining $|V|$ undetermined diagonal entries mean that $GF(2)$-RANK can be determined, even naïvely, by computing a set of $O(2^{\sqrt{n}})$ binary matrix ranks, and with a bit more sophistication it would not be surprising to learn that the problem is tractable even in its non-fixed-parameter version. For a finite field $\mathbb{F}$ with 3 elements or more, on the other hand, it seems likely that with the parameter d no longer fixed, the problem $\mathbb{F}$-RANK becomes intractable.

4.6.3 The Tetrahub Wheels

This leaves the question of minimal obstructions to minimum rank 3 over infinite fields, which will include graphs from two infinite families.

Given $m \geq 2$, the *tetrahub wheel number* m, or m^{th} *tetrahub*, which will be denoted by TH_m, is a graph on $4m+4$ vertices containing an induced cycle on $4m$ vertices. The vertices of the cycle are labeled $c_0, c_1, c_2, \ldots, c_{4m-1}$ in cyclic order, and the additional four "hub" vertices are labeled h_0, h_1, h_2, and h_3. The cycle vertices are partitioned into four residue classes modulo 4, and each cycle vertex in residue class k is connected to the appropriate hub vertex h_k.[3] Figure 4.7 depicts TH_5.

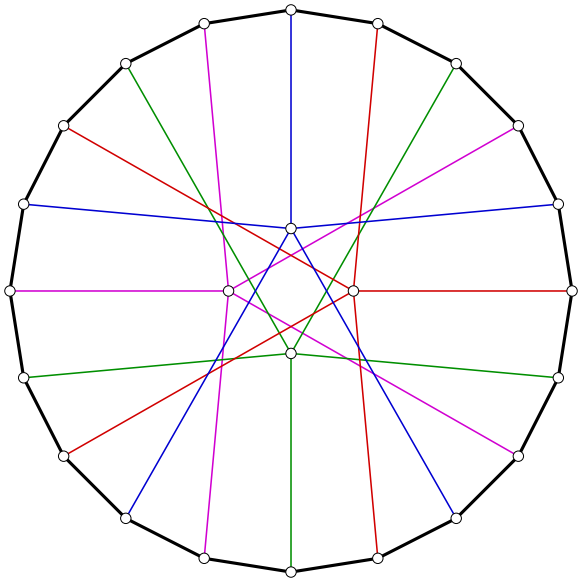


Figure 4.7: The fifth tetrahub, TH_5

Given $m \geq 2$, the *tetrahub wheel number* m *with axle,* or m^{th} *tetrahub with axle,* which will be denoted THA_m, is a graph on $4m+6$ vertices which has TH_m as an induced subgraph. The additional two vertices a_0 and a_1, which form the "axle" of the wheel, are connected to each other and to the hubs of the same parity by the five edges

$$\{a_0, a_1\}, \{a_0, h_0\}, \{a_0, h_2\}, \{a_1, h_1\}, \text{ and } \{a_1, h_3\}.$$

Figure 4.8 depicts THA_6.

Remark 4.9 (Symmetries and special cases)**.** The cycle vertices of a tetrahub, with or without axle, all have degree 3, as do the axle vertices of a tetrahub with axle. The hub vertices of TH_m have degree m, and the hub vertices of THA_m have degree $m+1$. The

[3]According to this nomenclature, the ordinary wheel graphs would be called "monohub" wheels.

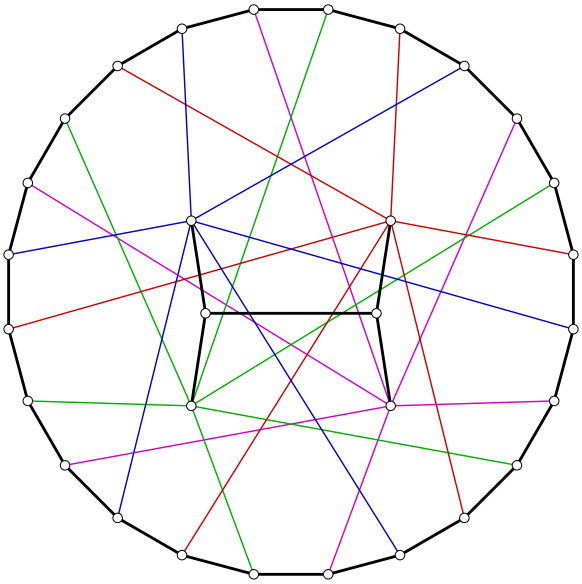


Figure 4.8: The sixth tetrahub with axle, THA_6

symmetry group of the graph must include at least the dihedral group of order $8m$ acting on the cycle, and since symmetries preserve degree, it is not hard to see that these are all possible symmetries for TH_m with $m \neq 3$ or for THA_m with $m \neq 2$. The special case THA_2 turns out to have additional symmetries, and indeed, THA_2 is the Heawood graph whose symmetry group of order $336 = (21)(8m)$ is transitive on its vertices. The special case TH_3 also turns out to have additional symmetries: TH_3 is the Möbius-Kantor graph whose symmetry group of order $96 = (4)(8m)$ is also transitive on its vertices. The Heawood graph was the starting point in constructing these families. It is the incidence graph of the Fano plane or projective plane of order 2, which is well-known (see [75], for example) to be a minimal rank-3 matroid that cannot be realized by independence relations of vectors in $\mathbb{R}^3$ or $\mathbb{C}^3$ (since the deletion of any point from the Fano plane can be realized in this way). Marking each point-line incidence by a 0 in the corresponding row and column thus gives a pattern Y with $\mathrm{mr}(Y) > 3$, and Y proves to be locally minimal (with respect to row or column deletion) for that property. It is natural to next examine the projective plane of order 3, which similarly gives a pattern Y with $\mathrm{mr}(Y) > 3$. In this case, however, the pattern Y is far from locally minimal; it suffices to use 8 of the 13 points and 8 of the 13 lines, leaving the well-studied Möbius-Kantor configuration, which can be represented in $\mathbb{C}^3$ but not in $\mathbb{R}^3$. Examination of these two cases, and in particular of the sequence of algebraic identities that forces $\mathrm{mr}(Y) > 3$, led to the generalizations TH_m and THA_m.

We will use the fact that TH_m and THA_m are bipartite, and also the fact that in each case, taking a matrix with the pattern of the complement graph and examining the off-diagonal block yields a rectangular pattern Y that is reduced. The graphs are bipartite by

construction, with one side of the partition taking in the even-numbered cycle vertices, then the odd-numbered hub vertices, then possibly the even-numbered axle vertices. The pattern Y from the complement is a reduced rectangular pattern because no vertex in the bipartite graph is adjacent to every vertex on the other side of the partition. In fact, every vertex is non-adjacent to at least two vertices on the other side of the partition, from which we can deduce the same two facts for every single-vertex deletion of TH_m or of THA_m.

The following lemma summarizes the necessary facts about minimum ranks of the complements of these graphs and their single-vertex deletions.

Lemma 4.10. *Given an infinite field $\mathbb{F}$ and an integer $m \geq 2$,*

- *let G be the complement of TH_m,*
- *let G' be any single-vertex deletion of G,*
- *let H be the complement of THA_m, and*
- *let H' be any single-vertex deletion of H,*

and consider the following cases:

1. *The characteristic of $\mathbb{F}$ is m.*
2. *The characteristic of $\mathbb{F}$ is other than m, and the field $\mathbb{F}$ contains an m^{th} root of unity.*
3. *The characteristic of $\mathbb{F}$ is other than m, and the field $\mathbb{F}$ does not contain an m^{th} root of unity.*

Then the following statements all hold:

- *In Case 1 and Case 2, the minimum rank of G over $\mathbb{F}$ is 3.*
- *In Case 3, the minimum rank of G over $\mathbb{F}$ is 4.*
- *In Case 3, the minimum rank of G' over $\mathbb{F}$ is 3.*
- *In Case 1, the minimum rank of H over $\mathbb{F}$ is 3.*
- *In Case 2, the minimum rank of H over $\mathbb{F}$ is 4.*
- *In Case 2, the minimum rank of H' over $\mathbb{F}$ is 3.*

In particular, for an infinite field $\mathbb{F}$ not of characteristic m and not containing a primitive m^{th} root of unity, the complement of TH_m is a minimal obstruction to minimum rank 3, and for an infinite field $\mathbb{F}$ not of characteristic m and containing a primitive m^{th} root of unity, the complement of THA_m is a minimal obstruction to minimum rank 3. This will suffice, once Lemma 4.10 is proven, to give the main results of this section.

Theorem 4.11. *Over any infinite field $\mathbb{F}$, there is no finite list of graphs such that a graph has minimum rank 3 over $\mathbb{F}$ if and only if it has no induced subgraph from the list.*

Proof. For any infinite field $\mathbb{F}$, we exhibit, supported by Lemma 4.10, an infinite sequence of minimal obstructions O_m to minimum rank 4, indexed by an infinite collection of positive integers $m \geq 2$.

In those cases where the characteristic p of $\mathbb{F}$ is positive, we leave O_m undefined for m equal to p. Otherwise, we take O_m to be the complement of TH_m in the case that $\mathbb{F}$ does not contain a primitive m^{th} root of unity, and otherwise we take O_m to be complement of THA_m. In particular:

- For $\mathbb{F} = \mathbb{R}$ we have the infinite sequence $\overline{\mathrm{THA}_2}, \overline{\mathrm{TH}_3}, \overline{\mathrm{TH}_4}, \ldots$.
- For $\mathbb{F} = \mathbb{C}$ we have the infinite sequence $\overline{\mathrm{THA}_2}, \overline{\mathrm{THA}_3}, \overline{\mathrm{THA}_4}, \ldots$.

□

Proof of Lemma 4.10. Suppose that a value x exists which is a primitive m^{th} root of unity, and let a be any number. Then we have

$$a = x^m a.$$

Letting this single equation be the system S, we construct the graph G_S corresponding to the triangulation $q_0 = x$, $q_1 = a$, and

$$\begin{aligned} & q_2 = q_0 \cdot q_1 && = xa \\ & q_3 = q_0 \cdot q_2 && = x^2 a \\ & q_4 = q_0 \cdot q_3 && = x^3 a \\ & \vdots \qquad \vdots && \vdots \\ & q_m = q_0 \cdot q_{m-1} && = x^{m-1} a \\ a = \quad & q_1 = q_0 \cdot q_m && = x^m a. \end{aligned}$$

□

Now let $\mathbb{F}$ be any infinite field. For each integer $i > 1$, let $B_i = \mathrm{TH}_{2i+2}$ if $\mathbb{F}$ has no primitive root of unity of order $2i + 2$, and let $B_i = \mathrm{THA}_{2i+2}$ otherwise.

We form a sequence of graphs indexed by the even integers m starting at $m = 4$, and for each such m, we select the 4-hub m-wheel if $\mathbb{F}$ has no primitive m^{th} root of unity, and the 4-hub m-wheel with axle otherwise.

Since no field has even characteristic greater than 2, each B_i will have constraint rank 4 over $\mathbb{F}$. Similarly, every single-vertex deletion of every graph in the list will have constraint rank 3 over $\mathbb{F}$. The theorem thus holds for every infinite field $\mathbb{F}$.

4.6.4 Concluding Remarks

With the proofs of Theorem 4.3 and Theorem 4.11, we have shown that, unlike the 0-RANK, 1-RANK, and 2-RANK problems, which can be solved in polynomial time by checking a

finite list of forbidden subgraphs, the complexity class of the 3-RANK problem is equivalent to $\exists\,\mathbb{R}$, which is NP-hard and in PSPACE. This confirms a long-held intuition amongst those who have worked on the minimum rank problem: that the d-RANK problem is fundamentally different for $d \leq 2$ than for $d \geq 3$. In the course of proving our results, we have also established a concrete method for interpreting any system of equations in the context of projective geometry.

These hardness results indicate that, for large instances, it is computationally intractable to solve d-RANK exactly for $d \geq 3$; however, it does not spell the end of applications for minimum rank and related matrix completion problems. Rather, it indicates a need for approximate algorithms and methods (like those found in Chapter 5) to bound the minimum rank of large graphs.

4.7 Appendix: Triangulation and Slack Variables

We now provide an explicit description of the two pre-processing steps described in Section 4.3.1. In Section 4.7.1, we detail a procedure to transform the set of polynomial equations S on variables $\mathbf{x}$ into a set of equations T on variables $\mathbf{q}$, where all equations in T are either of the form $q_i = q_j + q_k$ or the form $q_i = q_j q_k$ (with i, j, k not necessarily distinct). In Section 4.7.2, we describe further procedures, involving the creation of auxiliary and slack variables, which ensure that we may assume the variables $\mathbf{q}$ are generic with respect to a set of specific constraints, namely that for all indices $i \neq j$, we have $q_i \neq 0$, $q_i \neq q_j$, and $q_i + q_j \neq 1$.

All procedures in Sections 4.7.1 and 4.7.2 can be done in time that is polynomial in the size of the input, where the input is a binary string that encodes the set of polynomial equations according to some standard format—including, for example, binary encoding of any integers that appear as coefficients or as exponents. Each of the base variables and slack variables is given a label $q_1, q_2, \ldots$, and further labels are given to every intermediate stage of the calculation.

4.7.1 First Pass

A set of polynomial equations with integer coefficients is *triangulated* if the variables are $q_1, q_2, \ldots, q_k$ and if every equation takes either the form $q_i = q_j + q_k$ or the form $q_i = q_j q_k$ for three indices $i, j, k \geq 1$, not necessarily distinct. We begin by detailing a polynomial time procedure to triangulate S. Our goal is to use simple substitutions and reversible algebraic manipulations to transform S into a set T of triangulated equations whose solutions are in one-to-one correspondence with those of S. This may involve the addition of a small number of *slack variables* whose particular value is a new degree of freedom that plays no role in the original equations and is taken to be nonzero. (The second pass will introduce a large number of slack variables entirely replacing the original variables.) It will also involve the addition of new labels for every intermediate stage of computation.

In order to satisfy the claim of polynomial overhead while also accomodating, if desired, the possibility of parenthesized expressions in the input, a first stage can be employed that replaces all parenthesized expressions in terms of new variables, beginning with innermost expressions. For example, the single equation

$$(a+b)^{200} = (c+d)^{200}$$

would be transformed to the three equations

$$\begin{aligned} x &= a+b \\ y &= c+d \\ x^{200} &= y^{200} \end{aligned}$$

where x and y are new variables. This is iterated recursively until all parentheses are eliminated, with only polynomial overhead.

The second stage is to move all monomials to the left or right side of every equation in such a way as to eliminate all negative coefficients; for example, the equation $3ab^2 - 2b + 5 = 0$ would become $3ab^2 + 5 = 2b$. We wish to avoid the situation where one side or the other of an equation is zero. This can be accomplished by adding a slack variable to both sides; for example, $a + b = 0$ would become $a + b + t = t$ where $t \neq 0$ is a new variable not appearing elsewhere in the equations. We also wish to avoid having any constant terms. This can be accomplished by multiplying both sides of an equation by a slack variable; for example, $3ab^2 + 5 = 2b$ would become $3ab^2s + 5s = 2bs$ where $s \neq 0$ is a new variable not appearing elsewhere in the equations.

It is now time to allocate labels $q_1, q_2, \ldots$, first to the base variables and then to all intermediate stages of computation on both sides of every equation. Any positive integer power of a variable is calculated by repeated multiplication, including repeated squaring when the exponent is large, since the input is expressed in binary notation and we have promised only polynomial-sized overhead. Similarly, any positive integer coefficient is calculated as a multiple of its monomial by repeated addition, including repeated doubling when the coefficient is large. An example will illustrate.

Example 4.12. *We encode the single equation*

$$2a^{26} = 17bc^2$$

by first assigning $q_1 = a$, $q_2 = b$, and $q_3 = c$. Then, on both sides of the equation, we allocate new labels to intermediate stages of the calculation, using the binary representations $26 = 11010_2$ and $17 = 10001_2$, as

$$
\begin{array}{llll}
q_4 = q_1 \cdot q_1 & (= a^2) & q_{11} = q_3 \cdot q_3 & (= c^2) \\
q_5 = q_4 \cdot q_4 & (= a^4) & q_{12} = q_2 \cdot q_{11} & (= bc^2) \\
q_6 = q_5 \cdot q_5 & (= a^8) & q_{13} = q_{12} + q_{12} & (= 2bc^2) \\
q_7 = q_6 \cdot q_6 & (= a^{16}) & q_{14} = q_{13} + q_{13} & (= 4bc^2) \\
q_8 = q_7 \cdot q_5 & (= a^{24}) & q_{15} = q_{14} + q_{14} & (= 8bc^2) \\
q_9 = q_8 \cdot q_4 & (= a^{26}) & q_{16} = q_{15} + q_{15} & (= 16bc^2) \\
q_{10} = q_9 + q_9 & (= 2a^{26}) & q_{10} = q_{16} + q_{12} & (= 17bc^2)
\end{array}
$$

Since the common label q_{10} terminates both calculations, this encodes the desired equation $2a^{26} = 17bc^2$.

If it happens that the same intermediate expression occurs in different equations (or in different parts of the same equation), then the already-determined labels for that subexpression should be used rather than repeating any of the steps leading up to the shared subexpression.

