The Framework of Analytic Combinatorics Applied to RNA Structures

Christie S. Burris

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Mathematics

Christian Reidys
Peter Haskell
John Rossi
Daniel Orr

May 8, 2018
Blacksburg, Virginia

Keywords: RNA structures, γ-structures, fatgraphs, Analytic Combinatorics

The Framework of Analytic Combinatorics Applied to RNA Structures

Christie S. Burris

(ABSTRACT)

In recent years it has been shown that the folding pattern of an RNA molecule plays an important role in its function, likened to a lock and key system. γ-structures are a subset of RNA pseudoknot structures filtered by topological genus that lend themselves nicely to combinatorial analysis. Namely, the coefficients of their generating function can be approximated for large n. This paper is an investigation into the length-spectrum of the longest block in random γ-structures. We prove that the expected length of the longest block is on the order $n - O(n^{\frac{1}{2}})$. We further compare these results with a similar analysis of the length-spectrum of rainbows in RNA secondary structures, found in Li and Reidys (2018). It turns out that the expected length of the longest block for γ-structures is on the order the same as the expected length of rainbows in secondary structures.

The Framework of Analytic Combinatorics Applied to RNA Structures

Christie S. Burris

(GENERAL AUDIENCE ABSTRACT)

Ribonucleic acid (RNA), similar in composition to well-known DNA, plays a myriad of roles within the cell. The major distinction between DNA and RNA is the nature of the nucleotide pairings. RNA is single stranded, to mean that its nucleotides are paired with one another (as opposed to a unique complementary strand). Consequently, RNA exhibits a knotted 3D structure. These diverse structures (folding patterns) have been shown to play important roles in RNA function, likened to a lock and key system. Given the cost of gathering data on folding patterns, little is known about exactly how structure and function are related. The work presented centers around building the mathematical framework of RNA structures in an effort to guide technology and further scientific discovery. We provide insight into the prevalence of certain important folding patterns.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The central dogma of molecular biology has pervaded scientific literature since the 1950s soon after Watson and Crick proposed the double helix structure of DNA, both of whom are predominantly responsible for its inception [3]. In essence, the central dogma states that genetic information is passed from DNA to messenger RNA (mRNA) which in turn translate into amino acids that assemble into proteins [24]. This simplified narrative has been questioned in recent years after the realization that only 1.5% of the human genome codes for proteins [14]. What then is the role of the other 98% of non-coding RNA (ncRNA)? The initial inclination to follow the central dogma that led to the term "junk" DNA eventually dissipated. Researchers are now aware of numerous regulatory functions carried out by ncRNA [16].

The major distinction between DNA and RNA, apart from their respective nucleotide compositions is the nature of the nucleotide pairings. A sequence and its set of pairings is referred to as a *secondary structure*. Similarly, a molecule's *tertiary structure* additionally makes reference to its spatial embedding. For instance, DNA is double-stranded; the well-known double helix tertiary structure is formed by two single strands held together by nucleotides paired laterally. RNA however is single stranded, to mean that its nucleotides are paired with one another. As a result, the set of RNA secondary structures appearing in nature is diverse in shape. In recent years it has been shown that the structure of RNA plays an important role in its function, likened to a lock and key system [8, 16].

The current state of technology is such that gathering structural information is substantially more expensive and time consuming than gathering sequential information. As such, many researchers have turned to modeling structures constrained by laws of thermodynamics in an effort to supplement experimental data. The software used to implement these models and the subsequent analysis is continuously being updated and optimized for efficiency and biological accuracy [22]. Much of this work is centered around identifying ways to reduce

the complexity of a given model.

Laws of biophysics provide the framework for sampling structures from sequences. Loosely speaking, the minimum free energy (mfe) of an RNA molecule is the amount of energy required to unpair its nucleotide chain; the lower the energy, the more energy is required to break apart the bonds. Thus it is reasonable to assume that a stable RNA structure appearing in nature is one that attains its mfe. Dynamic programming algorithms are used to approximate the minimum free energy for RNA molecules [8].

Numerous polynomial-time algorithms exist for predicting RNA secondary structures. However, the prediction of pseudoknot structures is NP-complete. Methods for increasing the efficiency of these algorithms rely on the assumption that certain matrices in dynamic programming routines are sparse [15]. Möhl et al. in [17] and Huang and Reidys in [13] study sparcification of pseudoknots in an effort to lower space requirements. One property that implies sparsity is the *polymer-zeta property* which asserts that two nucleotides of distance m form a base pair with probability bm^{-c} for some constants $b > 0,\ c > 1$. If this property holds, long-distance base pairs have low probability.

Another important property that characterizes RNA sequences is their 5'-3' distance. A molecule has a so-called direction indicated by the asymmetry at the ends of a strand. The 5'-3' distance is defined as the length of the shortest path traversed along paired nucleotides from the 5' end to the 3' end of the molecule. As such, this distance is considered a measure of 'circularization', an important phenomenon amongst viral and messenger RNA. Yoffe et. al in [25] shows that the 5'-3' distance of RNA molecules is necessarily small, and largely independent of their length and sequence. The results of [11] show that the 5'-3' distances of random RNA secondary structures are distinctively smaller than those of biological RNA molecules and mfe structures.

The purpose of this paper and others is to investigate the spectrum of the lengths between paired nucleotides as it relates to the 5'-3' distance and the polymer-zeta property. What follows is a combinatorial construction of RNA structures that lends itself well to asymptotic analysis of object parameters. Section 2 provides the relevant definitions and theorems to carry out the transfer from biological information to mathematical information. Section 3 provides the main result of this paper, the expected value of the length of certain irreducible components related to nucleotide pairings.

Chapter 2

Background

2.1 RNA Structures

In order to study RNA folding patters we turn their structures into mathematical objects. More specifically, we treat them as combinatorial graphs, referred to as *diagrams* (see Figure 2.1). The set of diagrams forms a combinatorial class which we define in the next section.

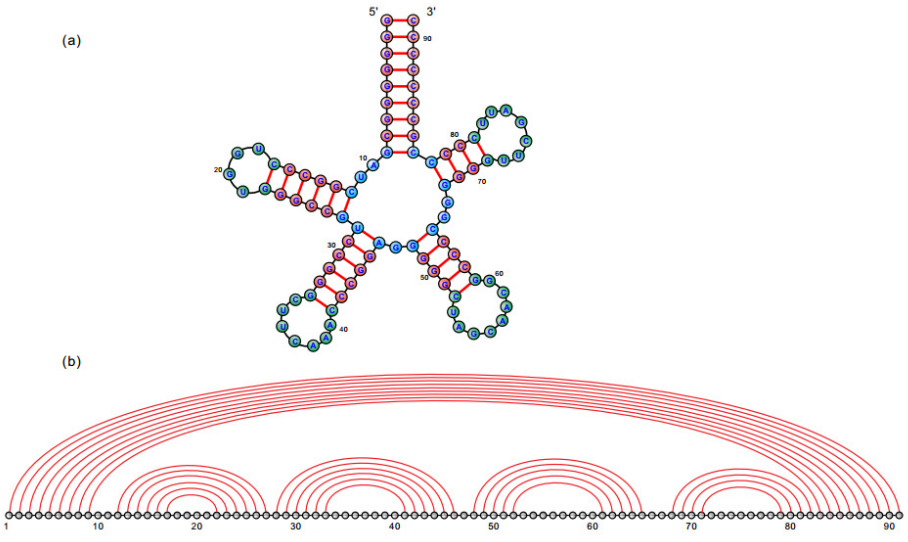


Figure 2.1: An RNA molecule represented as a contact graph (a) and as a diagram (b) [15].