4.7.2 Second Pass

For a given set of elementary equations in variables $q_1, q_2, q_3, \ldots$, the reduction to point-line incidences in the projective plane produces a bipartite graph whose complement G has the property such that G has minimum rank 3 if and only if there exists an assignment of the variables satisfying two sets of conditions: Firstly, all the elementary equations must be satisfied; and secondly, for $i \neq j$, we must have $q_i \neq 0$, $q_i \neq q_j$, and $q_i + q_j \neq 1$. We wish to remove the latter set of restrictions, which we do by way of a second pass that produces a new, larger set of variables, and a new set of elementary equations, such that even non-interesting solutions of S will correspond, for some generic choice of the slack variables, to an interesting solution in the higher-dimensional space of auxiliary and slack variables.

We do require a pre-processing stage between the first pass and the second pass that eliminates certain redundancies. The pre-processing stage examines all pairs of elementary additions and all pairs of elementary multiplications looking for certain patterns of repeated variables, and either eliminates a redundant equation or identifies a pair of variables q_i and q_ℓ that must be equal, in which case q_ℓ is eliminated and replaced everywhere by q_i. This substitution may in turn produce new examples of patterns of repeated variables, and so the process is run recursively until no examples remain. The patterns flagged by the pre-processing stage are listed below. When an elementary equation such as $q_i = q_j + q_k$ is listed, the intention is for the preprocessor to scan both for that equation and for the elementary equation $q_i = q_k + q_j$, which is mathematically equivalent but distinct as an elementary equation, and similarly for $q_i = q_j \cdot q_k$ and its mathematical equivalent $q_i = q_k \cdot q_j$.

- $q_i = q_j + q_k$ and $q_i = q_j + q_k$. The second equation is redundant and can be removed.

- $q_i = q_j + q_k$ and $q_\ell = q_j + q_k$. This pair of elementary additions implies the mathematical relationship $q_\ell = q_i$. The preprocessor should remove the variable q_ℓ, replacing it by q_i in every elementary equation where it occurs. The second equation becomes redundant as a result and is also eliminated.
- $q_i = q_j \cdot q_k$ and $q_i = q_j \cdot q_k$. Eliminate the redundant second elementary multiplication.
- $q_i = q_j \cdot q_k$ and $q_\ell = q_j \cdot q_k$. Eliminate q_ℓ, replacing it everywhere by q_i, and then eliminate the second redundant equation.
- $q_i = q_i + q_j$ and $q_j = q_i + q_i$. The first equation implies that the numerical value of q_j must be 0, after which the second equation implies that $2q_i = 0$. Over any field of characteristic other than 2 (in particular, in the case of most interest $\mathbb{F} = \mathbb{R}$), we can conclude that $q_i = 0$ and in particular that $q_j = q_i$, meaning that the preprocessor should replace q_j by q_i and eliminate the second redundant equation. In the special case where $\mathbb{F}$ has characteristic 2, we do not eliminate a variable, but in that case the two equations are equivalent and the second, redundant elementary addition should be eliminated.
- $q_i = q_j + q_k$ and $q_j = q_i + q_k$. Substituting the second equation into the first yields the mathematical relationship $q_i = q_i + 2q_k$, which in characteristic other than 2 yields $q_k = 0$ and therefore $q_j = q_i$, meaning that the preprocessor should replace q_j by q_i and eliminate a redundant equation. In characteristic 2, no variable is eliminated, but the second equation is redundant and is eliminated.

At the end of the pre-processing phase, there will remain no pair of elementary additions and no pair of elementary multiplications that fit any of these patterns of repeated variables.

We are now ready for the second pass. A few slack variables may have been introduced in the first pass, each of which is supposed to take a generic value that in particular is nonzero, and each of which was assigned a name q_i for some value of i. These we set aside to keep unchanged. For all other original variables q_i, we introduce an auxiliary variable p_i and a slack variable s_i. The relation $p_i = q_i + s_i$ is not quite sufficient, because in the case of a solution of S for which q_i was equal to 0, we would have $p_i = s_i$ and thus two named variables with equal values, which is among the unwanted coincidences that we must avoid. We thus employ an additional universal slack variable named t. The original named variables and elementary equations will be expressed in terms of auxiliary variables and slack variables using the following relation:

$$p_i = q_i + s_i + t \quad \text{or} \quad q_i = p_i - s_i - t.$$

Note that under this choice we must avoid ever giving a name to any quantity $s_i + t$, since it could, when q_i happens to equal 0, be forced to the same value as the named quantity p_i.

The original addition and multiplication equations must be re-written in such a way that the original variables q_i are never mentioned, which will produce longer equations that themselves must be decomposed in terms of additional intermediate variables. We tackle the

easier case of addition first. An elementary addition equation takes the form $q_i = q_j + q_k$. We have

$$\begin{aligned} & q_i = q_j + q_k \\ \iff \quad & p_i - s_i - t = p_j - s_j - t + p_k - s_k - t \\ \iff \quad & p_i + s_j + s_k + t = p_j + p_k + s_i. \end{aligned}$$

This must be decomposed into elementary addition equations with new named variables $a_1(i,j,k), \ldots$, each of which can be expressed in terms of the old variables q_i and the slack variables s_i and t. Care must be taken to ensure that no new variable is the sum of a single slack variable s_i with the universal slack variable named t. Concretely, we decompose the equations as follows:

$$\begin{aligned} a_1(i,j,k) &= s_j + s_k & &= s_j + s_k \\ a_2(i,j,k) &= p_i + a_1(i,j,k) & &= q_i + s_i + s_j + s_k + t \\ a_3(i,j,k) &= p_j + p_k & &= q_j + q_k + s_j + s_k + 2t \\ a_4(i,j,k) &= a_2(i,j,k) + t & &= q_i + s_i + s_j + s_k + 2t \\ a_4(i,j,k) &= a_3(i,j,k) + s_i & &= q_j + q_k + s_i + s_j + s_k + 2t \end{aligned}$$

Note that the new variable named $a_4(i,j,k)$ occurs on the left hand side of two equations. This gives it two different expanded expressions in terms of old registers and slack variables, and those two expressions differ only by the identity $q_i = q_j + q_k$.

Now we rewrite an elementary multiplication equation, which starts in the form $q_i = q_j \cdot q_k$. We will find that some of the intermediate named quantities reoccur in other multiplication equations, and so we begin by defining the following named new variables for any j such that q_j ever occurs as a factor (i.e., on the right-hand side) in an elementary multiplication equation:

$$\begin{aligned} p'_j &= p_j \cdot t & &= q_j t + s_j t + t^2 \\ s'_j &= s_j \cdot t & &= s_j t. \end{aligned}$$

We also define two universal new named variables

$$\begin{aligned} t' &= t \cdot t & &= t^2 \\ t'' &= t + t' & &= t + t^2. \end{aligned}$$

We have

$$\begin{aligned} & q_i = q_j \cdot q_k \\ \iff \quad & p_i - s_i - t = (p_j - s_j - t)(p_k - s_k - t) \\ \iff \quad & p_i + p_k s_j + p_j s_k + p_j t + p_k t = p_j p_k + s_i + s_j s_k + t + s_j t + s_k t + t^2 \\ \iff \quad & p_i + (p_k \cdot s_j) + (p_j \cdot s_k) + p'_j + p'_k = (p_j \cdot p_k) + s_i + (s_j \cdot s_k) + s'_j + s'_k + t'' \end{aligned}$$

which concretely, in the case $j \neq k$, we decompose as

$$\begin{aligned}
m_1(i,j,k) &= p_j \cdot s_k & &= q_j s_k + s_j s_k + s_k t \\
m_2(i,j,k) &= p_k \cdot s_j & &= q_k s_j + s_j s_k + s_j t \\
m_3(i,j,k) &= p_j \cdot p_k & &= q_j q_k + q_k s_j + q_j s_k + s_j s_k + q_j t + q_k t + s_j t + s_k t + t^2 \\
m_4(i,j,k) &= s_j \cdot s_k & &= s_j s_k \\
m_5(i,j,k) &= p_i + m_1(i,j,k) & &= q_i + s_i + q_j s_k + s_j s_k + t + s_k t \\
m_6(i,j,k) &= m_5(i,j,k) + m_2(i,j,k) & &= q_i + s_i + q_k s_j + q_j s_k + 2 s_j s_k + t + s_j t + s_k t \\
m_7(i,j,k) &= m_6(i,j,k) + p'_j & &= q_i + s_i + q_k s_j + q_j s_k + 2 s_j s_k + t + q_j t + 2 s_j t + s_k t + t^2 \\
m_8(i,j,k) &= m_3(i,j,k) + s_i & &= q_j q_k + s_i + q_k s_j + q_j s_k + s_j s_k + q_j t + q_k t + s_j t + s_k t + t^2 \\
m_9(i,j,k) &= m_8(i,j,k) + m_4(i,j,k) & &= q_j q_k + s_i + q_k s_j + q_j s_k + 2 s_j s_k + q_j t + q_k t + s_j t + s_k t + t^2 \\
m_{10}(i,j,k) &= m_9(i,j,k) + s'_j & &= q_j q_k + s_i + q_k s_j + q_j s_k + 2 s_j s_k + q_j t + q_k t + 2 s_j t + s_k t + t^2 \\
m_{11}(i,j,k) &= m_{10}(i,j,k) + s'_k & &= q_j q_k + s_i + q_k s_j + q_j s_k + 2 s_j s_k + q_j t + q_k t + 2 s_j t + 2 s_k t + t^2 \\
m_{12}(i,j,k) &= m_7(i,j,k) + p'_k & &= q_i + s_i + q_k s_j + q_j s_k + 2 s_j s_k + t + q_j t + q_k t + 2 s_j t + 2 s_k t + 2 t^2 \\
m_{12}(i,j,k) &= m_{11}(i,j,k) + t'' & &= q_j q_k + s_i + q_k s_j + q_j s_k + 2 s_j s_k + t + q_j t + q_k t + 2 s_j t + 2 s_k t + 2 t^2
\end{aligned}$$

Note that the final new named variable $m_{12}(i,j,k)$ occurs on the left hand side of two equations. This gives it two different expanded expressions in terms of the q_i and slack variables, and those two expressions differ only by the identity $q_i = q_j q_k$.

The case of an elementary multiplication in which $j = k$, that is to say $q_i = q_j \cdot q_j$, is decomposed identically except that no variable $m_2(i,j,j)$ is defined; in its place the variable $m_1(i,j,j)$ with identical value is used. (The effect is the same as though $m_2(i,j,j)$ were defined as above, followed by an additional pass of the pre-processing phase that eliminates redundant variables.)

The complete set of new variables y_i allocated in the second pass is thus as follows:

- A global slack variable t and two new named variables t' and t'' derived from it.
- For each of the old variables q_i, an auxiliary variable p_i and a slack variable s_i.
- For each of the old variables q_j that ever occurs as a factor in an elementary multiplication equation $q_i = q_j \cdot q_k$, new named variables p'_j and s'_j.
- Four new intermediate quantities $a\_(i,j,k)$ for each of the original elementary additions $q_i = q_j + q_k$.
- Twelve (or in some cases eleven) new intermediate quantities $m\_(i,j,k)$ for each of the original elementary multiplications $q_i = q_j \cdot q_k$.

Theorem 4.13. *Given a set S of polynomial equations in multiple variables with integer coefficients, perform the first pass to produce a system of elementary addition equations and elementary multiplication equations together with a list of variables named q_i. Perform the pre-processing to remove obvious redundancies, and then perform the second pass to produce a second list of new variables named y_i.*

Suppose that $\mathbf{x}$ *is a solution of the system* S *of polynomial equations with integer coefficients, and let* $\mathbf{q}$ *be the uniquely determined set of values for the set of first-pass variables* q_i*. Let the slack variable* t *and the slack variables* s_i *be indeterminates. Let* $\mathbf{y} = \{y_i\}$ *denote the complete set of new variables from the second pass, each value of which depends only on* $\mathbf{q}$ *and a choice of values for the indeterminates* t *and* s_i*, and thus depends only on the particular solution* $\mathbf{x}$ *of* S *and the indeterminates. Then, for a generic choice of the indeterminates* t *and* s_i*, and in particular for at least one choice of these values, there will be no instance of*

1. *a second-pass variable* y_i *taking the value* 0*,*
2. *a second-pass variable* y_i *taking the value* 1*,*
3. *a pair of distinct second-pass variables* y_i *and* y_j *taking the same value, or*
4. *second-pass variables* y_i *and* y_j *(not necessarily distinct) such that* $y_i + y_j = 1$*.*

Proof. The setup of the claim can be interpreted as a game in which an adversary ("Player 1") selects the values for all q_i variables, and the other player ("Player 2") must find values for the s_i and slack variables t, t' to prevent any of the four listed conditions from holding.

In particular, there is a finite set of equations to check to ensure there are sufficient degrees of freedom in the q_i and slack variables. Player 1 succeeds with a choice that always satisfies one of these equations. For each such equation, consider the monomials in s_i variables, and check each for the presence of one or more q_i variables; Player 1 fails if there exists some monomial with no q_i factors. We describe an algorithm to check this condition.

- For each individual expression defining a m_i or a_i variable: check each monomial in s_i variables for the presence of one or more q_i term. If all monomials have at least one such term, continue; else return false. This corresponds to checking Condition 1 in Theorem 4.13: if there is at least one q_i factor in each monomial, then for any choice of s_i variables, there exists some assignment of q_i and slack variables such that the expression is not identically zero. Condition 2 follows identically from the same property.
- For each pair of expressions p_1 and p_2 defining m_i or a_i variables, there are three variable indices (i, j, and k) per expression, some subsets of which may be equal (indicating repeated variables).

  Some sets of repeated variables, however, do not occur due to the pre-processing step; in particular, there are no repetitions that lead to two identical q_i variables. For example, for a pair of elementary addition equations $q_{i_1} = q_j + q_k$ and $q_{i_2} = q_k + q_j$, the pre-processing step would consolidate q_{i_1} and q_{i_2} into one variable. This situation also arises if the indices $\{i, j, k\}$ used in the first elementary equation are a permutation of the indices used in the second. Therefore, we remove from consideration any such combination of variables.

  For each remaining configuration of the variable indices, consider the polynomial expressions $p_1 - p_2$ (corresponding to Condition 3 in Theorem 4.13) and $p_1 + p_2 - 1$

(corresponding to Condition 4). Again check each monomial in s_i variables for the presence of one or more q_i factor. If all monomials have at least one such factor, continue. Otherwise, return that Player 1 wins.

Upon reaching the end of the algorithm, indicating that all pairs satisfy the monomial condition, return that Player 2 wins, indicating that the conditions of the theorem hold.

We verify that the condition on monomial terms holds for all of the above expressions via computer algebra in SageMath; see [65] and the corresponding implementation in Appendix B. □

Chapter 5

Zero Forcing with Random Sets

Into the white [...]
Change will surely come.

- Caligula's Horse, *Into the White*

This chapter is based on the work in Curtis, Gan, Haddock, Lawrence, and Spiro [66], which is under submission and available on arXiv.

5.1 Introduction

While Chapter 4 showed that determining the exact minimum rank of a graph is computationally hard over the reals, a natural next question is whether it is possible to instead bound the minimum rank with some easier-to-compute quantity. The *zero forcing process*, first introduced by Burgarth and Giovannetti [24], provides one such lower bound.

To define the zero forcing process, first let G be a graph with vertex set $V(G)$, with each vertex initially colored either blue or white. If u is a blue vertex of G and the neighborhood $N_G(u)$ of u contains exactly one white vertex v, then we may change the color of v to blue. This iterated procedure for coloring a graph is called "zero forcing". A *zero forcing set* B is a subset of vertices of G such that, if G initially has all of the vertices of B colored blue, then the zero forcing process may eventually color all of $V(G)$ blue. We let $\mathrm{ZFS}(G)$ denote the set of all zero forcing sets of G. The *zero forcing number* $Z(G)$ is the minimum cardinality of a zero forcing set in G; that is, $Z(G) = \min_{B \in \mathrm{ZFS}(G)} |B|$.

The zero forcing number $Z(G)$ was found by an AIM research group [52] to be a bound for the maximum nullity of a graph G; or equivalently, a lower bound on the minimum rank of G. In this context, the zero forcing process can be interpreted as a series of deductions about constraints on null vectors of the adjacency matrix of G.

In addition to its independent mathematical interest and use as a tool for studying minimum rank, zero forcing has subsequently found many real-world applications, including

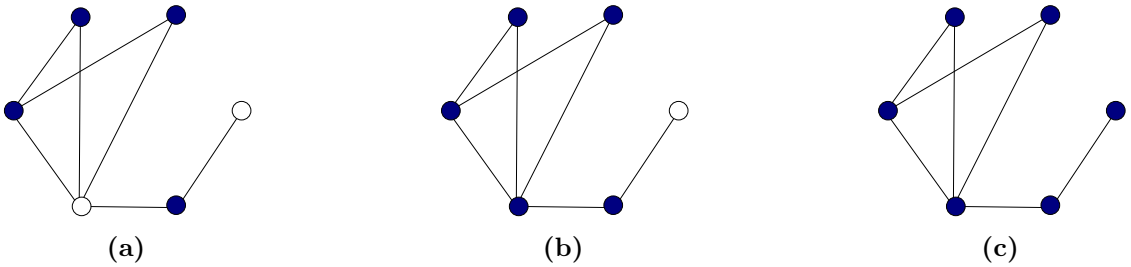


Figure 5.1: An example of zero forcing on a graph.

models of rumor spreading [41] and power grid domination [93], and as a result, zero forcing and many variants thereof have become an active area of research in recent years.

5.1.1 Zero Forcing with Randomness

In this chapter, we consider a randomized version of the zero forcing process. There are seemingly two natural ways to define such a random process: one can either use a deterministic set of blue vertices B together with forces that occur randomly, or one can use a random set of vertices B together with deterministic forces. The process known as *probabilistic zero forcing*, which was introduced by Kang and Yi [63], is of the former type and is by now well studied, see for example [28, 48, 73, 41, 60]. In this chapter we introduce a process of the latter type which we call *random set zero forcing*. This version of zero forcing can be interpreted as investigating the distribution of zero forcing sets among all subsets of the vertices in a graph. For example, in the application of power grid domination, random set zero forcing analyzes the *random initialization* scenario: if phase measurement units (PMUs) are placed on n randomly chosen nodes of an electric power network, determine the probability that the entire graph is monitored.

5.1.2 Main Results

Given a graph G and real number $0 \leq p \leq 1$, we define the random set $B_p(G) \subseteq V(G)$ by including each vertex of G independently and with probability p. For example, $B_1(G) = V(G)$, $B_0(G) = \emptyset$, and $B_{1/2}(G)$ is equally likely to be any subset of $V(G)$.

The central question we wish to ask is: Given G and p, what is (approximately) the probability that $B_p(G)$ is a zero forcing set of G? For example, one general bound we can prove is the following.

Proposition 5.1. *Let G be an n-vertex graph with minimum degree at least $\delta \geq 1$. For all p, we have*

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \leq \delta n p^{\delta}.$$

In fact, we prove a slightly stronger version of this theorem that holds for graphs with "few" vertices of degree less than δ; see Theorem 5.13. Proposition 5.1 can be viewed as a probabilistic analog of the basic fact that $Z(G) \geq \delta$ if G is a graph with minimum degree δ.

For many graphs G, it will happen that there exists a p such that $B_{p'}(G)$ is very unlikely to be a zero forcing set if p' is much smaller than p, and that $B_{p'}(G)$ is very likely to be zero forcing if p' is much larger than p; see for example Figure 5.2. This line of inquiry is motivated by the study of thresholds in random graphs, which is one of the fundamental topics in probabilistic combinatorics (see, for example, [47]). In fact, recalling that whether an initial vertex set B is zero forcing is a nontrivial monotone property on subsets of V, it follows from the foundational work of Bollobás and Thomason [18] that a threshold function for the zero forcing property in any given graph exists. With this in mind, we define the *threshold probability* $p(G)$ for an individual graph G to be the unique p such that $\Pr[B_p(G) \in \mathrm{ZFS}(G)] = \frac{1}{2}$.

Many other results from classical zero forcing also have probabilistic analogs for random set zero forcing. For example, it is straightforward to show that if G is an n-vertex graph, then $Z(G) \leq Z(\overline{K_n}) = n$ with equality if and only if $G = \overline{K_n}$. In the random setting, it is also easy to show the analogous result that for all n-vertex graphs G and $0 \leq p \leq 1$, we have

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \geq \Pr[B_p(\overline{K_n}) \in \mathrm{ZFS}(\overline{K_n})] = p^n,$$

with equality holding if and only if either $p \in \{0, 1\}$ or $G = \overline{K_n}$. In fact, we can use Observation 5.2 stated below to give the exact result $p(\overline{K_n}) = 2^{-1/n}$. Moreover, Observation 5.2 allows us to reduce our focus to connected graphs.

Observation 5.2. *Let G be the disjoint union of the graphs G_1 and G_2. Then*

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] = \Pr[B_p(G_1) \in \mathrm{ZFS}(G_1)] \cdot \Pr[B_p(G_2) \in \mathrm{ZFS}(G_2)].$$

Let G be a graph on $n \geq 2$ vertices with no isolated vertices. It is well known that every subset of $V(G)$ of size $n-1$ is a zero forcing set of G, and that $Z(G) = n-1$ if and only if $G = K_n$, the complete graph on n vertices (see, for example, [57]). With these observations it is not difficult to prove the following proposition (see Appendix 5.6).

Proposition 5.3. *If G is a graph on n vertices with no isolated vertices, then $p(G) \leq p(K_n)$. Moreover, $p(K_n) = 1 - \Theta(n^{-1})$.*

While it is straightforward to determine the graphs with the largest threshold probabilities, the analogous problem for smallest thresholds appears much harder. Intuitively, the path graph P_n is a natural candidate for the minimizer, since it is known that P_n is the unique n-vertex graph with zero forcing number 1. A proof of the following basic result can be found in Appendix 5.6.

Proposition 5.4. *The threshold probability of the path on n vertices satisfies*

$$p(P_n) = \Theta(n^{-1/2}).$$

In the classical setting, it is well known that amongst n-vertex graphs, the path P_n is the unique graph with the smallest zero forcing number. We conjecture that an analog of this result holds in the random setting.

Conjecture 5.5. *If G is an n-vertex graph and $0 \leq p \leq 1$, then*

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \leq \Pr[B_p(P_n) \in \mathrm{ZFS}(P_n)],$$

with equality holding if and only if either $p \in \{0, 1\}$ or $G = P_n$.

While we do not prove this conjecture in full, we provide some partial results; in particular, we prove the conjecture when restricted to trees and with n sufficiently large.

Theorem 5.6. *If T is an n-vertex tree with n sufficiently large, then for all $0 \leq p \leq 1$,*

$$\Pr[B_p(T) \in \mathrm{ZFS}(T)] \leq \Pr[B_p(P_n) \in \mathrm{ZFS}(P_n)],$$

with equality holding if and only if either $p \in \{0, 1\}$ or $T = P_n$.

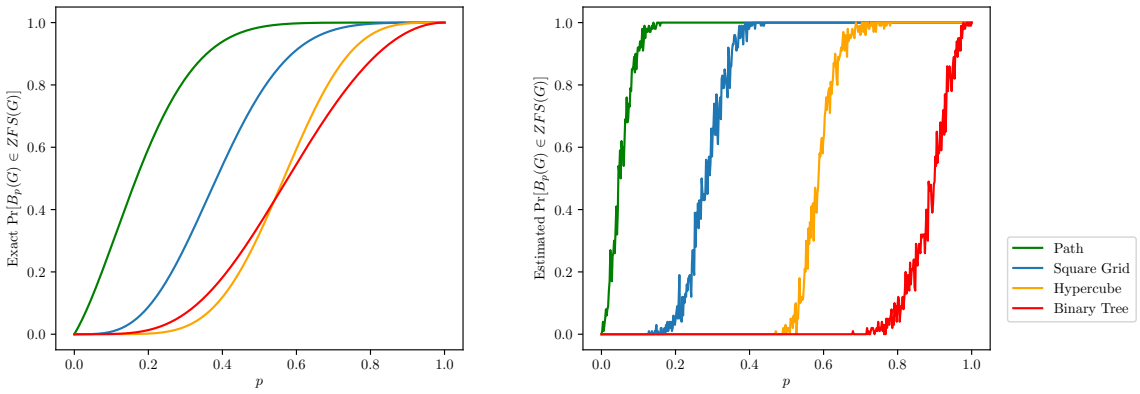


Figure 5.2: Exact (left) and Monte Carlo estimates (right) of $\Pr[B_p(G) \in \mathrm{ZFS}(G)]$ for the path, square grid, hypercube, and left-complete binary tree graphs on 16 and 256 vertices respectively.

We note that if Conjecture 5.5 were true, then in particular we would have $p(G) > p(P_n)$ for all n-vertex graphs $G \neq P_n$; we subsequently prove that this is true up to a constant factor.

Theorem 5.7. *If G is an n-vertex graph, then*

$$p(G) = \Omega(p(P_n)) = \Omega(n^{-1/2}).$$

In essence this result says that, for sufficiently large n, a random set of significantly less than $n^{1/2}$ vertices of any n-vertex graph G is very unlikely to be a zero forcing set.

Conjecture 5.5 can be viewed as a weakened version of a conjecture involving the number of zero forcing sets of a given size. To this end, we observe that if G is an n-vertex graph and $z(G;k)$ is the number of zero forcing sets of G of size k, then

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] = \sum_{k=1}^{n} z(G;k)p^k(1-p)^{n-k}. \tag{5.1}$$

The notation $z(G;k)$ follows that of Boyer et al. in [20] who introduced the study of zero forcing polynomials and found many explicit formulas for $z(G;k)$, including:

$$z(P_n;k) = \binom{n}{k} - \binom{n-k-1}{k}. \tag{5.2}$$

Observe that, by (5.1), Conjecture 5.5 is a weakened version of the following conjecture.

Conjecture 5.8 ([20])**.** *If G is an n-vertex graph, then for all k,*

$$z(G;k) \leq z(P_n;k) = \binom{n}{k} - \binom{n-k-1}{k}.$$

It was shown in [20] that Conjecture 5.8 holds whenever G contains a Hamiltonian path, but other than this very little is known. By extending our proof of Proposition 5.1, we prove Conjecture 5.8 whenever k is sufficiently small, as a function of the minimum degree of G.

Proposition 5.9. *If G is an n-vertex graph with minimum degree $\delta \geq 3$, then for all $k \leq (2\delta)^{-1/\delta} n^{1-1/\delta}$ we have $z(G;k) \leq z(P_n;k)$.*

We additionally show that this implies Conjecture 5.8 whenever G has sufficiently large minimum degree.

Corollary 5.10. *If G is an n-vertex graph with minimum degree $\delta \geq \log_2(n) + 2\log_2\log_2(n)$, then $z(G;k) \leq z(P_n;k)$ for all k.*

5.1.3 Organization and Notation

The remainder of this chapter is organized as follows. In Section 5.2, we provide a general bound on the probability that $B_p(G)$ is zero forcing given the minimum degree. In Section 5.3, we prove that the threshold probability for an n-vertex graph G is $\Omega(n^{-1/2})$. In Section 5.4, we prove Theorem 5.6, that amongst trees on sufficiently many vertices, paths have the largest probability of $B_p(G)$ being a zero forcing set. We conclude with some remarks and open questions in Section 5.5.

5.2 Bounds Using Degrees

In this warmup section, we give bounds on the probability that $B_p(G)$ is a zero forcing set in terms of the degree sequence of G. Our most general bound of this form is the following, where here and throughout $d(v)$ denotes the degree of v in the graph G.

Lemma 5.11. *Let G be an n-vertex graph with at least one edge and $p \in [0,1]$. Then*

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \leq \sum_{v \in V(G)} d(v) p^{d(v)}.$$

Proof. Let A be the event that $B_p(G) = V(G)$. For any $v \in V(G)$, let F_v be the event that v and exactly $d(v) - 1$ of its neighbors are in $B_p(G)$. We claim that for $B_p(G)$ to be a zero forcing set, either A or F_v for some v must occur. Indeed, if $B_p(G) \neq V(G)$ and $B_p(G)$ is a zero forcing set, then there must be some blue vertex v in $B_p(G)$ that forces a white vertex to be blue. In particular, if v is the first vertex which performs such a force, then it and exactly $d(v) - 1$ of its neighbors must be in $B_p(G)$. This proves our claim. Thus by the union bound we have

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \leq \Pr[A \cup \bigcup F_v] \leq \Pr[A] + \sum_{v \in V(G)} \Pr[F_v].$$

As each vertex is included in $B_p(G)$ independently and with probability p, we have

$$\Pr[F_v] = p \cdot \binom{d(v)}{d(v)-1} p^{d(v)-1}(1-p) = d(v) p^{d(v)}(1-p).$$

Plugging this into the bound above and using $\Pr[A] = p^n$ gives

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \leq p^n + \sum_{v \in V(G)} d(v) p^{d(v)}(1-p). \tag{5.3}$$

By assumption, G contains a vertex u with $d(u) \geq 1$. For this vertex we have

$$d(u)p^{d(u)}(1-p) = d(u)p^{d(u)} - d(u)p^{d(u)+1} \leq d(u)p^{d(u)} - p^n,$$

where this last step used $d(u) \geq 1$ and $d(u) + 1 \leq n$ (which always holds for n-vertex graphs). Plugging this bound into (5.3), and using the bound $d(v)p^{d(v)}(1-p) \leq d(v)p^{d(v)}$ for every other term of the sum gives the desired result. □

We also make use of the following, which can be proven using calculus.

Observation 5.12. *If d is a positive integer and $p \leq e^{-1/d}$, then*

$$xp^x \leq dp^d \ \forall\ x \geq d.$$

This result quickly gives Proposition 5.1, which we restate below.

Proposition 5.1. *Let G be an n-vertex graph with minimum degree at least $\delta \geq 1$. For all p, we have*

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \leq \delta n p^{\delta}.$$

Proof. When $G = K_2$ the theorem is equivalent to $2p(1-p) + p^2 \leq 2p$, i.e. that $-p^2 \leq 0$, so the result holds. From now on we assume G has at least 3 vertices. For all p and $n \geq 3$, we have

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \leq e^{-1}\delta n$$

since $e^{-1}\delta n \geq 1$. This implies the result when $p \geq e^{-1/\delta}$.

Observation 5.12 together with Lemma 5.11 and the fact that $d(v) \geq \delta$ for all v gives

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \leq \sum_{v \in V(G)} d(v) p^{d(v)} \leq \delta n p^{\delta}. \qquad \square$$

We next prove a slightly stronger version of this result which holds for graphs with "few" vertices of degree less than a given degree d. Proposition 5.1 can also be seen as a corollary of the following theorem.

Theorem 5.13. *Let G be an n-vertex graph without isolated vertices. Suppose that there exist integers $1 \leq d \leq n$ and $N \geq 0$ such that G contains at most N^k vertices of degree k for all $1 \leq k < d$. Then for all $p \leq e^{-1/d}$, we have*

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \leq 4pN + dnp^d.$$

Proof. The result is trivial if $pN > \frac{1}{2}$, so we can assume $pN \leq \frac{1}{2}$. Using Observation 5.12 together with Lemma 5.11 and the assumptions on G, we find

$$\begin{aligned}\Pr[B_p(G) \in \mathrm{ZFS}(G)] &\leq \sum_{v \in V(G)} d(v) p^{d(v)} \leq \sum_{\substack{v \in V(G) \\ d(v) < d}} d(v) p^{d(v)} + \sum_{\substack{v \in V(G) \\ d(v) \geq d}} dp^d \\ &\leq \sum_{k=1}^{d-1} k(pN)^k + dnp^d \leq \sum_{k=1}^{\infty} k(pN)^k + dnp^d.\end{aligned}$$

Note that in general we have $\sum_{k=1}^{\infty} kc^k = \frac{c}{(1-c)^2}$ provided $|c| < 1$. Applying this with $c = pN \leq \frac{1}{2}$ gives the desired result. $\square$

Recall that $z(G;k)$ is the number of zero forcing sets of G of size k. We now prove analogs of these results for $z(G;k)$.

Lemma 5.14. *Let G be an n-vertex graph with at least one edge. Then for all non-negative integers k,*

$$z(G;k) \leq \sum_{v \in V(G)} d(v) \binom{n - d(v)}{k - d(v)}.$$

Proof. The result is trivial if $k = n$. For $k < n$, every zero forcing set S must contain some vertex v of positive degree and exactly $d(v) - 1$ of its neighbors in order to have a vertex force. Thus every zero forcing set of size k can be constructed by first including a vertex v, then including exactly $d(v) - 1$ of its neighbors, then arbitrarily including $k - d(v)$ additional vertices. In total the number of ways to construct such a set is

$$\sum_{v \in V(G)} \binom{d(v)}{d(v)-1}\binom{n-d(v)}{k-d(v)} = \sum_{v \in V(G)} d(v)\binom{n-d(v)}{k-d(v)},$$

so G has at most this many zero forcing sets of size k. □

We next need the following lower bound on $z(P_n; k)$.

Lemma 5.15. *For all non-negative integers k we have*

$$z(P_n; k) \geq \frac{k^2}{n+k^2}\binom{n}{k}.$$

Proof. Recall from (5.2) that $z(P_n; k) = \binom{n}{k} - \binom{n-k-1}{k}$ for all k. Observe that

$$\begin{aligned}\frac{\binom{n-k-1}{k}}{\binom{n}{k}} =& \frac{(n-k-1)(n-k-2)\cdots(n-2k)}{n(n-1)\cdots(n-k+1)} = \prod_{i=0}^{k-1}\left(1 - \frac{k+1}{n-i}\right) \\ &\leq \left(1 - \frac{k}{n}\right)^k \leq \frac{1}{1+k^2/n} = \frac{n}{n+k^2},\end{aligned}$$

where the last inequality follows from the Bernoulli inequality: $(1-x)^n < \frac{1}{(1+x)^n} < \frac{1}{1+nx}$ for $x \in (0,1)$ and $n > 0$, where x in this case is $\frac{k}{n}$. This implies

$$z(P_n; k) = \binom{n}{k} - \binom{n-k-1}{k} \geq \left(1 - \frac{n}{n+k^2}\right)\binom{n}{k} = \frac{k^2}{n+k^2}\binom{n}{k}.$$ □

We now prove Proposition 5.9, which we restate below.

Proposition 5.9. *Let G be an n-vertex graph with minimum degree $\delta \geq 3$ and $k \leq (2\delta)^{-1/\delta} n^{1-1/\delta}$. Then $z(G; k) \leq z(P_n; k)$.*

Proof. By (5.2), for $k \geq n/2$ we have $z(P_n; k) = \binom{n}{k}$. Thus we may assume throughout that $k \leq n/2$.