Each nucleotide in an RNA molecule is represented as a labeled vertex $\{1, \ldots, n\}$. As such the length of the molecule is the size of the vertex set. The edges (i, j) in a diagram correspond to paired nucleotides in a sequences. Just as the nucleotides form a strand of phosphodiester bonds, the vertices of the diagram are represented geometrically by a horizontal line, referred

to as a *backbone*, which consists of the vertices $1, \ldots, n$ and distinguished edges $\{i, i+1\}$ for all $1 \leq i \leq n-1$. Additional pairings are represented as arcs in the upper half-plane, including non-phosphodiester bonds $(i, i+1)$ which we refer to as *1-arcs*. The length of an arc (i, j) is given by $j - i$.

Notice that the RNA structure in Figure 2.1 transfers to a diagram with no crossing arcs. This notion of crossing is essential to the mathematical and computational study of RNA structures.

Definition 1. Two arcs (i, j) and (i', j') are *crossing* if $i < i' < j < j'$.

Definition 2. An *RNA secondary structure* is a molecule whose corresponding diagram satisfies the following conditions.

1. *non-existence of 1-arcs*: if (i, j) is an arc, then $j - i \geq 2$.
2. *non-existence of base triples*: any two arcs do not share a common vertex.
3. *non-existence of crossing arcs*: any two arcs (i, j) and (i', j') are noncrossing, i.e. either $i < j < i' < j'$ or $i' < j' < i < j$.

Namely, an RNA structure S is contained in the class $\mathcal{S}$ if and only if the above conditions are held.

Definition 3. A *stack of length* τ is a maximal sequence of τ parallel arcs, $\big((i, j), (i+1, j-1), (i+2, j-2), \ldots (i+\tau, j-\tau)\big)$. Further, an RNA structure S is τ-*canonical* if it has minimum stack-length τ.

Remark. Stacks of length one are energetically unstable. Stacks of at least 2 or 3 are typically found in biological structures. The generating functions defined in later sections filter by minimum arc and stack length.

Definition 4. An arc in a diagram is considered a *rainbow* if it is maximal with respect to the partial order $(i, j) \leq (i', j') \iff i' \leq i < j \leq j'$.

Definition 5. A diagram with no crossing arcs is *irreducible* if it contains a rainbow connecting the first and last vertex.

In order for graphs to lend themselves nicely to the type of analysis performed in this paper, the class of diagrams must be realized as a combination of three combinatorial operators on fundamental combinatorial objects.

i. $\mathcal{A} = \mathcal{B} + \mathcal{C}$.

ii. $\mathcal{A} = \mathcal{B} \times \mathcal{C}$.

iii. $\mathcal{A} = \text{SEQ}(\mathcal{B})$.

(i) refers to the union of classes, (ii) to the Cartesian product of classes and (iii) to the union of Cartesian products $\mathcal{B}^n$ for all $n > 0$.

We go into further detail on this construction later. However we note here that if a combinatorial class, in this case the class of RNA structures of a particular kind, can be realized as a combination of these three operators, then the transfer from combinatorial objects to generating functions can be carried out in a fairly simple way. This transfer is referred to as the symbolic transfer or symbolic enumeration.

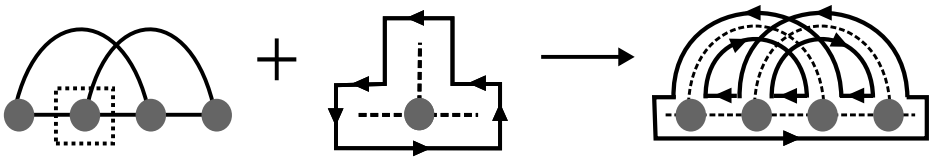
The class of RNA structures with no restriction on crossing arcs poses a challenge as the structures do not clearly contain irreducible substructures of a fundamental kind. By fundamental we mean structures with polynomial generating functions filtered by arcs or vertices. However, there are meaningful subclasses that allow for crossing in such a way that they can be constructed via symbolic enumeration.

One way to filter pseudoknot diagrams is by topological genus, which was first proposed in [19] and studied further in [23]. Topological objects hold combinatorial information in the form of key invariants that in turn provide information on the structure of the underlying graph. The idea is to pass from an drawing of a structure in $\mathbb{R}^2$ to a surface that admits an embedding — a drawing without crossings. From this notion of embedding on a higher dimensional surface, one defines the genus of a diagram as the minimal integer g such that the graph can be embedded on a surface of genus g. By definition, secondary structures are planar graphs of genus 0 since they are drawn in $\mathbb{R}^2$ without crossings.

The particular type of graph embedding employed in this treatment of pseudoknot structures is called a *combinatorial embedding* — an embedding that uniquely defines cyclic orderings of edges incident to each vertex. Consequently, the boundary components of an embedding are defined as the natural cyclic orders of the edges. In our case, this notion is straight forward since each edge is seen exactly once when traversing the boundaries. Combinatorial embeddings encode diagrams onto closed, orientable surfaces.

A combinatorial embedding is commonly represented by a *fatgraph*, or *ribbon graph* — a topological surface formed by a series of "fattenings" of the underlying diagram. Namely, vertices are inflated to discs and edges are inflated to ribbons with two boundaries referred to as half-edges (see Fig. 2.2). Associated with each vertex is a cyclic ordering of the half-edges. The resulting object is an oriented surface with boundary such that the half-edges that form the boundary of a ribbon are oriented in opposite directions (Fig. 2.2c). A more formal definition involving permutations is given below.

The power of this perspective is the equivalence of each of these constructions. Namely, every combinatorial embedding defines a unique 2-cell embedding of the underlying diagram on a closed, orientable surface, and vice versa. As a result, an RNA diagram can be easily classified by constructing its ribbon graph and applying the key invariants granted by the combinatorial embeddings.



(a) (b) (c)

Figure 2.2: (a) A diagram D, (b) inflation of a vertex in D, and (c) the corresponding oriented surface with the underlying diagram (dashed).

Definition 6. A *fatgraph*, $\mathbb{D}$, is a triple (H, σ, α) where H is a set of labeled half-edges, σ is the vertex permutation, and α is a fixed-point free involution.

Suppose a diagram D contains m arcs. To label the half-edges of $\mathbb{D}$, collapse the backbone of the underlying diagram into a single vertex and draw the ribbons as loops around this vertex. Label the half-edges $1, \ldots, 2m$ in numerical order as they appear on the right side of a ribbon when traversing the vertex clockwise. The counter clockwise cycle around the vertex is the vertex permutation σ. α is the composition of 2-cycles representing pairs of half edges that form the ribbons of $\mathbb{D}$. Note that the permutation $\gamma = \sigma \circ \alpha$ is a composition of the boundary components on $\mathbb{D}$. Namely, the group $\langle \sigma, \alpha \rangle$ acts transitively on H.

For example, consider the inflation of an underlying graph D drawn in Fig. 2.3a with the single vertex ribbon graph drawn in 2.3b. The vertex permutation σ is determined by the counter clockwise cycle around the single vertex ribbon graph. In this case, $\sigma = (654321)$. The fixed-point free involution α associated with $\mathbb{D}$ is $\alpha = (13)(25)(46)$. The boundary components (1245) and (36) are determined by the composition $\gamma = \sigma \circ \alpha = (1245)(36)$.