Observe that for all t,

$$\binom{n-t}{k-t}\Big/\binom{n}{k} = \frac{k(k-1)\cdots(k-t+1)}{n(n-1)\cdots(n-t+1)} \leq (k/n)^t,$$

with this last step using the fact that $(k-i)/(n-i) \leq k/n$ for $i \geq 1$ if and only if $k \leq n$. Using this and Lemma 5.14 gives

$$z(G;k) \leq \sum_v d(v)(k/n)^{d(v)} \binom{n}{k}.$$

Because $\delta \geq 3$, we have $k \leq e^{-1/\delta} n$, so by Observation 5.12 we have

$$z(G;k) \leq \delta n (k/n)^{\delta} \binom{n}{k}. \tag{5.4}$$

First consider the case $k \leq \sqrt{n}$. By (5.4) and Lemma 5.15, to prove $z(G;k) \leq z(P_n;k)$, it suffices to have $\delta n(k/n)^{\delta} \leq k^2/2n$, or equivalently $n/k \geq (2\delta)^{1/(\delta-2)}$. Since $k \leq \sqrt{n}$, it suffices to prove $n \geq (2\delta)^{2/(\delta-2)}$, and this is true for $3 \leq \delta \leq n$ and $n \geq 5$. Thus we may assume $k \geq \sqrt{n}$. In this case (5.4) and Lemma 5.15 imply $z(G;k) \leq z(P_n;k)$ provided

$$\delta n (k/n)^{\delta} \leq \frac{1}{2},$$

and this holds precisely when $k \leq (2\delta)^{-1/\delta} n^{1-1/\delta}$. □

With Proposition 5.9 we can prove Corollary 5.10, which we restate below.

Corollary 5.10. *Let G be an n-vertex graph with minimum degree $\delta \geq \log_2(n) + 2\log_2\log_2(n)$. Then $z(G;k) \leq z(P_n;k)$ for all k.*

Proof. The result trivially holds for $k \geq n/2$, so it suffices to prove the result for $k \leq n/2$. By Proposition 5.1, it suffices to show

$$n/2 \leq (2\delta)^{-1/\delta} n^{1-1/\delta},$$

or equivalently $n \leq 2^{\delta}(2\delta)^{-1}$. And indeed, for $n \geq 9$ the minimum degree condition implies

$$2^{\delta}(2\delta)^{-1} \geq n \frac{(\log_2(n))^2}{2(\log_2(n) + 2\log_2\log_2(n))} \geq n.$$

For $n \leq 8$ one can check that $\lceil \log_2(n) + 2\log_2\log_2(n) \rceil \geq n-1$, so our hypothesis on δ implies G is complete and the result is immediate. In either case we conclude the result. □

5.3 Bounds on Threshold Probabilities

In this section we prove that for any n-vertex graph G, the threshold probability $p(G)$ is asymptotically at least that of P_n, i.e. $p(G) = \Omega(n^{-1/2})$. At a high level, our proof revolves around finding a graph $\tilde{G}$ which has minimum degree 2 and $p(\tilde{G}) \approx p(G)$. Because $\tilde{G}$ has

minimum degree 2, Proposition 5.1 implies $p(G) \approx p(\tilde{G}) = \Omega(n^{-1/2})$. We begin with a preliminary result regarding graphs containing pendant paths.

We say that a path $v_1 \cdots v_k$ in a graph G is a *pendant path*[1] provided $k \geq 2$, $d_G(v_1) = 1$, $d_G(v_i) = 2$ for $1 < i < k$, and $d_G(v_k) > 2$ (where $d_G(v)$ is the degree of v in G). We refer to the vertex v_1, the vertex of degree one, as the *pendant vertex*, and to v_k, the vertex of degree at least 3, as the *anchor vertex*. Observe that the only tree that does not contain a pendant path is the path graph.

Lemma 5.16. *Let G be an n-vertex graph. If there exists a vertex $w \in V(G)$ that is the anchor of two distinct pendant paths in G, then $p(G) = \Omega(n^{-1/2})$.*

Proof. Let $w \in V(G)$ and assume that w is the anchor of two distinct pendant paths in G, i.e., there exist distinct pendant paths $u_1 \cdots u_k w$ and $u_{k+1} \cdots u_\ell w$ in G. Let

$$I = \{j \in \mathbb{Z} : 1 < j < k \text{ or } k+1 < j < \ell\}$$

and for each $i \in I$, let A_i be the event that $u_i, u_{i+1} \in B_p(G)$. Let A' be the event that $B_p(G) \cap \{u_1, u_k, u_{k+1}, u_\ell\} \neq \emptyset$. Observe that if $B_p(G) \in \mathrm{ZFS}(G)$, then A' or some A_i event occurs. Thus,

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \leq \Pr\left[\bigcup_{i \in I} A_i \cup A'\right] \leq \Pr[A'] + \sum_{i \in I} \Pr[A_i] \leq 4p + np^2.$$

Thus to have $\Pr[B_p(G) \in \mathrm{ZFS}(G)] = \frac{1}{2}$, we must have $np^2 + 4p \geq \frac{1}{2}$, which implies $p = \Omega(n^{-1/2})$. □

With this lemma, we see that when proving Theorem 5.7 we may assume each vertex is the anchor of at most one pendant path. The next lemma allows us to assume that none of these paths are too long (unless G consists of a single path). In order to prove the next lemma, we recall various definitions and notation related to forcing chains which can be found, for example, in [12].

Let G be a graph and $B \subseteq V(G)$. Using B as the initial set of blue vertices, apply the color change rule and record the forces. If a vertex v forces u we write $v \to u$. The *chronological list of forces* $\mathcal{F}$ is the ordered list of forces, written in the order they were performed, that produces the final coloring generated by B in G. We shall sometimes use $\mathcal{F}$ to denote the *set of forces* that produces the final coloring generated by B in G. A *forcing chain* of $\mathcal{F}$ is a sequence of vertices $(v_1, \ldots, v_k)$ such that $v_i \to v_{i+1}$ for $1 \leq i \leq k-1$. A *maximal forcing chain* of $\mathcal{F}$ is a forcing chain that is not a proper subsequence of another forcing chain of $\mathcal{F}$. The *reversal* of B for $\mathcal{F}$ is the set of all vertices that do not perform a force, i.e., the set of all vertices that are the last element in a maximal forcing chain of $\mathcal{F}$.

[1] Most authors do not impose any conditions on the degree of v_k in the definition of a pendant path, but this formulation will be more useful to us.

Lemma 5.17. *Let G be an n-vertex graph and let $\{v_1, \ldots, v_M\}$ denote a set of vertices of degree 1 in G. Let $\tilde{G}$ be the graph obtained from G by adding a clique on $\{v_1, \ldots, v_M\}$. Then*

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] \leq \Pr[B_p(\tilde{G}) \in \mathrm{ZFS}(\tilde{G})] + pM.$$

Proof. We begin by showing that if $B \in \mathrm{ZFS}(G)$ and $B \notin \mathrm{ZFS}(\tilde{G})$, then $v_i \in B$ for some $i = 1, \ldots, M$. Let $B \subseteq V(G)$ and suppose that $v_i \notin B$ for all i. Assume that $B \in \mathrm{ZFS}(G)$ and let $\mathcal{F}$ be the set of forces for B in G. Since each v_i is a pendant vertex in G and each $v_i \notin B$, the set $\{v_1, \ldots, v_M\}$ is contained in the reversal of B for $\mathcal{F}$. This, and the fact that the neighborhood of each vertex in $V(G) \setminus \{v_1, \ldots, v_M\}$ is unchanged by adding a clique to $\{v_1, \ldots, v_M\}$, implies $\mathcal{F}$ is a set of forces for B in $\tilde{G}$. Thus, $B \in \mathrm{ZFS}(\tilde{G})$ and hence we have shown that if $B \in \mathrm{ZFS}(G)$ and $B \notin \mathrm{ZFS}(\tilde{G})$, then $v_i \in B$ for some $i = 1, \ldots, M$.

Let A_i be the event that $v_i \in B_p(G)$. By the preceding argument and the union bound,

$$\Pr[B_p(G) \in \mathrm{ZFS}(G) \wedge B_p(G) \notin \mathrm{ZFS}(\tilde{G})] \leq \Pr\left[\cup_{i=1}^{M} A_i\right] \leq pM.$$

We also have

$$\begin{aligned}
&\Pr[B_p(G) \in \mathrm{ZFS}(G) \wedge B_p(G) \notin \mathrm{ZFS}(\tilde{G})] \\
&= \Pr[B_p(G) \in \mathrm{ZFS}(G)] - \Pr[B_p(G) \in \mathrm{ZFS}(G) \wedge B_p(G) \in \mathrm{ZFS}(\tilde{G})] \\
&\geq \Pr[B_p(G) \in \mathrm{ZFS}(G)] - \Pr[B_p(G) \in \mathrm{ZFS}(\tilde{G})].
\end{aligned}$$

Combining both inequalities gives the result. □

The 2-*core* of a graph G, denoted $C_2(G)$, is obtained from G by repeatedly removing all isolated vertices and all vertices of degree 1 from G until no further removals are possible. See [16] for basic facts about 2-cores. We say that T is a *pendant tree* of a graph G if T is a maximal induced subgraph of G such that T is a tree, and if there exists a unique vertex $w \in V(T)$ contained in $C_2(G)$. The vertex w is called the *anchor vertex* of T. It is known that a vertex v is in $C_2(G)$ if and only if v is contained in a cycle or a path between cycles. Thus, the 2-core of G can be obtained by removing all non-anchor vertices of each pendant tree and all components of G that are trees.

Let G be a graph and $B \subseteq V(G)$. We define $C_2(B, G)$ to be the set of vertices that are either contained in $B \cap C_2(G)$ or are anchor vertices of a pendant tree T such that $B \cap V(T)$ is a zero forcing set of T. When G is clear from context we simply write $C_2(B)$. These definitions are illustrated in Figure 5.3. The motivation for these definitions is found in Lemma 5.19.

Before proving our next lemma, we note the following observation about zero forcing on graphs with a cut vertex. Observation 5.18 follows from some of the concepts introduced in [79]. We write $G[S]$ to denote the induced subgraph of the graph G on $S \subseteq V(G)$.

Observation 5.18. *Let G be a graph with cut vertex w, and let $W_1 \cup W_2 \cup \{w\}$ be a partition of $V(G)$ such that W_1 and W_2 are the disjoint union of connected components of $G - w$. Let $G_i = G\big[V(W_i) \cup \{w\}\big]$ for $i = 1, 2$. Then $B \cap V(G_i) \in \mathrm{ZFS}(G_i)$ for some index i, and $B \cap V(G_i) \cup \{w\}$ is a zero forcing set of G_i for each $i = 1, 2$.*

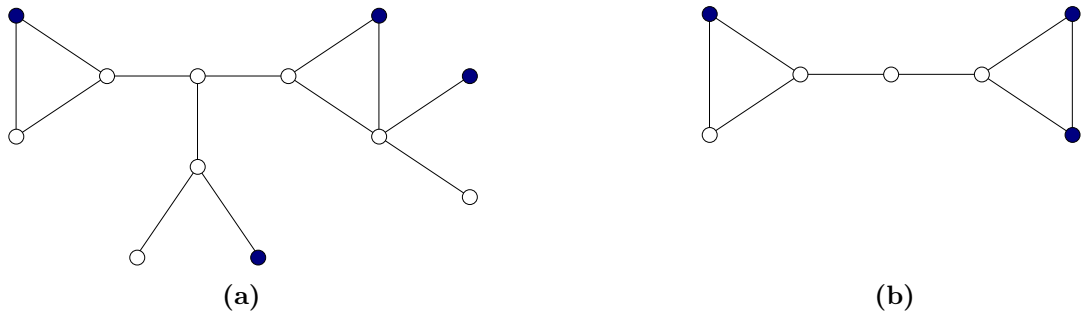


Figure 5.3: (a) Graph G with zero forcing set B colored blue. (b) Graph $C_2(G)$ with zero forcing set $C_2(B)$ colored blue.

Lemma 5.19. *Let G be a graph and $B \subseteq V(G)$. If B is a zero forcing set of G, then $C_2(B)$ is a zero forcing set of $C_2(G)$.*

Proof. Assume for contradiction that there exists a pair (B, G) such that B is a zero forcing set of G and $C_2(B)$ is not a zero forcing set of $C_2(G)$. Moreover, choose a minimal counterexample (B, G) such that G has as few vertices as possible.

We begin with a few observations to simplify the proof. The 2-core of G is the disjoint union of the 2-cores of each connected component of G. If $C_2(G)$ is the null graph, then it is vacuously true that $C_2(B)$ is a zero forcing set of $C_2(G)$. If $C_2(G) = G$, then $C_2(B) = B$ and hence $C_2(B)$ is a zero forcing set of $C_2(G)$. We may therefore assume that G is connected and that G contains a pendant tree.

Let T be a pendant tree of G with anchor vertex w, and let G_T be the induced subgraph of G on $(V(G) \setminus V(T)) \cup \{w\}$. Let

$$B_T = \begin{cases} B \cap V(G_T) & \text{if } B \cap V(T) \notin \mathrm{ZFS}(T) \\ (B \cap V(G_T)) \cup \{w\} & \text{if } B \cap V(T) \in \mathrm{ZFS}(T). \end{cases}$$

Since w is a cut vertex, Observation 5.18 implies that B_T is a zero forcing set of G_T. By our assumption of (B, G) being a vertex minimal counterexample, we have that $C_2(B_T, G_T)$ is a zero forcing set of $C_2(G_T)$ since G_T has strictly fewer vertices than G. Observing that $C_2(G_T) = C_2(G)$ and $C_2(B_T, G_T) = C_2(B, G)$, we have that $C_2(B, G)$ is a zero forcing set of $C_2(G)$. This gives a contradiction, proving the result. □

We can now prove the main result of this section, which we restate below.

Theorem 5.7. *If G is an n-vertex graph, then*

$$p(G) = \Omega(p(P_n)) = \Omega(n^{-1/2}).$$

Proof. Observe that if H is a connected component of G, then $p(G) \geq p(H)$. Also, by Lemma 5.16, if G is a tree that is not a path, or if G has a pendant tree that is not a pendant path,

then $p(G) = \Omega(n^{-1/2})$. We may therefore assume that G is connected, G contains a cycle, and every pendant tree of G is a pendant path. Note that the definition of pendant trees implies every vertex v is the anchor of at most one pendant path in G, and that every anchor vertex is contained in $C_2(G)$. Let $p = cn^{-1/2}$, where $c < 1$ is a positive constant which we specify later.

Let $\{P_1, \dots, P_M\}$ be the set of all pendant paths in G on at least $100n^{1/2} + 1$ vertices, and let $v_i \in P_i$ denote the vertex of P_i of degree 1. Let $\tilde{G}$ be the graph obtained from G by adding a clique on $\{v_1, \dots, v_M\}$. Observe that $M(100n^{1/2} + 1) \leq n$, and hence $M < n^{1/2}/100$. Thus by Lemma 5.17,

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] < \Pr[B_p(\tilde{G}) \in \mathrm{ZFS}(\tilde{G})] + .01.$$

Observe that $C_2(\tilde{G})$ is nonempty since G contains a cycle, and by Lemma 5.19,

$$\Pr[B_p(\tilde{G}) \in \mathrm{ZFS}(\tilde{G})] \leq \Pr[C_2(B_p(\tilde{G})) \in \mathrm{ZFS}(C_2(\tilde{G}))].$$

Thus it suffices to prove $\Pr[C_2(B_p(\tilde{G})) \in \mathrm{ZFS}(C_2(\tilde{G}))] < 0.49$ for n sufficiently large.

For $v \in V(C_2(\tilde{G}))$, let A_v denote the event that $v \in C_2(B_p(\tilde{G}))$. We claim that

$$\Pr[A_v] \leq 2p + (100n^{1/2})p^2 = (2c + 100c^2)n^{-1/2} := q. \tag{5.5}$$

Indeed, $\Pr[A_v] = p$ if v is not the anchor of some pendant path. If v is the anchor of the pendant path P, then for A_v to occur, either an endpoint of P or two consecutive vertices of P must be in $B_p(\tilde{G})$, and a union bound gives the result since P is assumed to have length at most $100n^{1/2} + 1$.

Because the A_v events are independent of each other, our bound above implies that

$$\Pr[C_2(B_p(\tilde{G})) \in \mathrm{ZFS}(C_2(\tilde{G}))] \leq \Pr[B_q(\tilde{G}) \in \mathrm{ZFS}(C_2(\tilde{G}))].$$

Since $C_2(\tilde{G})$ has at most n vertices and minimum degree at least 2, Proposition 5.1 implies

$$\Pr[B_q(\tilde{G}) \in \mathrm{ZFS}(C_2(\tilde{G}))] \leq 2nq^2.$$

Taking $c = 1/17$ and recalling the definition of q in (5.5) gives the desired result. □

5.4 Bounds for Trees

In this section we prove $\Pr[B_p(T) \in \mathrm{ZFS}(T)] \leq \Pr[B_p(P_n) \in \mathrm{ZFS}(P_n)]$ whenever T is an n-vertex tree with n sufficiently large. We will break our proof into two cases, namely when $p = \Omega(n^{-1})$ and $p = O(n^{-1})$. The intuition for this choice is that when $p \ll n^{-1}$, the probability that $B_p(P_n)$ is zero forcing is roughly the probability of choosing an endpoint, while for $p \gg n^{-1}$ it is roughly the probability of choosing two consecutive vertices. Therefore, we will need two different arguments for these two regimes.

5.4.1 Large p

The following provides a concrete statement agreeing with the intuition outlined above.