As mentioned previously, ribbon graphs can be viewed as embeddings on a closed, orientable surface from which key variants are extracted by first relabeling the oriented half-edges such that the two half-edges that form the boundary of a ribbon have the same label and a clockwise orientation is denoted by an inverse (Fig. 2.4b). This gives rise to a polygonal presentation

$$\langle S \mid W_1, \ldots, W_l \rangle$$

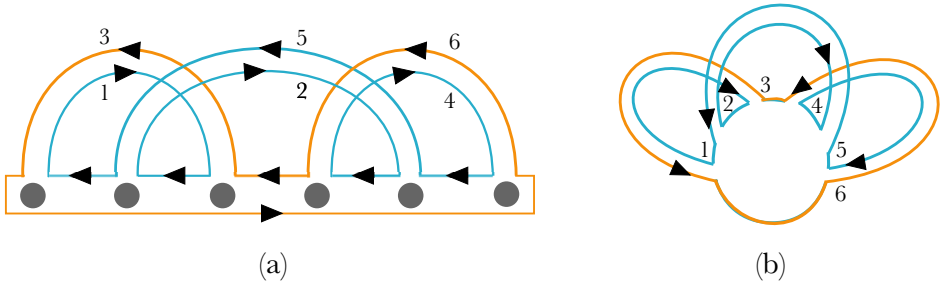


Figure 2.3: (a) A fatgraph with half-edges labeled based on the vertex permutation of the drawing on the right. (b) The corresponding fatgraph with the backbone collapsed into a single vertex. Half-edges of the same color lie on the same boundary component.

where S is the set of labels on half-edges and W_i corresponds to the traversal of a boundary component. Namely, each cycle formed when traversing the directed half-edges of a fatgraph is a boundary component and can be realized as a polygonal region.

Recall that consolidating—replacing every occurrence of ab by a and $b^{-1}a^{-1}$ by a^{-1} if a and b always appear in this manner—is a transformation that leaves the realization topologically invariant. Consequently, the half-edges of the backbone can be consolidated and a set of boundary components consisting only of arc ribbons remains.

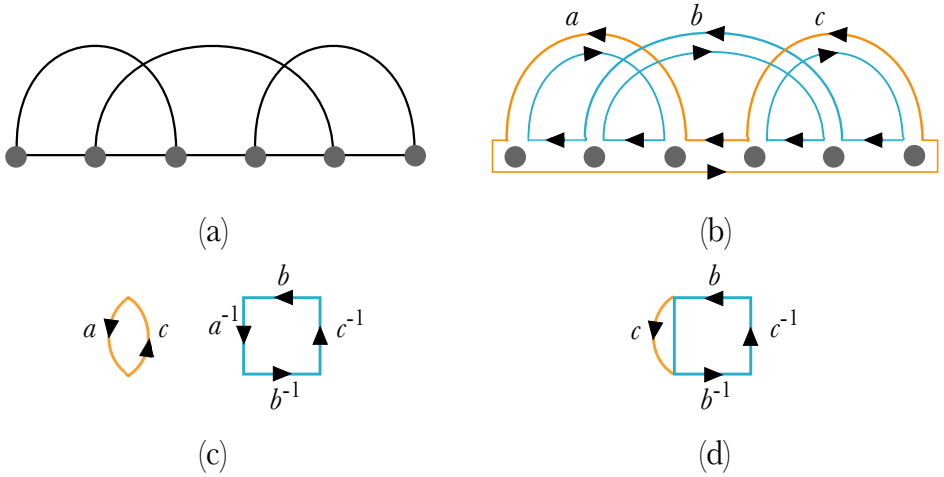


Figure 2.4: An irreducible shadow in (a) is inflated to its surface (b). The boundary components of the surface are represented as polygonal regions (c). Finally, a and a^{-1} are glued (d). The resulting polygonal region is realized as torus, a surface of genus $g = 1$.

It is well known that there is a correspondence between closed surfaces and polygonal regions with an even number of edges and an equivalence relation that identifies each edge with exactly one other edge. This is precisely the construction of fatgraphs presented above. Thus a fatgraph can be realized as an orientable closed surface, $X_{\mathbb{D}}$. For an example of this

transfer, see Figure 2.4.

Certain topological invariants such as the Euler characteristic and genus are well-defined for $X_{\mathbb{D}}$. The Euler characteristic is given by

$$\chi(X_{\mathbb{D}}) = v - m + r$$

where v is the number of vertices, m the number of edges, and r the number of boundary components in $\mathbb{D}$. Genus is then defined by the Euler characteristic,

$$g(X_{\mathbb{D}}) = 1 - \frac{1}{2}\chi(X_{\mathbb{D}}).$$

Lemma 1. *Removing isolated vertices, inserting a parallel arc, and removing noncrossing arcs in a diagram do not affect the Euler characteristic.*

Proof. Consider a diagram D with Euler characteristic $\chi(X_{\mathbb{D}})$. Recall from our discussion on the polygonal presentation that the backbone of D can be consolidated or collapsed into a single vertex. This is the case since, if the vertex v_i is unpaired, the subwords $e_{i-1}e_i$ and $e_i^{-1}e_{i-1}^{-1}$ appear in some W_j in which case the subwords consolidate to a single letter. Viewed differently, the addition or removal of a vertex and an edge does not change the Euler characteristic.

Suppose the arc $\tilde{a} = (i+1, j-1)$ is inserted under an arc $a = (i, j)$. The boundary component originally containing a^{-1} is replaced by $\tilde{a}^{-1}$. Furthermore, a new boundary component $a^{-1}\tilde{a}$ is formed. The addition of one edge and one boundary component cancel out in the computation of the Euler characteristic. Namely, $\tilde{a}$ can be glued to $\tilde{a}^{-1}$ in the original boundary component and a^{-1} remains as before.

Consider a noncrossing arc $a = (i, j)$ in D. Originally, this arc creates the boundary component $a^{-1}e_{j-1}^{-1}\cdots e_i^{-1}$ where $e_i^{-1}, \ldots, e_{j-1}^{-1}$ are the inner backbone edges underneath a. The half-edge a is traversed in a different boundary component. When computing the polygonal presentation of $X_{\mathbb{D}}$, a^{-1} glues with a and is replaced by $e_{j-1}^{-1}\cdots e_i^{-1}$. This is precisely the presentation if the arc a is removed entirely. Viewed differently, the Euler characteristic with a removed subtracts one edge and one boundary component. Thus removing noncrossing arcs does not change the Euler characteristic. □

Corollary 1.1. *The relation between genus and the number of boundary components is solely determined by the number of arcs in the upper half-plane. Namely,*

$$2 - 2g = 1 - m + r, \tag{2.1}$$

where m is the number of arcs (or equivalently, ribbons) and r is the number of boundary components.

Remark. Corollary 1.1 offers a means of computing the genus of a diagram with minimal effort. Namely, a diagram is fattened and the half-edges are traversed to compute the number of boundary components. This information along with the number of arcs is all that is required to compute the genus. For simple diagrams this process can be carried out by hand. However, given an arbitrary graph G, it is NP-complete to find the smallest g such that G has a combinatorial embedding on a genus-g surface [21].

The genus of ribbon graphs is used to classify subsets of pseudoknot structures. One natural subclass is the set of structures with fixed genus [7]. Another subclass that lends itself well to random sampling is the set of k-noncrossing structures, studied in [2,20]. 10 other subclasses are collected in [18].

Section 3 analyzes the subclass known as γ-*structures*. For fixed γ, a γ-structure is composed of irreducible components (of a more general form than irreducibility defined above) whose individual genus is bounded by γ and contain no bonds of length one (1-arcs). Note that the genus of a particular structure is not bounded since the genus of a composition of nested structures is additive.

An *isolated arc* is a stack of length 0. The stack $\big((i,j),(i{+}1,j{-}1),(i{+}2,j{-}2),\ldots(i{+}\tau,j{-}\tau)\big)$ of length τ induces 2τ *P-intervals*, $[i,i+1],[i+1,i+2],\ldots[j-1,j]$. Namely, a P-interval is any interval between 2 arcs that are nested in a stack. The interval $[i+\tau,j-\tau]$ is called a σ-*interval*. Note that if a subinterval of a σ-interval is a P-interval, it is no longer a σ-interval (see Figure 2.5). Distinguishing between these intervals becomes relevant when deriving generating functions.

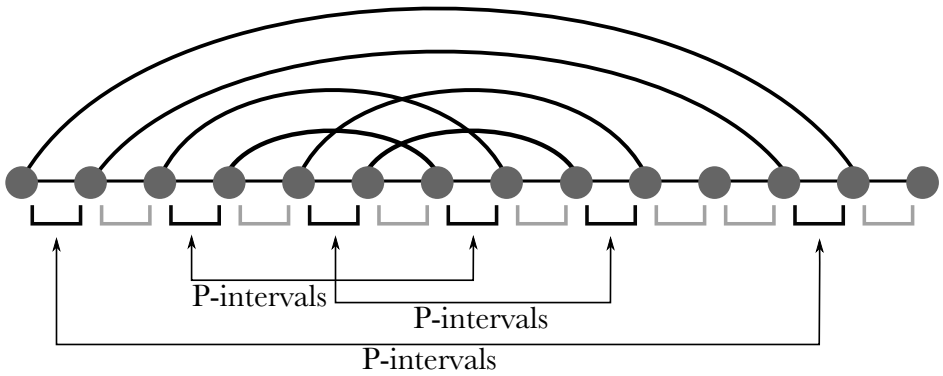


Figure 2.5: Secondary structure with labeled P-intervals. All other intervals are σ-intervals.

The *shadow* of the diagram is obtained by removing all noncrossing arcs, deleting all isolated vertices, and collapsing all induced stacks (i.e. maximal subsets of subsequent, parallel arcs) to single arcs. By construction, all arcs in a shadow are maximal. A *matching* is a

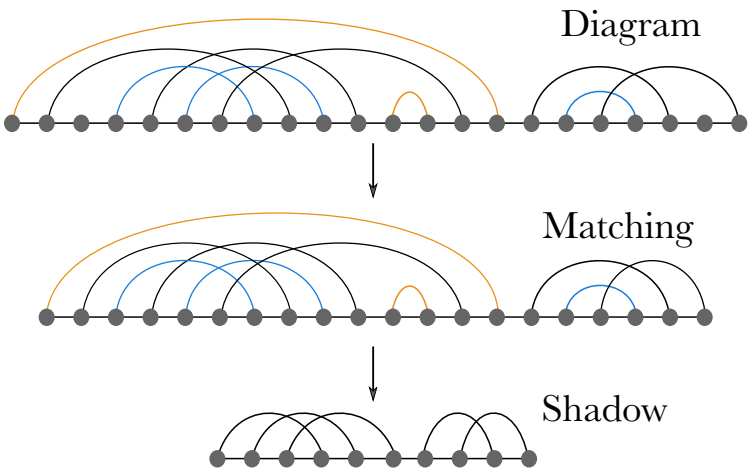


Figure 2.6: The shadow of a diagram is obtained by removing isolated vertices (this intermediate step produces a matching), removing non-crossing arcs (orange), and collapsing all stacks (blue) and resulting stacks into isolated arcs.

diagram that contains only vertices of degree 3. As the name implies, all vertices are paired. Figure 2.6 describes the construction of a shadow from its corresponding diagram.

One can define an equivalence relation on arcs of a diagram as follows. For any arcs $\alpha_1, \alpha_k \in E$, $\alpha_1 \sim \alpha_k$ iff there exists a sequence of arcs $(\alpha_1, \ldots, \alpha_{k-1}, \alpha_k)$ such that α_i and α_{i+1} are crossing for all i. This equivalence relation in turn gives us a notion of irreducibility. Namely, a crossing diagram is *irreducible* if for any two arcs α_i, α_j in its edge set E, $\alpha_i \sim \alpha_j$. By construction, the backbone of a diagram can be partitioned into irreducible components which are referred to as *blocks*.

γ-diagrams and γ-matchings are such that their irreducible components have genus at most γ. A γ-shape is a γ-matching that contains only isolated arcs. The objects of interest to us are τ-canonical γ-structures, γ-diagrams with minimum stack length τ and no 1-arcs.

Remark. In Section 3.1, the generating function for arbitrary γ is derived for γ-structures. However, in this paper we restrict our attention to $\gamma = 1$ because all four shadows of genus 1 are irreducible which yields a relatively simple generating function for 1-structures.

2.2 Analytic Combinatorics

As mentioned previously, we ask that the RNA structures we study lend themselves nicely to symbolic enumeration. Here we describe why that is so, and how as a result we can extract powerful information on parameters of RNA structures. What follows is a brief introduction

to the field of Analytic Combinatorics, taken from Flajolet and Sedgewick [6], as it applies to the analysis of RNA graph parameters. Throughout this section and the rest of the paper, we refer to definitions and theorem from [6] followed by their corresponding chapter or page number. The work of many, including Li and Reidys in [15], and the generalization in Section 3 apply the framework provided in [6].

Analytic Combinatorics is the study of large combinatorial objects. For any field of mathematics, one can identify the objects and operators with which the field is concerned. Here the objects of study are generating functions, formal power series whose coefficients carry information on a parameter of a combinatorial class. The operators are symbolic and analytic transfers.

Symbolic transfers allow us to pass from combinatorial constructions to generating functions without the cumbersome task of solving a recurrence relation.

Definition 7. A *combinatorial class*, $\mathcal{C}$, is a finite or denumerable set on which a size function $w : \mathcal{C} \to \mathbb{Z}_{\geq 0}$ is defined such that the number of elements of any given size is finite. Of note are two classes: the class containing one object of size 0, denoted $\mathcal{E}$, and the class containing one object of size 1, denoted $\mathcal{Z}$.

Three common constructions arise in the describing classes of RNA structures.

$$\mathcal{A} = \mathcal{B} \sqcup \mathcal{C}$$

refers to $\mathcal{A}$ as the disjoint union of classes $\mathcal{B}$ and $\mathcal{C}$. Similarly,

$$\mathcal{A} = \mathcal{B} \times \mathcal{C}$$

if the elements of $\mathcal{A}$ can be realized as the ordered pair of an object from $\mathcal{B}$ and an object from $\mathcal{C}$. The size of a pair $a = (b, c)$ is computed $w(a) = w(b) + w(c)$. Finally, the sequence class

$$\mathcal{A} = \mathrm{SEQ}(\mathcal{B})$$

is the infinite union of the classes $\mathcal{B}$, $\mathcal{B} \times \mathcal{B}$, etc. together with the empty class $\mathcal{E}$. Namely,

$$\mathrm{SEQ}(\mathcal{B}) = \mathcal{E} + \sum_{n \geq 1} \mathcal{B}^n$$

The size of an object $a = (b_1, b_2, \ldots, b_j)$ is $w(a) = w(b_1) + w(b_2) + \cdots + w(b_j)$. Note that this definition of size is well-defined only if $\mathcal{B}$ contains no empty word.