Lemma 5.20. *Let $v_1, \dots, v_n$ be vertices of a graph G. If $\frac{4}{n-2} < p < 1$ and $n \geq 11$, then*

$$\Pr[v_1 \in B_p(G) \vee v_n \in B_p(G)] < \Pr[v_i, v_{i+1} \in B_p(G) \text{ for some } i].$$

Proof. Define

$$p_e := \Pr[v_1 \in B_p(G) \vee v_n \in B_p(G)] = 1 - (1-p)^2,$$
$$p_m := \Pr[v_i, v_{i+1} \in B_p(G) \text{ for some } i] \geq 1 - (1-p^2)^{\lfloor \frac{n-1}{2} \rfloor},$$

where this last inequality holds because p_m is strictly more than the probability of having at least one of the pairs (v_i, v_{i+1}) with i odd in $B_p(G)$; see the proof of Proposition 5.4 for a more formal argument. Thus it is sufficient to show that

$$(1-p^2)^k < (1-p)^2, \tag{5.6}$$

where $k := \lfloor \frac{n-1}{2} \rfloor$, for $n \geq 18$. Using the standard bounds, $\frac{x}{1+x} < \ln(1+x) < x$ for $|x| < 1$, and so

$$-k \ln(1-p^2) > kp^2 \quad \text{and} \quad \frac{2p}{(1-p)} > -2\ln(1-p).$$

Thus, Equation (5.6) follows if $kp^2 > \frac{p}{1-p}$, or equivalently, if $kp^3 - kp^2 + p < 0$. Since this polynomial has roots at $p = 0$ and $p = \frac{k \pm \sqrt{k(k-4)}}{2k}$, we have that $kp^3 - kp^2 + p < 0$ in the range $p \in \left(\frac{2}{k}, 1 - \frac{2}{k}\right)$, and thus, for $p \in \left(\frac{4}{n-2}, 1 - \frac{4}{n-2}\right)$ if $n \geq 10$. Finally, we can check directly that for $n \geq 11$, $(1-p^2)^{\frac{n-2}{2}} < (1-p)^2$ for all $p \in (0.55, 1)$, and this range includes $(1 - \frac{4}{n-2}, 1)$ for all $n \geq 11$.

□

Analogous to Proposition 5.16, we can show that graphs with a vertex at the end of two short pendant paths are harder to zero force than paths.

Lemma 5.21. *Let G be an n-vertex graph that has a vertex w which is the endpoint of two pendant paths $u_1 \cdots u_s w$ and $v_1 \cdots v_t w$. If $p \geq 8/(n-s-t)$ and $n - s - t \geq 14$, then*

$$\Pr[B_p(G) \in \mathrm{ZFS}(G)] < \Pr[B_p(P_n) \in \mathrm{ZFS}(P_n)].$$

Proof. Let $w_1, \dots, w_r$ be an arbitrary ordering of $V(G) \setminus \{u_1, \dots, u_s, v_1, \dots, v_t\}$. Relabel the vertices of P_n so that its vertices along the path are $u_1 \cdots u_s w_1 \cdots w_r v_t \cdots v_1$. Because $V(G) = V(P_n)$, we can couple our random variables so that $B_p := B_p(G) = B_p(P_n)$. Let $F = B_p \cap \{u_1, \dots, u_{s-1}, v_1, \dots, v_{t-1}\}$. It suffices to show for all $S \subseteq \{u_1, \dots, u_{s-1}, v_1, \dots, v_{t-1}\}$ that

$$\Pr[B_p(G) \in \mathrm{ZFS}(G) | F = S] \leq \Pr[B_p(P_n) \in \mathrm{ZFS}(P_n) | F = S], \tag{5.7}$$

with strict inequality for at least one such set. If S contains two consecutive vertices u_i and u_{i+1}, two consecutive vertices v_j and v_{j+1}, or u_1 or v_1, then $B_p \in \mathrm{ZFS}(P_n)$ so Equation (5.7) holds trivially. Thus from now on we can assume this is not the case.

With this assumption, the vertex u_s in G can only be colored blue by B_p if at least one of u_s or v_t is in B_p (this is because u_s is adjacent to u_{s-1}, which does not enact any forces by assumption on S, and to w, which can only enact a force if at least one of u_s, v_t are colored blue at some point). On the other hand, B_p will be a zero forcing set for P_n provided B_p contains two consecutive vertices from $\{u_s, w_1, \cdots, w_r, v_t\}$.

By applying Lemma 5.20 to the vertex set $\{u_s, w_1, \ldots, w_r, v_t\}$, we see that if $p > \frac{8}{n-s-t+2}$ and $n - s - t + 2 \geq 16$, then

$$\Pr[B_p \text{ contains } u_s \text{ or } v_t] < \Pr[B_p \text{ contains a consecutive pair from } \{u_s, w_1, \cdots, w_r, v_t\}].$$

As these two events are independent given the random set F, we conclude that for $p \geq 8/(n-s-t) > 8/(n-s-t+2)$ and $n - s - t + 2 \geq 16$,

$$\Pr[B_p(G) \in \mathrm{ZFS}(G) | F = S] < \Pr[B_p(P_n) \in \mathrm{ZFS}(P_n) | F = S],$$

and from this we conclude the result. □

With this we can prove Theorem 5.6 for the case $p = \Omega(n^{-1})$.

Proposition 5.22. *If $T \neq P_n$ is an n-vertex tree with $n \geq 42$, and if $\frac{24}{n} < p < 1$, then*

$$\Pr[B_p(T) \in \mathrm{ZFS}(T)] < \Pr[B_p(P_n) \in \mathrm{ZFS}(P_n)].$$

Proof. Let u, v be any two leaves of T which are at a shortest distance from each other. Observe that the path between u, v consists of two pendant paths, say $uu_2, \cdots u_s w$ and $vv_2, \cdots v_t w$. Because T is not a path, there either exists exactly one leaf $\ell \neq u, v$, or at least two leaves $i, j \neq u, v$.

Suppose for contradiction that $s + t > \frac{2n}{3}$. If T has exactly three leaves, then

$$d(\ell, u) = d(\ell, w) + s < n/3 + s,$$

where this inequality used the fact that none of the internal vertices along the path from ℓ to w use any of the vertices along the path from u to v, of which there are more than $2n/3 + 1$. By a symmetric argument we have $d(\ell, v) < n/3 + t$. In particular, we must have

$$d(\ell, u) + d(\ell, v) < 2n/3 + s + t < 2(s + t) = 2d(u, v),$$

and hence at least one of $d(\ell, u), d(\ell, v)$ is smaller than $d(u, v)$, a contradiction to our choice of u, v. Similarly if T has at least four leaves, then $d(i, j) < \frac{n}{3}$, which again gives a contradiction. Thus we can assume $s + t \leq 2n/3$. With this, our hypothesis implies $p \geq \frac{8}{n-s-t}$ and $n - s - t + 2 \geq 16$, so we can apply Lemma 5.21 to give the desired result. □

5.4.2 Small p

We will prove our result for small p by upper bounding $z(T;k)$, which we recall is the number of zero forcing sets of T of size k.

Lemma 5.23. *If $T \neq P_n$ is an n-vertex tree, then*

$$z(T;k) \leq \frac{13k^4}{n^2}\binom{n}{k}.$$

Proof. Let Δ denote the maximum degree of T and ℓ the number of leaves of T. Observe that the zero forcing number of a graph is always at least the minimum number of paths needed to cover the vertices of the graph. In particular, every zero forcing set for the tree T has size at least $\ell/2$. It is also known (see for example [74]) that every zero forcing set for a tree T has size at least $\Delta - 1$. Thus for $k < \max\{\Delta - 1, \ell/2\}$, we have $z(T;k) = 0$ and the bound trivially holds. From now on we assume $\max\{\Delta - 1, \ell/2\} \leq k$. The bound is also trivial when $k = n$, so we may assume $k < n$. Lastly, we may also assume $\Delta \geq 3$ since T is not a path.

We first count the number of sets S of size k which have two pairs of vertices u, v and x, y with $u \sim v,\ x \sim y$ and $\{u, v\} \cap \{x, y\} = \emptyset$ (where $\sim$ denotes vertex adjacency). In this case the number of sets S is at most

$$(n-1)^2\binom{n-4}{k-4},$$

since one can choose each pair (which is just an edge in T) in at most $n - 1$ ways.

We next count the number of sets S of size k which contain three vertices u, v, w with $u \sim v \sim w$. The number of such S is at most

$$(n-1)(2\Delta-2)\binom{n-3}{k-3},$$

since one can first choose two adjacent vertices in $n - 1$ ways, then a third vertex which is adjacent to at least one of these in at most $2\Delta - 2$ ways, and then the remaining vertices in $\binom{n-3}{k-3}$ ways.

We next count the number of zero forcing sets S that contain no two adjacent vertices. Because $k < n$ and S is a zero forcing set, at least one vertex of S must be able to force. Because S contains no adjacent vertices, this is only possible if S contains a leaf. Choose such a leaf u_1 to include in S, which can be done in ℓ ways. Let $u_1u_2 \cdots u_s$ be the unique path in T with $\deg(u_i) = 2$ for $1 < i < s$ and $\deg(u_s) \neq 2$.

Claim 5.24. *The set S either contains a leaf $v \neq u_1$, or a neighbor of u_s other than u_{s-1}.*

Proof. Assume this were not the case. Because S contains no two adjacent vertices, no additional leaves, and no other neighbor of u_s, it is not difficult to see that the only vertices that will be colored blue by S are $S \cup \{u_2, \ldots, u_s\}$. Because S is a zero forcing

set, we must have $V(T) = S \cup \{u_2, \ldots, u_s\}$. However, by assumption the only leaves that could be in $S \cup \{u_2, \ldots, u_s\}$ are u_1 and u_s, but $T \neq P_n$ contains at least three leaves, so $V(T) \neq S \cup \{u_2, \ldots, u_s\}$, giving the desired contradiction. □

In total then, we see that the number of choices for such a set S is at most

$$\ell(\ell + \Delta - 1)\binom{n-2}{k-2},$$

where the terms in the expression above count the number of choices for u_1, followed by the number of choices for some additional leaf or neighbor of u_s, followed by the number of arbitrary sets of $k-2$ vertices.

It remains to count S that have exactly one pair of adjacent vertices. One can first choose the adjacent pair $u_1, v_1 \in S$ in at most $n-1$ ways. If $\deg(u_1) = 2$, then let $u_1 \cdots u_s$ be the unique path from u_1 with u_2 the neighbor of u_1 not equal to v_1, and with $\deg(u_i) = 2$ for all $i < s$ and $\deg(u_s) \neq 2$. If $\deg(u_1) \neq 2$, then we simply consider the 1-vertex path u_1. Analogously define the path $v_1 \cdots v_t$. As in the previous case, because S contains no other pair of adjacent vertices, it must contain at least one leaf or one neighbor of either u_s or v_t that is not u_{s-1} or v_{t-1}. In total then the number of choices for such an S is at most

$$(n-1)(\ell + 2\Delta - 2)\binom{n-3}{k-3}.$$

Putting together all the above cases, wee see that in total, $z(T;k)$ is at most

$$\begin{aligned}
&(n-1)^2\binom{n-4}{k-4} + (n-1)(2\Delta-2)\binom{n-3}{k-3} + \ell(\ell+\Delta-1)\binom{n-2}{k-2} + (n-1)(\ell+2\Delta-2)\binom{n-3}{k-3} \\
&\leq n^2 \cdot \frac{k^4}{n^4}\binom{n}{k} + n(2\Delta-2)\cdot\frac{k^3}{n^3}\binom{n}{k} + \ell(\ell+\Delta-1)\cdot\frac{k^2}{n^2}\binom{n}{k} + n(\ell+2\Delta-2)\cdot\frac{k^3}{n^3}\binom{n}{k},
\end{aligned}$$

where this last inequality used the fact that $n - 1 \leq n$ and that $\binom{n-t}{k-t} \leq (k/n)^t\binom{n}{k}$ for all integers $t \geq 0$. Using our assumptions $\Delta - 1 \leq k$ and $\ell \leq 2k$, we find that the above expression is at most $(1+2+6+4)\frac{k^4}{n^2}\binom{n}{k}$ as desired. □

With this we can prove the following.

Proposition 5.25. *For every $C > 0$, there exists an integer n_0 such that for all $n \geq n_0$, if $T \neq P_n$ is an n-vertex tree and $0 < p \leq \frac{C}{n}$, then*

$$\Pr[B_p(T) \in \mathrm{ZFS}(T)] < \Pr[B_p(P_n) \in \mathrm{ZFS}(P_n)].$$

Proof. By the previous lemma and the trivial bound $\binom{n}{k} \leq n^k/k!$, we have

$$\Pr[B_p(T) \in \mathrm{ZFS}(T)] \leq \sum_k z(T;k)p^k \leq \sum_k \frac{13k^4n^{k-2}}{k!}p^k \leq p \cdot 13Cn^{-1}\sum_k \frac{k^4C^{k-2}}{k!}.$$

The above sum is convergent, so for n sufficiently large we find

$$\Pr[B_p(T) \in \mathrm{ZFS}(T)] \leq \frac{1}{2}p \leq \Pr[B_p(P_n) \in \mathrm{ZFS}(P_n)],$$

where this latter inequality is strict provided $p > 0$. □

With Propositions 5.22 and 5.25 we can prove Theorem 5.6, which we restate below.

Theorem 5.6. *If T is an n-vertex tree with n sufficiently large, then for all $0 \leq p \leq 1$,*

$$\Pr[B_p(T) \in \mathrm{ZFS}(T)] \leq \Pr[B_p(P_n) \in \mathrm{ZFS}(P_n)],$$

with equality holding if and only if either $p \in \{0, 1\}$ or $T = P_n$.

Proof. The equality of the result trivially holds for either $p \in \{0, 1\}$ or $T = P_n$. If $T \neq P_n$ with n sufficiently large, by Proposition 5.22, the result holds for $24/n < p < 1$, and by Proposition 5.25 the result holds for $0 < p \leq 24/n$. □

5.5 Concluding Remarks

Our work described above suggests several open problems regarding the threshold probability $p(G)$, which we recall is the unique $p \in [0, 1]$ such that $\Pr[B_p(G) \in \mathrm{ZFS}(G)] = \frac{1}{2}$. For example, we conjecture the following refinement of Theorem 5.7.

Conjecture 5.26. *If G is an n-vertex graph which contains a clique of size k, then*

$$p(G) = \Omega(\sqrt{k/n}).$$

This conjecture can be viewed as a probabilistic analog of the classical result that $Z(G) \geq k$ if G has a clique of size k, which was proved by Butler and Young [25]. The motivation for the bound $\Omega(\sqrt{k/n})$ comes from considering a graph G which consists of a clique on k vertices, with each of these vertices attached to a path of length roughly n/k. For this graph, a given vertex of the clique will be forced by the path it is connected to with probability roughly $1 - e^{-p^2 n/k}$, so if p is much smaller than $\sqrt{k/n}$, then almost none of the clique vertices in G will be colored blue. Thus $p(G) = \Omega(\sqrt{k/n})$ in this case[2].

Another natural problem is to compute $p(G)$ for various natural families of graphs. For example, Table 5.1 summarizes the order of magnitude of $p(G)$ for many such families. However, one case for which we do not understand $p(G)$ is when G is the n-dimensional hypercube Q_n.

[2]In fact, a sharper analysis shows that $p(G) = \Omega(\sqrt{k \log(k)/n})$ for k not too large in terms of n. We suspect that Conjecture 5.26 can be strengthened to include this $\log(k)$ term, but for ease of presentation we have written the conjecture as is.

Table 5.1: Thresholds for graph families.

Family	Description	Threshold Probability
K_n	Complete graph on n vertices	$1 - \Theta(n^{-1})$
nK_1	Graph on n isolated vertices	$2^{-1/n}$
$K_{n_1,\cdots,n_k}$	Complete multipartite graph	$1 - \Theta_k(\min_i\{n_i^{-1}\})$
P_n	Path on n vertices	$\Theta(n^{-1/2})$
C_n	Cycle on n vertices	$\Theta(n^{-1/2})$
W_n	Wheel on n vertices	$\Theta(n^{-1/3})$

Problem 5.27. *Does there exist a constant c such that $p(Q_n) \sim c$? If so, what is this constant?*

Because $Z(Q_n) = 2^{n-1}$, we must have $c \geq .5$ if it exists, but beyond this we know nothing about c. The empirical plot in Figure 5.2 of the probability that $B_p(Q_8)$ is zero forcing suggests that c might be at least .58. Another family of graphs whose zero forcing properties we do not understand are grid graphs.

Problem 5.28. *Determine the order of magnitude of $p(P_m \Box P_n)$, where $P_m \Box P_n$ denotes the $m \times n$ grid graph.*

Assuming $2 \leq m \leq n$, we can apply Theorem 5.13 with $d = 4$ and $N \approx n^{1/3}$ to show $p(P_m \Box P_n) = \Omega(\min\{n^{-1/3}, (mn)^{-1/4}\})$. The best general upper bound we have is $p(P_m \Box P_n) = O(n^{-1/2m})$, since at this point it is fairly likely that $B_p(P_m \Box P_n)$ contains two consecutive P_m paths, which forces the entire graph. For small m we suspect that our upper bound is closer to the truth than our lower bound, but for large m the situation is unclear.

Another natural problem is to study how graph operations affect the threshold probability of a graph. It is perhaps intuitive that $p(C_n) \approx p(P_n)$ since C_n can be formed from P_n by adding a single edge. In fact, it is easily verified that $p(C_n) = \Theta(n^{-1/2})$ by using an argument similar to the proofs in Appendix 5.6. However, there are examples where this intuition fails. Indeed, let $v_1, \ldots, v_n$ be the vertices of P_n, and let R_n denote the graph obtained from P_n by adding the edge v_1v_3 (see Figure 5.4).