The symbolic transfer can be summarized by the following dictionary from combinatorial class to generating function.

$$\begin{cases} \mathcal{E} & \Longrightarrow \boldsymbol{E}(z) = 1 \\ \mathcal{Z} & \Longrightarrow \boldsymbol{Z}(z) = z \\ \mathcal{A} = \mathcal{B} \sqcup \mathcal{C} & \Longrightarrow \boldsymbol{A}(z) = \boldsymbol{B}(z) + \boldsymbol{C}(z) \\ \mathcal{A} = \mathcal{B} \times \mathcal{C} & \Longrightarrow \boldsymbol{A}(z) = \boldsymbol{B}(z) \cdot \boldsymbol{C}(z) \\ \mathcal{A} = \mathrm{SEQ}(\mathcal{B}) & \Longrightarrow \boldsymbol{A}(z) = \frac{1}{1-\boldsymbol{B}(z)} \end{cases}$$

The last transfer holds only when $\mathcal{B}$ does not contain an empty structure. To see this, consider

$$\mathcal{A} = \mathrm{SEQ}(\mathcal{B}) = \{\epsilon\} \sqcup \mathcal{B} \sqcup (\mathcal{B} \times \mathcal{B}) \sqcup \cdots$$

where $\{\epsilon\}$ is the null object. By the first two transfers,

$$\boldsymbol{A}(z) = 1 + \boldsymbol{B}(z) + \boldsymbol{B}(z)^2 + \cdots = \sum_{i \geq 0} \left(\boldsymbol{B}(z)\right)^i .$$

Loosely speaking, the sum can be realized as a geometric series and the transfer seems obvious. However this representation of the geometric series is usually followed by a bound on $\boldsymbol{B}(z)$ in accordance with the notion of convergence which is only valid if the objects under scrutiny are analytic objects. Here the objects are formal power series with coefficients in the ring $\mathbb{Q}$. One arrives at the equation in the transfer by considering the notion of inverses in the ring $\mathbb{Q}[[x]]$. The transfer holds if and only if $1 - \boldsymbol{B}(z)$ has an inverse in $\mathbb{Q}[[x]]$. Namely, if $1 - \boldsymbol{B}(z)$ has a nonzero constant term, hence the reason why $\mathcal{B}$ must not contain an empty structure.

The analytic transfer allows us to pass from analytic functions (realized as convergent power series) to coefficient asymptotics. How then do we arrive at an analytic function when the original generating functions are formal objects? If it is possible to extract a functional equation whose solution is the generating function, then considering the equation as an analytic function allows us to pass to coefficient asymptotics. The guiding principles of this transfer are the following, taken from [6].

1. The *location* of a function's singularities dictates the *exponential growth* of its coefficients.

2. The *nature* of a function's singularities determines the associate *subexponential factor*.

The first principle is an immediate consequence of the definition of power series convergence. The second principle is the topic of Chapter 6 in [6]. Moreover, there is a correspondence between the asymptotic expansion of a function near its dominant singularities and the

asymptotic expansion of the function's coefficients. See the appendix for a closer look at the definitions behind this framework. Below are two examples of analytic transfer.

$$\begin{cases} f(z) = \frac{1}{1-(z/\rho)} & \implies [z^n]f(z) = \rho^{-n} \\ f(z) = \frac{1}{(1-(z/\rho))^\alpha} & \implies [z^n]f(z) \approx \frac{n^{\alpha-1}}{\Gamma(\alpha)}\rho^{-n} \quad (\alpha \in \mathbb{C}\backslash\mathbb{Z}_{\leq 0}) \end{cases}$$

The first is a well-known transfer of the geometric series. The second transfer is key for the analysis of RNA structures. We state the transfer as a theorem to refer back to in Section 3. A proof sketch can be found in the appendix.

Theorem 2. ***(Flajolet, Sedgewick)*** *Let α be an arbitrary complex number in $\mathbb{C}\backslash\mathbb{Z}_{\leq 0}$. The coefficient of z^n in*

$$f(z) = (1-z)^{-\alpha}$$

admits for large n a complete asymptotic expansion in descending powers of n,

$$[z^n]f(z) \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)}(1 + O(n^{-1})).$$

A similar theorem vital to the transfer from generating function to asymptotic estimation of coefficients involves Big-Oh and little-oh.

Theorem 3. ***(Flajolet, Sedgewick)*** *Let $\alpha, \beta \in \mathbb{R}$ and let $f(z)$ be a function that is Δ-analytic.*

(i) Assume that $f(z)$ satisfies in the intersection of a neighborhood of 1 with its Δ-domain the condition

$$f(z) = O((1-z)^{-\alpha}).$$

Then one has:

$$[z^n]f(z) = O(n^{\alpha-1}).$$

(ii) Assume that $f(z)$ satisfies in the intersection of a neighborhood of 1 with its Δ-domain the condition

$$f(z) = o((1-z)^{-\alpha}).$$

Then one has:

$$[z^n]f(z) = o(n^{\alpha-1}).$$

Recall that we have shifted from viewing generating functions as *formal* objects to *analytic* objects. Via symbolic enumeration one can derive a functional form for a generating function. Treating the functional form as a complex function in one (or more) variables one can

then seek its asymptotic expansion near its dominant singularity, a process referred to as singularity analysis. By Theorems 2 and 3 one can carry out a term-by-term transfer of the asymptotic expansion to arrive at an asymptotic estimate of the coefficients.

It is also possible that an explicit functional form for the generating function can not be found, or rather need not be found. This is the case in Section 3 for γ-matchings. Theorem 4 gives an alternative to carrying out explicit singularity analysis given an implicit formula for the generating function. Namely, the generating function $y(z)$ satisfies $y(z) = F(z, y)$.

Theorem 4. ***(Flajolet, Sedgewick)*** *Let $F(z, w)$ be a bivariate function analytic at $(z, w) = (z_0, w_0)$. Assume the conditions:*

$$F(z_0, w_0) = 0, \quad F_z(z_0, w_0) \neq 0, \quad F_w(z_0, w_0) = 0, \quad F_{ww}(z_0, w_0) \neq 0$$

Choose an arbitrary ray angle θ emanating from z_0. Then there exists a neighborhood Ω of z_0 such that at every point z of Ω with $z \neq z_0$ and z not on the ray, the equation $F(z, y) = 0$ admits two analytic solutions $y_1(z)$ and $y_2(z)$ that satisfy, as $z \to z_0$:

$$y_1(z) = y_0 + \delta\sqrt{1 - z/z_0} + O(1 - z/z_0), \quad \delta := \sqrt{\frac{2z_0 F_z(z_0, w_0)}{F_{ww}(z_0, w_0)}}$$

and similarly for y_2 whose expansion is obtained by changing the square root to a negative square root.

Remark. Theorem 4 demonstrates the universality of the square-root singularity type. Combinatorial classes that are defined by symbolic enumeration of subclasses with finite generating functions admit a recursion as well as a generating function with an implicit functional form. The Analytic Implicit Function Theorem (AIFT) asserts that $F(z, w)$ sufficiently smooth admits a unique solution

$$F(z, f(z)) = 0, \quad |z| < \rho$$

where $f(z)$ is analytic in the neighborhood $|z| < \rho$. Locally, near $(\rho, f(\rho))$, $F(z, w)$ has a Taylor expansion

$$F + (w - f(\rho))F_w + (z - \rho)F_z + \frac{1}{2}(w - f(\rho))^2 F_{ww},$$

(lower order terms omitted). The assumption $F = F_w = 0$ of AIFT simlifes the above equation such that the solution to $F(z, w)$ near $z = \rho$ admits a square-root form.

Each of the three Theorems from [6] make statements regarding the asymptotic behavior of the coefficients of a function's series expansion near its singularity. To arrive at the singular

expansion of our generating functions, we first investigate the nature of their singularities. i.e. locate their dominant singularity and determine uniqueness. To do so we employ Pringsheim's Theorem which is useful for our analysis since generating functions necessarily have non-negative coefficients.

Theorem 5. *(Pringsheim's Theorem) If $f(z)$ is representable at the origin by a series expansion that has non-negative coefficients and radius of convergence R, then the point $z = R$ is a singularity of $f(z)$.*

Until now we have implicitly only considered generating functions of a single variable (OGFs) as they carry information on the size parameter. However, we are not restricted to asymptotic coefficient analysis of OGFs. Analytic combinatorics also provides a framework that makes it easy to study properties of parameters of graphs such as: How many substructures of a particular kind appear in a random graph? To answer this type of question, we introduce the bivariate generating function

$$\boldsymbol{F}(z, u) = \sum_{n,b} f(n, b) z^n u^b,$$

where $f(n, b)$ is the count of the number of objects of size n with the additional property of having b substructures of a particular kind.