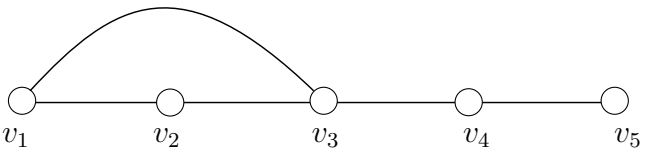


Figure 5.4: The triangle with pendant path on five vertices, R_5.

Observe that any zero forcing set of R_n must contain either v_1 or v_2. Thus,

$$\Pr\left[B_p(R_n) \in \mathrm{ZFS}(R_n)\right] \leq \Pr\left[v_1 \in B_p(R_n) \vee v_2 \in B_p(R_n)\right] \leq 2p.$$

This implies $p(R_n) \in [1/4, 1]$, and hence $p(R_n) = \Theta(1)$.

In the deterministic setting, it is well known that the zero forcing number $Z(G)$ of a graph G changes by at most one if a single edge or vertex is removed from G. This is far from true for $p(G)$. Indeed, recall that $p(P_n) = \Theta(n^{-1/2})$. Let P'_n be obtained by deleting the edge v_1v_2. Since P'_n has K_1 as a connected component, by Observation 5.2 we have $p(P'_n) \geq p(K_1) = \frac{1}{2}$. A similar result holds if one deletes v_2.

Similarly, deleting edges or vertices can dramatically decrease $p(G)$. Consider the triangle with pendant path R_n, which has $p(R_n) = \Theta(1)$. If one deletes v_1, then the resulting graph is P_{n-1}, which has $p(P_{n-1}) = \Theta(n^{-1/2})$. If one deletes the edge v_1v_3, then the resulting graph is P_n, which has $p(P_n) = \Theta(n^{-1/2})$.

Lastly, one could consider randomized versions of variants of the classical zero forcing number. For example, under so-called "skew zero forcing" (which was originally introduced in [4]), one can easily generalize Proposition 5.1 to give an upper bound of roughly $\delta np^{\delta-1}$ on the probability that a random starting set $B_p(G)$ is a skew zero forcing set for G. It would also be interesting to consider probabilistic zero forcing with a random set of vertices initially colored blue.

5.6 Appendix: Threshold Probability Calculations

In this Appendix we provide proofs of Propositions 5.3 and 5.4. We make use of the following inequalities. Recall that

$$1 - x \leq e^{-x} \tag{5.8}$$

for all real values x, and

$$\left(1 - \frac{c}{n}\right)^n \geq 1 - c \tag{5.9}$$

for $|c| \leq n$ and $n \geq 1$.

Proposition 5.3. *If G is a graph on n vertices with no isolated vertices, then $p(G) \leq p(K_n)$. Moreover, $p(K_n) = 1 - \Theta(n^{-1})$.*

Proof. The result is immediate for $n = 1$, so assume $n \geq 2$. Define

$$f(p) = n(1-p)p^{n-1} + p^n,$$

which is the probability that $B_p(G)$ contains at least $n-1$ vertices. Since every subset of $V(G)$ of size $n-1$ is a zero forcing set, we have for $p \in [0, 1]$,

$$\Pr\left[B_p(G) \in \mathrm{ZFS}(G)\right] \geq f(p) = \Pr\left[B_p(K_n) \in \mathrm{ZFS}(K_n)\right],$$

where this equality used the fact that a set S is a zero forcing set of K_n if and only if $|S| \geq n-1$. Since $\Pr[B_p(G) \in \mathrm{ZFS}(G)]$ and $\Pr[B_p(K_n) \in \mathrm{ZFS}(K_n)]$ are increasing functions of p, we conclude that $p(G) \leq p(K_n)$.

We now prove the asymptotic result. Let $p = 1 - c/n$, where $c \leq n$ is positive. By (5.9),

$$f(p) \geq p^n \geq 1 - c,$$

which implies $f(p) > 1/2$ if $p > 1 - \frac{1}{2n}$. Similarly, by (5.8),

$$f(p) = c(1-c/n)^{n-1} + (1-c/n)^n \leq ce^{-c+c/n} + e^{-c} \leq ce^{-c/2} + e^{-c},$$

where the last inequality holds since $n \geq 2$. Thus $f(p) < 1/2$ for $n \geq 5$ and $p < 1 - \frac{5}{n}$. We conclude $p(K_n) = 1 - \Theta(n^{-1})$. □

Proposition 5.4. *The threshold probability of the path on n vertices satisfies*

$$p(P_n) = \Theta(n^{-1/2}).$$

Proof. Let $v_1, \ldots, v_n$ denote the vertices of P_n. By convention, we assume the vertices $v_1, \ldots, v_n$ of P_n are in *path order*, i.e., the edges of P_n are $v_i v_{i+1}$ for $1 \leq i \leq n-1$. Note that $S \subseteq V(P_n)$ is a zero forcing set if and only if S contains an endpoint or S contains two consecutive vertices (see Figure 5.5).

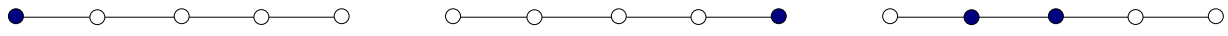


Figure 5.5: Three zero forcing sets for P_5

Define the random variable X to be the number of indices $i \in \{1, 2, \cdots, n-1\}$ such that $v_i, v_{i+1} \in B_p(P_n)$. Markov's inequality yields

$$\Pr[X \geq 1] \leq \mathbb{E}[X] = (n-1)p^2.$$

Since $B_p(P_n) \in \mathrm{ZFS}(P_n)$ if and only if either $X \geq 1$ or at least one of $v_1, v_n \in B_p(P_n)$, a union bound now implies

$$\Pr[B_p(P_n) \in \mathrm{ZFS}(P_n)] \leq (n-1)p^2 + 2p.$$

This quantity is less than $1/2$ provided $p = cn^{-1/2}$ for any $c < 1/4$. Thus $p(P_n) = \Omega(n^{-1/2})$.

Next, for $i \in \{1, 2, \ldots, n-1\}$, let A_i be the event that $v_i, v_{i+1} \in B_p(P_n)$, and define $A = \bigcup_{i \text{ odd}} A_i$. Then,

$$\begin{aligned}\Pr[B_p(P_n) \in \mathrm{ZFS}(P_n)] \geq \Pr[A] &= 1 - \prod_{i \text{ odd}} (1 - \Pr[A_i]) \\ &= 1 - (1-p^2)^{\lfloor (n-1)/2 \rfloor} \geq 1 - e^{-p^2 \lfloor (n-1)/2 \rfloor},\end{aligned}$$

where the first equality follows from the fact that these events are independent, and the last step uses (5.8). This probability will be greater than $1/2$ for $p = Cn^{-1/2}$ with C sufficiently large. We conclude that $p(P_n) = \Theta(n^{-1/2})$. □

Bibliography

[1] Mikkel Abrahamsen, Linda Kleist, and Tillmann Miltzow. "Training Neural Networks is ER-complete". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 18293–18306.

[2] Ethan Akin. *The geometry of population genetics*. Vol. 31. Springer Science & Business Media, 2013.

[3] Dan Alistarh et al. "Time-space trade-offs in population protocols". In: *Proceedings of the twenty-eighth annual ACM-SIAM symposium on discrete algorithms*. SIAM. 2017, pp. 2560–2579.

[4] Mary Allison et al. "Minimum rank of skew-symmetric matrices described by a graph". In: *Linear Algebra Appl.* 432.10 (2010). IMA-ISU research group on minimum rank, pp. 2457–2472. ISSN: 0024-3795. DOI: 10.1016/j.laa.2009.10.001. URL: https://doi.org/10.1016/j.laa.2009.10.001.

[5] Nima Anari, Shayan Oveis Gharan, and Cynthia Vinzant. "Log-concave polynomials, entropy, and a deterministic approximation algorithm for counting bases of matroids". In: *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2018, pp. 35–46.

[6] David Anderson et al. "Tier structure of strongly endotactic reaction networks". In: *arXiv preprint arXiv:1808.05328* (2018).

[7] David F Anderson. "A proof of the global attractor conjecture in the single linkage class case". In: *SIAM Journal on Applied Mathematics* 71.4 (2011), pp. 1487–1508.

[8] David F Anderson. "Boundedness of trajectories for weakly reversible, single linkage class reaction systems". In: *Journal of mathematical chemistry* 49.10 (2011), p. 2275.

[9] David F Anderson. "Global asymptotic stability for a class of nonlinear chemical equations". In: *SIAM Journal on Applied Mathematics* 68.5 (2008), pp. 1464–1476.

[10] David Angeli, Patrick De Leenheer, and Eduardo D Sontag. "A Petri net approach to the study of persistence in chemical reaction networks". In: *Mathematical biosciences* 210.2 (2007), pp. 598–618.

[11] Francesco Barioli et al. "On the minimum rank of not necessarily symmetric matrices: a preliminary study". In: *Electronic Journal of Linear Algebra* 18 (2009), p. 126.

[12] Francesco Barioli et al. "Zero forcing parameters and minimum rank problems". In: *Linear Algebra Appl.* 433.2 (2010), pp. 401–411. ISSN: 0024-3795. DOI: 10.1016/j.laa.2010.03.008. URL: https://doi.org/10.1016/j.laa.2010.03.008.

[13] Wayne Barrett, Jason Grout, and Raphael Loewy. "The minimum rank problem over the finite field of order 2: Minimum rank 3". In: *Linear Algebra and its Applications* 430.4 (2009), pp. 890–923. ISSN: 0024-3795. DOI: https://doi.org/10.1016/j.laa.2008.08.025. eprint: arXiv:math.CO/0612331. URL: https://www.sciencedirect.com/science/article/pii/S0024379508003984.

[14] Wayne Barrett, Hein van der Holst, and Raphael Loewy. "Graphs whose minimal rank is two". In: *Electron. J. Linear Algebra* 11 (2004), 258–280 (electronic). ISSN: 1081-3810.

[15] Wayne Barrett, Hein van der Holst, and Raphael Loewy. "Graphs whose minimal rank is two: the finite fields case". In: *Electron. J. Linear Algebra* 14 (2005), 32–42 (electronic). ISSN: 1081-3810.

[16] Allan Bickle. "The k-Cores of a Graph". PhD thesis. Kalamazoo, Michigan: Western Michigan University, Dec. 2010.

[17] Anders Björner. *Oriented matroids*. 46. Cambridge University Press, 1999.

[18] Béla Bollobás and Arthur G Thomason. "Threshold functions". In: *Combinatorica* 7.1 (1987), pp. 35–38.

[19] Ludwig Boltzmann. *Vorlesungen über gastheorie*. Vol. 1. JA Barth (A. Meiner), 1910.

[20] Kirk Boyer et al. "The zero forcing polynomial of a graph". In: *Discret. Appl. Math.* 258 (2019), pp. 35–48.

[21] John S Breese, David Heckerman, and Carl Kadie. "Empirical analysis of predictive algorithms for collaborative filtering". In: *arXiv preprint arXiv:1301.7363* (2013).

[22] Robert Brijder. "Computing with chemical reaction networks: a tutorial". In: *Natural Computing* 18 (2019), pp. 119–137.

[23] Robert Brijder, David Doty, and David Soloveichik. "Robustness of expressivity in chemical reaction networks". In: *DNA Computing and Molecular Programming: 22nd International Conference, DNA 22, Munich, Germany, September 4-8, 2016. Proceedings 22*. Springer. 2016, pp. 52–66.

[24] Daniel Burgarth and Vittorio Giovannetti. "Full Control by Locally Induced Relaxation". In: *Phys. Rev. Lett.* 99 (10 Sept. 2007), p. 100501. DOI: 10.1103/PhysRevLett.99.100501. URL: https://link.aps.org/doi/10.1103/PhysRevLett.99.100501.

[25] Steve Butler and Michael Young. "Throttling zero forcing propagation speed on graphs". In: *Australas. J. Combin.* 57 (2013), pp. 65–71. ISSN: 1034-4942.

[26] Emmanuel Candes and Benjamin Recht. "Exact matrix completion via convex optimization". In: *Communications of the ACM* 55.6 (2012), pp. 111–119.

[27] John F. Canny. *Some Algebraic and Geometric Computations in PSPACE*. Tech. rep. UCB/CSD-88-439. EECS Department, University of California, Berkeley, Aug. 1988. URL: http://www2.eecs.berkeley.edu/Pubs/TechRpts/1988/6041.html.

[28] Yu Chan et al. "Using Markov chains to determine expected propagation time for probabilistic zero forcing". In: *Electron. J. Linear Algebra* 36 (2020), pp. 318–333. DOI: 10.13001/ela.2020.5127. URL: https://doi.org/10.13001/ela.2020.5127.

[29] Alexander L Chistov and D Yu Grigor'Ev. "Complexity of quantifier elimination in the theory of algebraically closed fields". In: *International Symposium on Mathematical Foundations of Computer Science*. Springer. 1984, pp. 17–31.

[30] Samuel Clamons, Lulu Qian, and Erik Winfree. "Programming and simulating chemical reaction networks on a surface". In: *Journal of the Royal Society Interface* 17.166 (2020), p. 20190790.

[31] Gheorghe Craciun. "Polynomial dynamical systems, reaction networks, and toric differential inclusions". In: *SIAM Journal on Applied Algebra and Geometry* 3.1 (2019), pp. 87–106.

[32] Gheorghe Craciun. "Toric differential inclusions and a proof of the global attractor conjecture". In: *arXiv preprint arXiv:1501.02860* (2015).

[33] Gheorghe Craciun et al. "Toric dynamical systems". In: *Journal of Symbolic Computation* 44.11 (2009), pp. 1551–1565.

[34] William H Cunningham. "Testing membership in matroid polyhedra". In: *Journal of Combinatorial Theory, Series B* 36.2 (1984), pp. 161–188.

[35] Guoli Ding and Andreı Kotlov. "On minimal rank over finite fields". In: *Electron. J. Linear Algebra* 15 (2006), 210–214 (electronic). ISSN: 1081-3810.

[36] David Doty and Shaopeng Zhu. "Computational complexity of atomic chemical reaction networks". In: *Natural Computing* 17.4 (2018), pp. 677–691.

[37] David Doty and Shaopeng Zhu. "Computational complexity of atomic chemical reaction networks". In: *Natural Computing* 17.4 (2018), pp. 677–691.

[38] Marta Dueñas-Dıez and Juan Pérez-Mercader. "How chemistry computes: Language recognition by non-biochemical chemical automata. From finite automata to turing machines". In: *Iscience* 19 (2019), pp. 514–526.

[39] Jack Edmonds. "Matroids and the greedy algorithm". In: *Mathematical programming* 1 (1971), pp. 127–136.

[40] Jack Edmonds. "Submodular functions, matroids, and certain polyhedra". In: *Combinatorial Optimization—Eureka, You Shrink! Papers Dedicated to Jack Edmonds 5th International Workshop Aussois, France, March 5–9, 2001 Revised Papers*. Springer. 2003, pp. 11–26.

[41] Sean English, Calum MacRury, and Paweł Prałat. "Probabilistic zero forcing on random graphs". In: *European J. Combin.* 91 (2021), Paper No. 103207, 22. ISSN: 0195-6698. DOI: 10.1016/j.ejc.2020.103207. URL: https://doi.org/10.1016/j.ejc.2020.103207.

[42] Martin Feinberg. "Chemical reaction network structure and the stability of complex isothermal reactors—I. The deficiency zero and deficiency one theorems". In: *Chemical engineering science* 42.10 (1987), pp. 2229–2268.

[43] Martin Feinberg. "Complex balancing in general kinetic systems". In: *Archive for rational mechanics and analysis* 49.3 (1972), pp. 187–194.

[44] Martin Feinberg. *Foundations of chemical reaction network theory*. Springer, 2019.

[45] Martin Feinberg and Friedrich JM Horn. "Chemical mechanism structure and the coincidence of the stoichiometric and kinetic subspaces". In: *Archive for Rational Mechanics and Analysis* 66.1 (1977), pp. 83–97.

[46] Martin Feinberg and Friedrich JM Horn. "Dynamics of open chemical systems and the algebraic structure of the underlying reaction network". In: *Chemical Engineering Science* 29.3 (1974), pp. 775–787.

[47] Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

[48] Jesse Geneson and Leslie Hogben. *Propagation time for probabilistic zero forcing*. 2018. DOI: 10.48550/ARXIV.1812.10476. URL: https://arxiv.org/abs/1812.10476.

[49] Gilles Gnacadja et al. "Monotonicity of interleukin-1 receptor–ligand binding with respect to antagonist in the presence of decoy receptor". In: *Journal of theoretical biology* 244.3 (2007), pp. 478–488.

[50] Manoj Gopalkrishnan, Ezra Miller, and Anne Shiu. "A geometric approach to the global attractor conjecture". In: *SIAM Journal on Applied Dynamical Systems* 13.2 (2014), pp. 758–797.

[51] Aleksandr Nikolaevich Gorban and Ilya V Karlin. *Invariant manifolds for physical and chemical kinetics*. Vol. 660. Springer, 2005.

[52] AIM Minimum Rank-Special Graphs Work Group. "Zero forcing sets and the minimum rank of graphs". In: *Linear Algebra Appl.* 428.7 (2008), pp. 1628–1648. ISSN: 0024-3795. DOI: 10.1016/j.laa.2007.10.009. URL: https://doi.org/10.1016/j.laa.2007.10.009.