We are still interested in extracting coefficient asymptotics from these bivariate generating functions. However, a double coefficient extraction

$$f_{n,b} = [z^n u^b] \boldsymbol{F}(z, u)$$

is quite difficult. Instead, an equally powerful "horizontal" extraction does the trick in providing asymptotics that appear in moments. The goal becomes to estimate the horizontal generating function

$$f_n(u) := \sum_{b} f(n, b) u^b \equiv [z^n] \boldsymbol{F}(z, u)$$

as this term appears in the moment generating function which we discuss in the next section.

The method for studying $f_n(u)$ is known as perturbation analysis since $\boldsymbol{F}(z, 1) = \boldsymbol{F}(z)$. The variable u marking the parameter of interest is regarded as inducing a deformation of the OGF. The way in which such deformations affect the type of singularity of the counting generating functions can then be studied. It happens that the Hankel contours employed in the derivation of univariate asymptotic analysis have the additional nice property of producing uniform asymptotic expansions in the bivariate case. With uniformity of expansion, one can uniformly estimate $f_n(u)$ to mean that a small perturbation in u yields small perturbations in $f_n(u)$. Namely, for $u \in N_\epsilon(1)$,

$$f_n(1) \approx C n^{-\alpha} \rho^{-n} \Longrightarrow f_n(u) \approx C(u) n^{-\alpha(u)} \rho(u)^{-n}. \tag{2.2}$$

Definition 8. Let $\{f_u(s)\}_{u\in U}$ be a family of functions indexed by U. The asymptotic equivalence

$$f_u(s) = O(g(s)) \quad (s \to s_0),$$

is said to be *uniform with respect to* u if there exists an absolute constant K (independent of $u \in U$) and a fixed neighborhood $N(s_0)$ of s_0 such that

$$\forall u \in U, s \in N(s_0)\ , \quad |f_u(s)| \leq K|g(s)|.$$

An asymptotic expansion

$$f_u(s) = h_0(s) + h_1(s) + \ldots + O(h_m(s))$$

where $h_0(s) >> h_1(s) >> \ldots >> h_m(s)$ is uniform if for each m the Big-Oh error term is uniform.

What makes the method of analytic combinatorics so powerful is the universality of the schema that govern singularity analysis and transfer theorems. From simple derivations of generating functions through symbolic enumeration, we are given objects that, while counting profoundly distinct objects, have the same behavior near their dominant singularity, and thus similar asymptotic behavior of their coefficients.

2.3 Probabilistic Graph Theory

As the name implies, probabilistic graph theory treats graph theoretic objects as probabilistic ones to answer questions regarding how graphs with particular properties behave on average. A *measure space*, or a *probability space* $(\Omega, \mathbb{P})$ is the set of all possible graphs (given properties of interest) along with a σ-algebra $\mathcal{A}$ of subsets of Ω, endowed with a specified *probability function* $\mathbb{P} : \Omega \to [0, 1]$. The σ-algebra (i) contains the empty set, and (ii) is closed under complements and countable unions. Additionally, $\mathbb{P}$ is additive over finite and countable unions of disjoint sets, and satisfies $\mathbb{P}(\Omega) = 1$.

We say an *event* $A \subset \Omega$ is a subset of Ω that occurs with probability

$$\mathbb{P}(A) := \sum_{\omega \in A} \mathbb{P}(\omega).$$

To illustrate this construction, suppose you are interested in a certain property of graphs with n vertices and no additional specified restrictions (these are spanning subgraphs of the complete graph on n vertices). Label the set of all such graphs $\mathcal{G}_n$. You might choose to associate to $\mathcal{G}_n$ the uniform probability function. Namely, for $G \in \mathcal{G}_n$,

$$\mathbb{P}(G) = \frac{1}{|\mathcal{G}_n|}.$$

where $|\mathcal{G}_n| = 2^{\binom{n}{2}}$ since each graph is uniquely specified by whether or not each possible edge–of which there are $\binom{n}{2}$–appears.

Naturally, we wish to study how graphs look on average as the size of the graph grows without bound. A particular property of a graph, referred to as a parameter, can be studied by means of (discrete) random variables. A *random variable* $\mathbb{X}$ is a function from the sample space Ω to the real numbers, or in the case of combinatorial classes, the non-negative integers. In particular, to study the connectivity of a graph, the number of edges needed to be removed in order to disconnect the graph, one can define a random variable which assigns to each graph its connectivity number. Further analysis can provide insight into connectivity of graphs on average.

The event

$$(\mathbb{X} = m) = \{\omega \in \Omega : \ \mathbb{X}(\omega) = m\}$$

plays an important role in analyzing graph parameters. One can further define the events $\mathbb{X} < m$, $\mathbb{X} \leq m$, $\mathbb{X} > m$, and $\mathbb{X} \geq m$ analogously.

The *probability generating function* (pgf) of a discrete random variable $\mathbb{X}$ with values in $\mathbb{Z}_{\geq 0}$, is defined as

$$p(u) := \sum_b \mathbb{P}(\mathbb{X} = b)u^b.$$

The goal of this treatment as mentioned above is to determine whether a given combinatorial parameter holds on average as n becomes large. For this we introduce the following asymptotic notation. Given a sequence of probability spaces $\{(\Omega_n, \mathbb{P}_n)\}_{n\geq 1}$ we say that property A is satisfied *asymptotically almost surely* (a.a.s.) if

$$\mathbb{P}_n(A_n) \to 1, \quad \text{as } n \to \infty$$

where A_n is the subset of graphs of size n satisfying property A. Similarly, for real valued functions $f, g : \mathbb{N} \to \mathbb{R}$ we say $f \sim g$ if

$$\frac{f(n)}{g(n)} \to 1, \quad \text{as } n \to \infty.$$

One often considers satisfying property A as an event $\mathbb{X} = m$ where the random variable $\mathbb{X}$ indicates whether property A is satisfied. $\mathbb{X}_n$ can then be defined as the restriction of $\mathbb{X}$ to the subspace of graphs of size n. The above-mentioned notions of random variables, sequences of probability spaces, and asymptotic behavior provide the objects and analytic framework for studying graphs. In the next section we define moments of a distribution of a random variable, which we will see provide the quantification necessary to derive asymptotic

approximations.

In probability theory, *expectation* is the average value, or mean, of a random variable $\mathbb{X}$ denoted by $\mathbb{E}[\mathbb{X}]$. *Variance* $\mathbb{V}[\mathbb{X}]$ quantifies the notion of how concentrated the random variable is around its expected value. Expectation and variance are also known as the first and second moments of a distribution, respectively, and are defined below.

$$\mathbb{E}[\mathbb{X}] := \sum_{\omega \in \Omega} \mathbb{X}(\omega)\mathbb{P}(\omega)$$
$$\mathbb{V}[\mathbb{X}] := \mathbb{E}[(\mathbb{X} - \mathbb{E}[\mathbb{X}])^2]$$

Variance is more often expressed in the expanded form which is possible by linearity of expectation.

$$\mathbb{V}[\mathbb{X}] = \mathbb{E}[\mathbb{X}^2] - \mathbb{E}^2[\mathbb{X}].$$

Markov's Inequality provides an upper bound on the probability of an event by a ratio of the expectation which yields a relationship between these two quantities. The proof is straightforward from the definitions.

Markov's Inequality. Let $\mathbb{X}$ be a nonnegative random variable and m a positive real number. Then

$$\mathbb{P}(\mathbb{X} \geq m) \leq \frac{\mathbb{E}[\mathbb{X}]}{m}.$$

Another important theorem and the one used in the following chapters is due to Markov's teacher Chebyshev. Chebyshev's Inequality (eq. 2.3) provides an upper bound for the variance of a random variable about its mean.