[53] Marshall Hall. "Finite Projective Planes". In: *The American Mathematical Monthly* 62.7P2 (1955), pp. 18–24. DOI: 10.1080/00029890.1955.11988741. eprint: https://doi.org/10.1080/00029890.1955.11988741. URL: https://doi.org/10.1080/00029890.1955.11988741.

[54] Marshall Hall. "The Theory of Groups". In: Macmillan, 1959. Chap. 20.

[55] Godfrey H Hardy. "Mendelian proportions in a mixed population". In: *Science* 28.706 (1908), pp. 49–50.

[56] L. Hogben, J.C.H. Lin, and B.L. Shader. *Inverse Problems and Zero Forcing for Graphs*. Mathematical Surveys and Monographs. American Mathematical Society, 2022. ISBN: 9781470466558. URL: https://books.google.com/books?id=cKF9EAAAQBAJ.

[57] Leslie Hogben, Jephian C-H Lin, and Bryan L Shader. *Inverse Problems and Zero Forcing for Graphs*. Vol. 270. American Mathematical Society, 2022.

[58] Fritz Horn. "Necessary and sufficient conditions for complex balancing in chemical kinetics". In: *Archive for Rational Mechanics and Analysis* 49 (1972), pp. 172–186.

[59] Fritz Horn and Roy Jackson. "General mass action kinetics". In: *Archive for rational mechanics and analysis* 47.2 (1972), pp. 81–116.

[60] David Hu and Alec Sun. *Probabilistic Zero Forcing on Grid, Regular, and Hypercube Graphs*. 2020. DOI: 10.48550/ARXIV.2010.12343. URL: https://arxiv.org/abs/2010.12343.

[61] Vassilis Kalofolias et al. "Matrix completion on graphs". In: *arXiv preprint arXiv:1408.1717* (2014).

[62] Varun Kanade and Alistair Sinclair. "A simpler proof of single-linkage case for quadratic systems". Unpublished.

[63] Cong X. Kang and Eunjeong Yi. "Probabilistic zero forcing in graphs". In: *Bull. Inst. Combin. Appl.* 67 (2013), pp. 9–16. ISSN: 1183-1278.

[64] K. H. Kim and F. W. Roush. "Kapranov rank vs. tropical rank". In: *Proc. Amer. Math. Soc.* 134.9 (2006), 2487–2494 (electronic). ISSN: 0002-9939. eprint: arXiv:math.CO/0503044v2.

[65] Rachel Lawrence. *MinRank3 package*. https://github.com/RachelLawrence/MinRank3. 2023.

[66] Rachel Lawrence et al. "Zero Forcing with Random Sets". In: *arXiv preprint arXiv:2208.12899* (2022). URL: https://doi.org/10.48550/arXiv.2208.12899.

[67] N.R. Lebovitz. *Ordinary Differential Equations*. Mathematics Series. Brooks/Cole, 1999. ISBN: 9780534365523. URL: https://books.google.com/books?id=tVHakQEACAAJ.

[68] Alfred J Lotka. "Contribution to the theory of periodic reactions". In: *The Journal of Physical Chemistry* 14.3 (2002), pp. 271–274.

[69] Timothy W McKeithan. "Kinetic proofreading in T-cell receptor signal transduction." In: *Proceedings of the national academy of sciences* 92.11 (1995), pp. 5042–5046.

[70] A. Montina. *Output-sensitive algorithm for generating the flats of a matroid*. 2011. arXiv: 1107.4301 [math.CO].

[71] Nils E Napp and Ryan P Adams. "Message passing inference with chemical reaction networks". In: *Advances in neural information processing systems* 26 (2013).

[72] Harihar Narayanan. "A rounding technique for the polymatroid membership problem". In: *Linear algebra and its applications* 221 (1995), pp. 41–57.

[73] Shyam Narayanan and Alec Sun. "Bounds on expected propagation time of probabilistic zero forcing". In: *European J. Combin.* 98 (2021), Paper No. 103405, 16. ISSN: 0195-6698. DOI: 10.1016/j.ejc.2021.103405. URL: https://doi.org/10.1016/j.ejc.2021.103405.

[74] Mohammad Reza Oboudi. "On the zero forcing number of trees". In: *Iran. J. Sci. Technol. Trans. A Sci.* 45.3 (2021), pp. 1065–1070. ISSN: 1028-6276. DOI: 10.1007/s40995-021-01112-5. URL: https://doi.org/10.1007/s40995-021-01112-5.

[75] James Oxley. *Matroid Theory*. Oxford University Press, Feb. 2011. ISBN: 9780198566946. DOI: 10.1093/acprof:oso/9780198566946.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780198566946.001.0001.

[76] Yuri Rabinovich, Alistair Sinclair, and Avi Wigderson. Unpublished manuscript, extended version of [77]. 1993.

[77] Yuri Rabinovich, Alistair Sinclair, and Avi Wigderson. "Quadratic dynamical systems (preliminary version)". In: *FOCS*. 1992, pp. 304–313.

[78] Benjamin Recht. "A simpler approach to matrix completion." In: *Journal of Machine Learning Research* 12.12 (2011).

[79] Darren D. Row. "A technique for computing the zero forcing number of a graph with a cut-vertex". In: *Linear Algebra and its Applications* 436.12 (2012), pp. 4423–4432. ISSN: 0024-3795. DOI: https://doi.org/10.1016/j.laa.2011.05.012. URL: https://www.sciencedirect.com/science/article/pii/S0024379511004174.

[80] Sayed Ahmad Salehi, Keshab K Parhi, and Marc D Riedel. "Chemical reaction networks for computing polynomials". In: *ACS synthetic biology* 6.1 (2017), pp. 76–83.

[81] Marcus Schaefer. "Complexity of some geometric and topological problems". In: *International Symposium on Graph Drawing*. Springer. 2009, pp. 334–344.

[82] Marcus Schaefer and Daniel Štefankovič. "Fixed points, Nash equilibria, and the existential theory of the reals". In: *Theory of Computing Systems* 60.2 (2017), pp. 172–193.

[83] Alexander Schrijver. *Combinatorial optimization: polyhedra and efficiency*. Vol. 24. Springer Science & Business Media, 2003.

[84] Dana Scott. "Measurement structures and linear inequalities". In: *Journal of mathematical psychology* 1.2 (1964), pp. 233–247.

[85] Guy Shinar and Martin Feinberg. "Concordant chemical reaction networks". In: *Mathematical biosciences* 240.2 (2012), pp. 92–113.

[86] Anne Shiu and Bernd Sturmfels. "Siphons in chemical reaction networks". In: *Bulletin of mathematical biology* 72 (2010), pp. 1448–1463.

[87] David Siegel and Debbie MacLean. "Global stability of complex balanced mechanisms". In: *Journal of Mathematical Chemistry* 27.1 (2000), pp. 89–110.

[88] David Soloveichik et al. "Computation with finite stochastic chemical reaction networks". In: *natural computing* 7 (2008), pp. 615–633.

[89] Eduardo D Sontag. "Structure and stability of certain chemical networks and applications to the kinetic proofreading model of T-cell receptor signal transduction". In: *IEEE transactions on automatic control* 46.7 (2001), pp. 1028–1047.

[90] James Joseph Sylvester. "XIX. A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 4.23 (1852), pp. 138–142.

[91] Marko Vasić, David Soloveichik, and Sarfraz Khurshid. "CRN++: Molecular programming language". In: *Natural Computing* 19 (2020), pp. 391–407.

[92] Charlie Wood. "How to Make the Universe Think for Us". In: *Quanta Magazine* (May 31, 2022). URL: `https://www.quantamagazine.org/how-to-make-the-universe-think-for-us-20220531/` (visited on 10/18/2023).

[93] Min Zhao, Liying Kang, and Gerard J Chang. "Power domination in graphs". In: *Discrete mathematics* 306.15 (2006), pp. 1812–1816.

Appendix A

A Linear Program for Invariants

Suppose we are given a mass-preserving, simplicial mass action system with a full-support initial condition. Is there an efficient method to determine which supports may contain limit points for the trajectory, and which are ruled out by invariants? Building on Theorem 3.10, we present a linear programming approach to determine whether a linear invariant exists to rule out a given support S for any limit points of the trajectory.

We begin with a seemingly simpler problem: Suppose we wish to show that, for a trajectory with start state $x_0 \in \mathbb{R}^N_+$ and for a specific candidate limit point z, x_0 and z are inconsistent with respect to some invariant; that is, that the value of some linear invariant $\iota(x)$ takes on different values at $\iota(x_0)$ and $\iota(z)$. The existence of any such ι for a given (x_0, z) pair is captured by asking whether $\Delta := z - x_0$ satisfies the following system.

$$\begin{aligned} &M\Delta = 0 \\ &\Delta_I < 0 \text{ for all } I \in \overline{S} \\ &\sum_{I \in \mathcal{S}} \Delta_I = 0 \end{aligned}$$

where $S := \mathit{support}(z)$, $\bar{S} := \mathcal{S} \setminus S$, and $M \in \mathbb{R}^{|E| \times |S \cup \bar{S}|}$ has rows m_e representing the incidence of element $e \in E$ in each species $I \in S \cup \bar{S}$. To see why, recall that an invariant $\iota(x)$ is a linear combination $\sum_e b_e q_e(x)$, where $q_e(x)$ is the marginal distribution for $e \in E$ at point x. There exists an invariant ruling out (x_0, z) when $q_e(x_0) \neq q_e(z)$ for one or more $e \in E$. Then, $(Mx)_e = m_e \cdot z = \sum_I (m_e)_I z_I = q_e(x)$. So if z is consistent with all invariants of the trajectory, we have

$$Mz = Mx_0$$

That is, we can rule out (x_0, z) when $M(z - x_0) \neq 0$. In order to show there exists *no* initial condition x_0 consistent with limit point z, it is sufficient to show that there is no x_0 satisfying the above equation. Letting $\Delta = z - x_0$, this condition becomes:

$$M\Delta = 0 \text{ has no solutions } \Delta$$

Finally, we add additional conditions to restrict to the case where z lies on the boundary of the state space, with support S, and x_0 on the interior of the state space, so for $I \in \bar{S}$, $z_I = 0$ and $x_0 > 0$; so we can require that

$$\Delta_I < 0 \text{ for all } I \in \bar{S}$$

Furthermore, the total quantity of species is conserved across all reactions, so we require $\sum_I z_I = \sum_I x_0$. That is,

$$\sum_I \Delta_I = 0$$

Restated, we seek some $\Delta \in \mathbb{R}^{|S \cup \bar{S}|+1}$ to satisfy the following (noting the additional entry, used for minimization, which we index as Δ_ϵ) :

$$\begin{aligned} &\textit{minimize } [0, \ldots, 0, -1] \cdot \Delta \\ &\textit{subject to: } \begin{bmatrix} M & \begin{matrix} 0 \\ \vdots \end{matrix} \\ 1 \ldots 1 & 0 \end{bmatrix} \Delta = 0 \\ &\left[\begin{array}{c|c|c} 0_{\bar{S} \times S} & I_{\bar{S} \times \bar{S}} & \begin{matrix} 0 \\ \vdots \end{matrix} \\ \hline 0_{\bar{S} \times S} & I_{\bar{S} \times \bar{S}} & \begin{matrix} 1 \\ \vdots \end{matrix} \end{array}\right] \Delta \leq 0 \end{aligned}$$

The first matrix incorporates both equality constraints, and the top half of the second matrix encodes a relaxed constraint $\Delta_I \leq 0$ for $I \in \bar{S}$. The bottom half of the second matrix puts constraints on Δ_ϵ:

$$\Delta_\epsilon + \Delta_I \leq 0 \text{ for all } I \in \bar{S}$$

Note that $\Delta_I \leq 0$ for all $I \in \bar{S}$, so this constraint enforces $\Delta_\epsilon \leq |\Delta_I|$. Specifically, because the system minimizes $-\Delta_\epsilon$, if $\Delta_I = 0$ for some $I \in \bar{S}$, then $\Delta_\epsilon = 0$. Otherwise, the optimal Δ will have Δ_ϵ as large as possible without exceeding the constraint bound.

Dual Formulation

For a complementary perspective, we can also solve the dual linear program: Minimize over vectors $w \in \mathbb{R}^{|E|+2|\bar{S}|+1}$, the indices of which will be denoted w_e for $e \in E$, w_1, w_I for $I \in \bar{S}$, and $w_{I'}$ for the second occurrences of $I \in \bar{S}$ (in that order). Define A_S and $A_{\bar{S}}$ to be the

rows of $M^\top$ corresponding to S and $\bar{S}$ respectively, so that $M^\top = \begin{bmatrix} A_S \\ A_{\bar{S}} \end{bmatrix}$. Then the dual is:

$$
\begin{aligned}
&\textit{minimize } 0 \\
&\textit{subject to: } w_I \geq 0 \text{ for } I \in \bar{S} \\
&\qquad\qquad w_{I'} \geq 0 \text{ for} I \in \bar{S} \\
&\left[\begin{array}{c|c|c|c} A_S & \begin{matrix}1\\ \vdots\end{matrix} & 0_{S\times\bar{S}} & 0_{S\times\bar{S}} \\ \hline A_{\bar{S}} & \begin{matrix}\vdots\\ 1\end{matrix} & I_{|\bar{S}|\times|\bar{S}|} & I_{|\bar{S}|\times|\bar{S}|} \\ \hline 0\cdots0 & 0 & 0\cdots0 & 1\cdots1 \end{array}\right] w = c
\end{aligned}
$$

Lemma A.1. *The primal linear program is feasible.*

Proof. $\Delta = \vec{0}$ is a solution. □

Lemma A.2. *If the primal linear program is bounded, its optimal objective value is 0.*

Proof. Notice that, for Δ a feasible solution to the primal, $c\Delta$ is also a feasible solution for all $c > 0$. Therefore, if there exists a feasible Δ with nonzero objective value $v < 0$, then for any value $v' < 0$ we can find $c > 0$ such that $c\Delta$ is feasible with objective value v', and so the objective value is unbounded. Note from the fact that $\Delta = \vec{0}$ is a feasible solution, that the objective value is ≤ 0, always. Thus, the only way to have a bounded objective is when the objective value is 0. □

Lemma A.3. *The dual linear program is either bounded feasible, or infeasible. In particular, the dual is feasible if and only if the primal is bounded.*

Proof. The objective of the dual is 0 for all w, so it is never unbounded. The dual is bounded feasible if and only if the primal is bounded feasible; we already know the dual is bounded and the primal is feasible, so the lemma follows. □

With this understanding, it is straightforward to implement the linear programming solution to determine whether there exists a full-support $x_0 \in \mathbb{R}^N_{>0}$ which is compatible (with respect to invariants) with the given candidate limit point z. We will see next that if z is not compatible with any such x_0, the dual in fact finds a vector w of the form specified in Theorem 3.10; and therefore, the invariant rules out not only z, but also all other candidate limit points with support S.

To see this, we first assume that for any $e \in I \in \bar{S}$, there is some $J \in S$ with $e \in J$; if not, the invariant q_e trivially rules out any limit point with support S. Now observe that a feasible solution w must have $\begin{bmatrix} A_S & \vec{1} \end{bmatrix} \vec{w_E} = 0$ (where $\vec{w_E}$ is w restricted to just the entries corresponding to $e \in E$ and w_1). First consider the trivial solution $w_e = 0$ for all $e \in E$ and for $e = 1$. The earlier assumption implies that $\begin{bmatrix} A_S & \vec{1} \end{bmatrix} w_E = 0$. Noting that any solution

must also satisfy $\begin{bmatrix} A_S & \vec{1} & I_{\bar{S}} & I_{\bar{S}} \end{bmatrix} w = 0$, with $w_I \geq 0$ for all $I \in \bar{S}$, we must also have $w_I = 0, w_{I'}$ for all $I \in \bar{S}$. But this contradicts the final equality constraint: $\sum_{I \in \bar{S}} w_{I'} = 1$. So any construction with $w_e = 0$ for all $e \in E$ and for $e = 1$ is not feasible. So if $\begin{bmatrix} A_S & \vec{1} \end{bmatrix}$ has a trivial nullspace, the dual is infeasible.

Suppose conversely that there is some nontrivial solution space to $\begin{bmatrix} A_S & \vec{1} \end{bmatrix} w_E = 0$. Again using the earlier assumption, this solution fully defines $\begin{bmatrix} A_{\bar{S}} & \vec{1} \end{bmatrix} w_E$. Now, we consider the signs of entries of $\begin{bmatrix} A_{\bar{S}} & \vec{1} \end{bmatrix} w_E$.