Chebyshev's Inequality. Let $\mathbb{X}$ be a nonnegative random variable and m a positive real number. Then

$$\mathbb{P}(\big|\mathbb{X} - \mathbb{E}[\mathbb{X}]\big| \geq m) \leq \frac{\mathbb{V}[\mathbb{X}]}{m^2}.$$

Another key feature in analytic combinatorics is *convergence in distribution* or convergence in law which provides the relationship between combinatorial parameters and asymptotic properties. We say that a limit law exists for a parameter if there is convergence as n grows of the corresponding family of cumulative distribution functions. In this case of RNA secondary structures, the limit law is *discrete* to mean convergence is established without standardizing the random variable.

Definition 9. The discrete random variables $\mathbb{X}_n$ supported by $\mathbb{Z}_{\geq 0}$ are said to converge in law to a discrete random variable $\mathbb{Y}$ supported by $\mathbb{Z}_{\geq 0}$, written $\mathbb{X}_n \Rightarrow \mathbb{Y}$ if for each $k \geq 0$,

$$\lim_{n \to \infty} \mathbb{P}(\mathbb{X}_n \leq k) = \mathbb{P}(\mathbb{Y} \leq k).$$

A nice feature of limit laws of the discrete kind is their equivalence to local limit laws. There exists a *local limit law* if, for each $k \geq 0$,

$$\lim_{n\to\infty} \mathbb{P}(\mathbb{X}_n = k) = \mathbb{P}(\mathbb{Y} = k).$$

We use the framework outlined above to extract important information regarding the structure of RNA secondary structures in the next section.

Chapter 3

γ-structures

We begin with a biological inquiry. Can we specify the length of irreducible components of γ-structures? The set of γ-structures is translated to a combinatorial class where the objects of study are now graphs with properties generalized from known binding blocks. Through symbolic enumeration, information on the combinatorial class is transfered to formal generating functions. We are then in a position to extract coefficients of the generating functions which provide a count of these objects of a particular length. Again the objects of study are transfered from generating functions, or more specifically formal power series, to complex analytic functions. Through singularity analysis we are able to compute asymptotic estimates of the coefficients. Finally, we employ techniques from probabilistic graph theory to produce likelihood estimates to yield a more rich understanding of the behavior of the original biological objects.

First, we construct the generating function for τ-canonical γ-structures. Next we use tools from Analytic Combinatorics and Probabilistic Graph Theory to derive the mean and variance for the length of the longest block. Further results on the length-spectrum and uniqueness follow.

3.1 Combinatorics and Generating Functions

The main result of this section is the generating function for γ-structures. The theorems can be found in Han et al. [10], however the proofs differ slightly. Han et al. does not restrict γ to be 1 and thus diverges from the work below to derive the generating function for irreducible shadows using the generating function for matchings filtered by genus.

The process of deriving $\boldsymbol{G}_\tau(z)$ for arbitrary γ begins with deriving the OGF for γ-matchings,

Table 3.1: Classes and generating functions associated with γ-structures

Objects	Generating Function	Filtration
$\mathcal{T}$: blocks*	OGF: $\boldsymbol{T}(z) = \sum_{n\geq 0} t(n)z^n$	number of vertices
$\mathcal{I}_g$: irreducible shadows of genus g	OGF: $\boldsymbol{I}_g(u) = \sum_{m=2g}^{6g-2} i_g(m)u^m$	number of arcs
$\mathcal{H}$: γ-matchings	OGF: $\boldsymbol{H}(u) = \sum_{2m\geq 0} h(n)u^m$ BGF: $\boldsymbol{H}(u,e) = \sum_{2m,s\geq 0} h(m,s)u^m e^s$	number of arcs and 1-arcs
$\mathcal{P}$: γ-shapes	BGF: $\boldsymbol{P}(z,e) = \sum_{2m,s\geq 0} p(m,s)u^m x^s$	number of arcs and 1-arcs
$\mathcal{G}_\tau$: τ-canonical γ-structures	OGF: $\boldsymbol{G}_\tau(z) = \sum_{n\geq 0} g_\tau(n)z^n$	number of vertices

* The blocks referred to here are the irreducible components of γ-structures, which differ from the irreducible components of γ-matchings discussed in Lemma 6 and Theorem 8.

then the corresponding BGF that additionally filters by 1-arcs. From γ-matchings we derive the BGF for γ-shapes. Finally, from γ-shapes we inflate arcs to stacks and introduce isolated vertices. We specify γ for the purpose of singularity analysis and asymptotic estimation, however it is not necessary to do so when deriving the generating functions for shapes and structures as we will see shortly.

Lemma 6. *Let ζ be a fixed irreducible shadow with $m \geq 2$ arcs. The generating function for blocks of γ-matchings whose maximal component is ζ filtered by the number of arcs is given by*

$$\boldsymbol{B}_\zeta(u) = \left(\frac{u}{1-u\boldsymbol{H}(u)^2}\right)^m \boldsymbol{H}(u)^{2m-1}. \tag{3.1}$$

Proof. $\boldsymbol{B}_\zeta(u)$ is obtained via symbolic enumeration by specifying the process of inflation of ζ to an arbitrary block. First observe that ζ contains $2m-1$ σ-intervals. A (possibly empty) γ-matching can be placed in each of these σ-intervals. This is represented symbolically by $\mathcal{H}^{2m-1}$. Note that the number of σ-intervals is preserved when inflating shadows to matchings.

By inflating the arcs of ζ in a particular way, we arrive at an expression for the nesting of γ-matchings into the P-intervals of an irreducible matching. Under each arc, nest an arc

with a (possibly empty) γ-matching attached to the outside of either end of the arc. A nested sequence of this kind,

$$\mathrm{SEQ}(\mathcal{R} \times \mathcal{H}^2),$$

produces the desired inflation (Fig. 3.1). Note that allowing for empty matchings on both ends of the inflated arc also produces stacks.

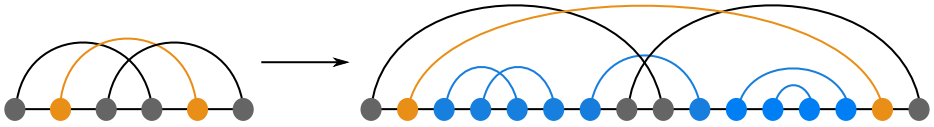


Figure 3.1: The (orange) arc is inflated.

The symbolic enumeration of all blocks with fixed irreducible shadow ζ is

$$\mathcal{B}_\zeta = (\mathcal{R} \times \mathrm{SEQ}(\mathcal{R} \times \mathcal{H}^2))^m \times \mathcal{H}^{2m-1}.$$

The generating function for the class of arcs $\mathcal{R}$ is $\boldsymbol{R}(u) = u$. Thus the generating function for blocks of γ-matchings whose maximal component is determined by ζ is

$$\boldsymbol{B}_\zeta(u) = \left(\frac{u}{1 - u\boldsymbol{H}(u)^2}\right)^m \boldsymbol{H}(u)^{2m-1}.$$

□

Remark. The enumeration of blocks with maximal shadow ζ depends only on the number of arcs m, and not on genus or the arrangement of arcs in the shadow. Thus the enumeration of blocks is the same for any irreducible shadow containing the same number of arcs. As such, the generating function for irreducible shadows arises in the generating function for γ-matchings.

The power of this construction stems from the realization that the generating function for irreducible shadows is a polynomial (as opposed to an infinite series). Lemma 7 below implies that the generating function for irreducible shadows of genus g is bounded below by $2g$ and above by $6g - 2$. Without this property, an explicit functional form for $\boldsymbol{H}(u)$ could not be determined.