Lemma A.4. *If for some w_E satisfying $\begin{bmatrix} A_S & \vec{1} \end{bmatrix} w_E = 0$, the signs of entries of $\begin{bmatrix} A_{\bar{S}} & \vec{1} \end{bmatrix} w_E$ are either all non-negative, or all non-positive, and additionally at least one entry is nonzero, then the dual is feasible.*

Proof. First, normalize w_E by some constant c_0 such that

$$\sum_{I \in \bar{S}} [(A_{\bar{S}})_I, 1] \cdot (c_0 w_E) = -1$$

(where $(A_{\bar{S}})_I$ is the row of $A_{\bar{S}}$ corresponding to species I). Note that c_0 may be negative or positive, depending on the values in w_E. Then set

$$\begin{aligned} w_{I'} &:= [(A_{\bar{S}})_I, 1] \cdot (-c_0 w_E) \geq 0 \\ w_I &:= 0 \end{aligned}$$

It follows that for each I,

$$\begin{aligned} c_0 \left([(A_{\bar{S}})_I, 1] \cdot w_E \right) + w_I + w'_I &= 0, \\ \begin{bmatrix} A_S & \vec{1} \end{bmatrix} c_0 w_E &= 0, \\ \text{and } \sum_{I \in \bar{S}} w_{I'} &= 1. \end{aligned}$$

□

Lemma A.5. *If there is no w_E satisfying $\begin{bmatrix} A_S & \vec{1} \end{bmatrix} w_E = 0$ such that the signs of entries of $\begin{bmatrix} A_{\bar{S}} & \vec{1} \end{bmatrix} w_E$ are either all non-negative or all non-positive, with at least one nonzero, then the dual is infeasible.*

Proof. Suppose that for all w_E such that $\begin{bmatrix} A_S & \vec{1} \end{bmatrix} w_E = 0$, then $\begin{bmatrix} A_{\bar{S}} & \vec{1} \end{bmatrix} w_E$ has mixed signs or is all zero. Then for any such w_E, in order to satisfy $\begin{bmatrix} A_{\bar{S}} & \vec{1} & I_{\bar{S}} & I_{\bar{S}} \end{bmatrix} w = 0$, there will be at least one row constraint which has $[(A_{\bar{S}})_I, 1] \cdot w_E \geq 0$ and thus requires

$$\begin{aligned} [(A_{\bar{S}})_I, 1] \cdot w_E + w_I + w_{I'} &= 0 \\ \Rightarrow \; w_I + w_{I'} &\leq 0 \end{aligned}$$

This is impossible due to the nonnegativity constraints on both w_I and $w_{I'}$, and the fact that $\sum_{I \in \bar{S}} w_{I'} = 1$, so at least one $w_{I'}$ is strictly positive. □

This confirms that the primal linear program is bounded if and only if there exists an invariant of the exact form specified in Theorem 3.10. In other words, if it is possible to rule out all possible interior initial conditions x_0 for the candidate limit point z via invariant compatibility constraints, there is in fact a single invariant which rules out not only z, but also any z' with $supp(z') = S := supp(z)$.

This method can be used, as in Example 2.39, to empirically determine whether there are any stationary supports S which contain candidate limit points. For potential stationary support S, simply check whether the primal is bounded (or equivalently, whether the dual is feasible), and rule out S in any case which receives a positive answer.

Taking this method one step further, all limit points of a trajectory must form a connected set which are all *LP-compatible* in the following sense:

Definition A.1 (LP-compatible)**.** Points z_1 and z_2 are *LP-compatible* if

$$\begin{bmatrix} M \\ 1 \dots 1 \end{bmatrix} (z_2 - z_1) = \vec{0}$$

Supports S_1 and S_2 are *LP-compatible* if there exists some z_1, z_2 such that $supp(z_1) = S_1$ and $supp(z_2) = S_2$ and

$$\begin{bmatrix} M \\ 1 \dots 1 \end{bmatrix} (z_2 - z_1) = \vec{0}$$

For some stationary supports S, it may be the case that $dim\ null(\begin{bmatrix} M_S \\ 1 \dots 1 \end{bmatrix}) = 0$, from which we can conclude that there is at most one limit point with this support (or supported on any subset of this support). By the connectedness of limit points, this means that if there exists a limit point with support $\subseteq S$, then it is the only limit point of the trajectory.

Appendix B

Code for Minimum Rank Proof

This appendix exhibits a Jupyter notebook implementing the algorithm discussed in the proof of Theorem 4.13. The code can also be found on Github in [65].

Input: A polynomial in two classes of variables, $s_0 \dots s_{n_s-1}$ and $q_0 \dots q_{n_q-1}$. The s-variables are controlled by Player 2 and the q-variables are controlled by Player 1.

Overview: As specified in the proof of Theorem 4.13: Check the coefficent for all different monomials in s-variables that occur in the polynomial. If at any point a monomial with a coefficient that doesn't use any q-variables is found, report it and return. (Note that any term in the coefficient which has a factor of an s-variable should be ignored in this accounting, since it is part of a different monomial.) If the algorithm finds such a monomial, Player 2 wins. Else, Player 1 can force p to be the zero polynomial by choosing values for the q-variables.

[12]:
```
def check_monomial_coefficient(coeff, q, s, debug=False):

    found_q_free_term = 0

    if coeff.operator() is not sage.symbolic.operators.add_vararg:
        operands = [coeff]
    else:
        operands = coeff.operands()

    if debug:
        print("Coefficient terms: " + str(operands))

    for coeff_term in operands: # each term in the coefficient sum

        s_free = 1
        q_free = 1
```

```
        for q_var in q:
            if coeff_term.degree(q_var) > 0:
                q_free = 0
                break
        for s_var in s:
            if coeff_term.degree(s_var) > 0:
                s_free = 0
                break

        if q_free == 0 and s_free == 1:
            # there is a true q-term in the coefficient,
            # so P2 doesn't win with this monomial
            return False

        if q_free == 1:
            found_q_free_term = 1

    if found_q_free_term == 0:
        return False
    return True
```

[13]:
```
# Step 1: Cycle through all monomials; that is, all combinations of
#         the t and s variables that appear in the polynomial.
# Step 2: For each one: Check the coefficient. Ignore terms with
#         both q's AND s/t's. If what remains has no q's, P2 wins.

def check_P2_win(p, q, s, debug=False):

    if type(p) is not sage.symbolic.expression.Expression:
        print("Warning: Input isn't a symbolic expression type")
        return (p != 0)

    if p.operator() is not sage.symbolic.operators.add_vararg:
        summands = [p]
    else:
        summands = p.expand().operands()
    if debug:
        print("Polynomial: " + str(p))
        print("Polynomial terms: " + str(summands))
```

```
    for term in summands:

        if debug:
            print()
            print("Current term: " + str(term))

        monomial = term
        # remove all the q factors
        for q_var in q:
            if monomial.degree(q_var) > 0:
                monomial = monomial.coefficient(q_var,
                                                monomial.degree(q_var))
        if monomial.operands() and \
           monomial.operator() is sage.symbolic.operators.mul_vararg:
            for operand in monomial.operands():
                if operand.is_constant():
                    monomial = monomial / operand

        coeff = p.coefficient(monomial)

        if debug:
            print("with monomial: " + str(monomial))
            print("which has coefficient: " + str(coeff))

        if coeff == 0: # no s variables, just a constant coefficient
            continue
        if check_monomial_coefficient(coeff, q, s, debug=debug):
            if debug:
                print("P2 wins with monomial " + str(monomial) +
                      " and coefficient (" + str(coeff) +")")
            return True

    return False
```

[14]:
```
# Set up an arbitrary instance for testing.
# s are the slack variables
# q are the adversarial "constants" chosen by P1

n_s = 5
s = list(var('s%d' % i) for i in range(n_s))
```

```
n_q = 3
q = list(var('q%d' % i) for i in range(n_q))

p = s[0]*q[0] + s[0]*q[1] + s[0]*s[1]*q[2] + \
    s[2]*s[2]*s[0]*q[1] + s[2]^2 + 1
```

[5]:
```
# Test case
# Returns True and exhibits the steps used to determine it
check_P2_win(p, q, s, debug=True)
```

```
Polynomial: q1*s0*s2^2 + q2*s0*s1 + q0*s0 + q1*s0 + s2^2 + 1
Polynomial terms: [q1*s0*s2^2, q2*s0*s1, q0*s0, q1*s0, s2^2, 1]

Current term: q1*s0*s2^2
with monomial: s0*s2^2
which has coefficient: q1
Coefficient terms: [q1]

Current term: q2*s0*s1
with monomial: s0*s1
which has coefficient: q2
Coefficient terms: [q2]

Current term: q0*s0
with monomial: s0
which has coefficient: q1*s2^2 + q2*s1 + q0 + q1
Coefficient terms: [q1*s2^2, q2*s1, q0, q1]

Current term: q1*s0
with monomial: s0
which has coefficient: q1*s2^2 + q2*s1 + q0 + q1
Coefficient terms: [q1*s2^2, q2*s1, q0, q1]

Current term: s2^2
with monomial: s2^2
which has coefficient: q1*s0 + 1
Coefficient terms: [q1*s0, 1]
P2 wins with monomial s2^2 and coefficient (q1*s0 + 1)
```

[5]: True

[15]:
```
# Set up polynomials used in the proof
def make_polynomials(s, q):
    return {
        "a1": s[1] + s[2],
        "a2": q[0] + s[0] + s[1] + s[2] + s[3],
        "a3": q[1] + q[2] + s[1] + s[2] + 2*s[3],
        "a4": q[0] + s[0] + s[1] + s[2] + 2*s[3],
        "a4_2": q[1] + q[2] + s[0] +s[1] + s[2] + 2*s[3],
        "pj_prime": q[1]*s[3] + s[1]*s[3] + s[3]^2,
        "sj_prime": s[1]*s[3],
        "t_prime": s[3]^2,
        "t_dblprime": s[3] + s[3]^2,
        "m1": q[1]*s[2] + s[1]*s[2] + s[2]*s[3],
        "m2": q[2]*s[1] + s[1]*s[2] + s[1]*t,
        "m3": q[1]*q[2] + q[2]*s[1] + q[1]*s[2] + s[1]*s[2] \
              + q[1]*t +q[2]*t + s[1]*t + s[2]*t + t^2,
        "m4": s[1]*s[2],
        "m5": q[0] + s[0] + q[1]*s[2] + s[1]*s[2] + t + s[2]*t,
        "m6": q[0] + s[0] + q[2]*s[1] + q[1]*s[2] + \
              2*s[1]*s[2] + t + s[1]*t + s[2]*t,
        "m7": q[0] + s[0] + q[2]*s[1] + q[1]*s[2] + 2*s[1]*s[2] \
              + t + q[1]*t + 2*s[1]*t + s[2]*t + t^2,
        "m8": q[1]*q[2] + s[0] + q[2]*s[1] + q[1]*s[2] + \
              s[1]*s[2] + q[1]*t + q[2]*t + s[1]*t + s[2]*t + t^2,
        "m9": q[1]*q[2] + s[0] + q[2]*s[1] + q[1]*s[2] + \
              2*s[1]*s[2] + q[1]*t + q[2]*t + s[1]*t + s[2]*t + t^2,
        "m10": q[1]*q[2] + s[0] + q[2]*s[1] + q[1]*s[2] + \
               2*s[1]*s[2] + q[1]*t + q[2]*t + 2*s[1]*t + \
               s[2]*t + t^2,
        "m11": q[1]*q[2] + s[0] + q[2]*s[1] + q[1]*s[2] + \
               2*s[1]*s[2] + q[1]*t + q[2]*t + 2*s[1]*t + \
               2*s[2]*t + t^2,
        "m12": q[0] + s[0] + q[2]*s[1] + q[1]*s[2] + 2*s[1]*s[2] \
               + t + q[1]*t + q[2]*t + 2*s[1]*t + 2*s[2]*t + 2*t^2,
        "m12_2": q[1]*q[2] + s[0] + q[2]*s[1] + q[1]*s[2] + \
                 2*s[1]*s[2] + t + q[1]*t + q[2]*t + 2*s[1]*t + \
                 2*s[2]*t + 2*t^2
    }
```

[16]:
```
s_letters = ["si", "sj", "sk", "si2", "sj2", "sk2", "t"]
s = list(var(s_letters[i]) for i in range(len(s_letters)))
```

```
q_letters = ["qi", "qj", "qk", "qi2", "qj2", "qk2"]
q = list(var(q_letters[i]) for i in range(len(q_letters)))
```

[17]:
```
# Print the setup for producing equations
# For easy verification that they match as stated in the paper

s1 = [si, sj, sk, t]
q1 = [qi, qj, qk]
polys1 = make_polynomials(s1, q1)
for key, value in polys1.items():
    print(key + ": " + str(value))

# sanity check that none of the individual polynomials
# can be forced to = identically 0
for name, p in polys1.items():
    if not check_P2_win(p, q, s, debug=False):
        print("Error: The polynomial " + name)
```

```
a1: sj + sk
a2: qi + si + sj + sk + t
a3: qj + qk + sj + sk + 2*t
a4: qi + si + sj + sk + 2*t
a4_2: qj + qk + si + sj + sk + 2*t
pj_prime: qj*t + sj*t + t^2
sj_prime: sj*t
t_prime: t^2
t_dblprime: t^2 + t
m1: qj*sk + sj*sk + sk*t
m2: qk*sj + sj*sk + sj*t
m3: qj*qk + qk*sj + qj*sk + sj*sk + qj*t + qk*t + sj*t + sk*t + t^2
m4: sj*sk
m5: qj*sk + sj*sk + sk*t + qi + si + t
m6: qk*sj + qj*sk + 2*sj*sk + sj*t + sk*t + qi + si + t
m7: qk*sj + qj*sk + 2*sj*sk + qj*t + 2*sj*t + sk*t + t^2 + qi + si + t
m8: qj*qk + qk*sj + qj*sk + sj*sk + qj*t + qk*t + sj*t + sk*t + t^2 + si
m9: qj*qk + qk*sj + qj*sk + 2*sj*sk + qj*t + qk*t + sj*t + sk*t + t^2 + si
m10: qj*qk + qk*sj + qj*sk + 2*sj*sk + qj*t + qk*t + 2*sj*t + sk*t + t^2 +␣
 ↪si
m11: qj*qk + qk*sj + qj*sk + 2*sj*sk + qj*t + qk*t + 2*sj*t + 2*sk*t + t^2␣
 ↪+ si
```

```
m12: qk*sj + qj*sk + 2*sj*sk + qj*t + qk*t + 2*sj*t + 2*sk*t + 2*t^2 + qi +␣
 ↪si + t
m12_2: qj*qk + qk*sj + qj*sk + 2*sj*sk + qj*t + qk*t + 2*sj*t + 2*sk*t +␣
 ↪2*t^2 + si + t
```

[11]:
```
# Main Theorem Step:
# Check each of the pairs of equations for the conditions from
#  the theorem. Some of the variables within an equation may be
#  equal; and similarly across equations.
# Consider all combinations, with the exception of those which
#  force all variables = 0 or repeated identical q variables.

import itertools
errors = False
counter = 0
# First decide which variables for the first equation are equal
# There are 3 variables (i, j, and k) any of which may be equal
# options_1 encodes these: eg. [0,0,1] indicates i = j =/= k.
options_1 = [[0,0,0], [0,0,1], [0,1,0], [0,1,1], [0,1,2]]
for o1 in options_1:
    # Now decide which variables for the second equation are equal
    # They may equal each other and/or the variables from options_1
    # Reuse variables from options_1 and/or use up to 3 new vars
    # All the possible ways to select three of those
    # (with replacement)
    options_2 = itertools.product(list(set(o1)) + [3,4,5], repeat=3)
    # For simplicity of implementation, options_2 does contain
    #  some isomorphic options, eg. [0,0,3] and [0,0,4].
    # If implementation needs to be more efficient,
    # remove duplicates.

    for o2 in options_2:

        # Total iterations inside the nested loops: 655
        counter = counter + 1

        # Skip variable settings that create two distict equations
        # which both represent the same basic add/multiply operation
        # (The preprocessing step prevents these from occurring.)

        if o1[1] == o2[1] and o2[2] == o2[2]:
```

```
            continue
        if o1[2] == o2[1] and o1[1] == o2[2]:
            continue
        if o1[0] == o1[2] and o2[1] == o2[2] and o1[0] == o2[1] \
            and o1[1] == o2[0]:
            continue
        if o2[0] == o2[2] and o1[1] == o1[2] and o2[0] == o1[1] \
            and o2[1] == o1[0]:
            continue
        if o1[0] == o2[1] and o1[1] == o2[0] and o1[2] == o2[2]:
            continue
        if o1[0] == o2[2] and o1[1] == o2[0] and o1[2] == o2[1]:
            continue
        if o1[0] == o2[1] and o1[1] == o2[2] and o1[2] == o2[0]:
            continue

        # Set up the polynomials

        s1 = [s[o1[0]], s[o1[1]], s[o1[2]], t]
        q1 = [q[o1[0]], q[o1[1]], q[o1[2]]]

        s2 = [s[o2[0]], s[o2[1]], s[o2[2]], t]
        q2 = [q[o2[0]], q[o2[1]], q[o2[2]]]


        polys1 = make_polynomials(s1, q1)
        polys2 = make_polynomials(s2, q2)

        # Check the pair of polynomials and record any pair for
        # which the theorem conditions are not met.

        for name1, p1 in polys1.items():
            for name2, p2 in polys2.items():
                if name2 in ["t_prime", "t_dblprime"]:
                    continue # these don't have any ijk-indexed
                    # variables so we don't need a second copy
                if name2 in ["pj_prime", "sj_prime"] and \
                    name1 == name2 and q1[1] == q2[1]:
                    continue # these two only have j variables.
                    #Only one generated per unique j.
                if not check_P2_win(p1 - p2, q, s, debug=False):
```

```
                    print("Error (p1 - p2): The polynomials " + \
                          name1 + " and " + name2 + \
                          " with variables " + str([s1, s2]))
                    errors = True
                if not check_P2_win(p1 + p2 - 1, q, s, debug=False):
                    print("Error (p1 + p2 - 1): The polynomials "\
                          + name1 + " and " + name2 + \
                          " with variables " + str([s1, s2]))
                    errors = True

print(errors)
```

```
False
```

The program returns `False`; that is, as claimed in the proof of Theorem 4.13, Player 2 wins for all the necessary pairs of polynomials.