Lemma 7. *(Anderson et al. [1]) A shadow of genus $g \geq 1$ has the following properties: (1) a shadow of genus g contains at least $2g$ and at most $6g - 2$ arcs, and (2) for any $2g \leq m \leq 6g - 2$, there exists a shadow of genus g containing exactly m arcs.*

Theorem 8. *(Han et al. [11]) The generating function for γ-matchings satisfies*

$$\boldsymbol{H}(u)^{-1} = 1 - \left(u\boldsymbol{H}(u) + \boldsymbol{H}(u)^{-1} \sum_{g=1}^{\gamma} \boldsymbol{I}_g\left(\frac{u\boldsymbol{H}(u)^2}{1 - u\boldsymbol{H}(u)^2}\right)\right),$$

or equivalently,

$$\boldsymbol{H}(u) - u\boldsymbol{H}(u)^2 - \sum_{g=1}^{\gamma} \boldsymbol{I}_g\Big(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\Big)\Big) = 1 \tag{3.2}$$

Proof. As previously mentioned, a γ-matching can be decomposed into a sequence of concatenated blocks.

$$\mathcal{H} = \mathrm{SEQ}(\mathcal{B})$$

Consider the partition $\mathcal{B} = \mathcal{B}_1 \sqcup \mathcal{B}_2$ of the class of blocks distinguishing a block's unique maximal component:

· $\mathcal{B}_1$: blocks whose maximal component is a single arc (Figure 3.2a),
· $\mathcal{B}_2$: blocks whose maximal component is an irreducible matching (Figure 3.2b).

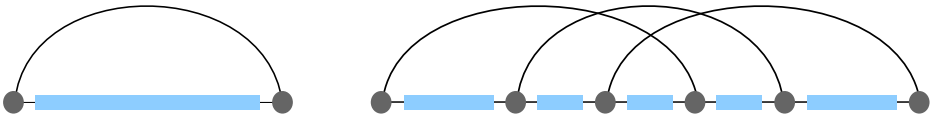


Figure 3.2: (left) The structure of blocks in class $\mathcal{B}_1$, and (right) an example of the structure of a block in class $\mathcal{B}_2$. The blue bar represents the position of γ-matchings nested in their maximal components.

For blocks in $\mathcal{B}_1$, the nested component is precisely a γ-matching. This is represented symbolically by

$$\mathcal{B}_1 = \mathcal{R} \times \mathcal{H},$$

and the associated generating function is $\boldsymbol{B}_1(u) = u\boldsymbol{H}(u)$.

Now consider blocks in $\mathcal{B}_2$. The generating function for blocks with fixed maximal shadow ζ is given by Eq. 3.1 in Lemma 6 as

$$\boldsymbol{B}_\zeta(u) = \left(\frac{u}{1-u\boldsymbol{H}(u)^2}\right)^m \boldsymbol{H}(u)^{2m-1}.$$

Since the generating function for blocks is the same across maximal shadows with the same number of arcs, summing over all possible genus and number m of arcs in an irreducible shadow gives the generating function for blocks in class $\mathcal{B}_2$:

$$\boldsymbol{B}_2(u) = \sum_{g=1}^{\gamma} \sum_{m=2g}^{6g-2} i_g(m) \left(\frac{u}{1-u\boldsymbol{H}(u)^2}\right)^m \boldsymbol{H}(u)^{2m-1}.$$

where $i_g(m)$ is the number of irreducible shadows with m arcs.

The empty diagram is taken to be a matching. Therefore $[z^0]\boldsymbol{H}(u) = 1$ and $\boldsymbol{H}(u)$ has an inverse in the ring of formal power series $\mathbb{Q}[[u]]$. Since $\mathcal{H} = \mathrm{SEQ}(\mathcal{B}_1 + \mathcal{B}_2)$, the generating function for γ-matchings is

$$\begin{aligned}\boldsymbol{H}(u)^{-1} &= 1 - \Big(u\boldsymbol{H}(u) + \sum_{g=1}^{\gamma}\sum_{m=2g}^{6g-2} i_g(m)\left(\frac{u}{1-u\boldsymbol{H}(u)^2}\right)^m \boldsymbol{H}(u)^{2m-1}\Big)\\ &= 1 - \Big(u\boldsymbol{H}(u) + \boldsymbol{H}(u)^{-1}\sum_{g=1}^{\gamma}\sum_{m=2g}^{6g-2} i_g(m)\left(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\right)^m\Big)\end{aligned}$$

The generating function for irreducible shadows is

$$\boldsymbol{I}_g(u) = \sum_{m=2g}^{6g-2} i_g(m)u^m.$$

Thus the inner sum in the generating function for γ-matchings can be realized as the composition

$$\boldsymbol{I}_g\left(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\right) = \sum_{m=2g}^{6g-2} i_g(m)\left(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\right)^m.$$

This observation and the following algebraic manipulation takes us to the form of $\boldsymbol{H}(u)$ given in the statement of the theorem.

$$\begin{aligned}\boldsymbol{H}(u)^{-1} &= 1 - \Big(u\boldsymbol{H}(u) + \boldsymbol{H}(u)^{-1}\sum_{g=1}^{\gamma}\sum_{m=2g}^{6g-2} i_g(m)\left(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\right)^m\Big)\\ \boldsymbol{H}(u)^{-1} &= 1 - \Big(u\boldsymbol{H}(u) + \boldsymbol{H}(u)^{-1}\sum_{g=1}^{\gamma}\boldsymbol{I}_g\Big(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\Big)\Big)\\ 1 &= \boldsymbol{H}(u) - \boldsymbol{H}(u)\Big(u\boldsymbol{H}(u) + \boldsymbol{H}(u)^{-1}\sum_{g=1}^{\gamma}\boldsymbol{I}_g\Big(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\Big)\Big)\\ 1 &= \boldsymbol{H}(u) - u\boldsymbol{H}(u)^2 - \sum_{g=1}^{\gamma}\boldsymbol{I}_g\Big(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\Big)\Big)\end{aligned}$$

□

Remark. Note that γ-matchings have no restriction on 1-arcs. Since γ-structures do not allow 1-arcs, it is necessary to identify all 1-arcs in a given γ-matching in order to ensure they are eliminated during the inflation process. The following generating function for γ-matchings additionally filters by 1-arcs.

Theorem 9. *(Han et al., 2012) The bivariate generating function $\boldsymbol{H}(u,e)$ for γ-matchings that marks the number of 1-arcs by the parameter e is given by*

$$\boldsymbol{H}(u,e) = \frac{1}{u+1-eu}\boldsymbol{H}\left(\frac{u}{(u+1-eu)^2}\right). \tag{3.3}$$

Proof. We begin by defining a recursion for the number of γ-matchings with a labeled 1-arc. A PDE equivalent to this recursion is derived and a solution $\tilde{\boldsymbol{H}}(u,e)$ is determined. We then show that $\tilde{\boldsymbol{H}}(u,e) = \boldsymbol{H}(u,e)$.

The number of γ-matching with $m+1$ arcs and $s+1$ 1-arcs with one labeled 1-arc is

$$(s+1)h(m+1,s+1).$$

Each of these marked γ-matchings can be formed from γ-matching with m arcs by inserting (and marking) a 1-arc between any of the $2m-1$ adjacent pairs of vertices, or on either end of the sequence. Either the 1-arc is inserted underneath an existing 1-arc, or not. If it is inserted underneath an existing 1-arc, the resulting γ-matching has 1 additional arc and no additional 1-arcs since the inserted 1-arc replaces a 1-arc in the original structure. The original γ-matching has $s+1$ 1-arcs, thus the number of places to insert the nested 1-arc is $s+1$. The count of such labeled γ-matchings is

$$(s+1)h(m,s+1).$$

Suppose the 1-arc is not inserted underneath an existing 1-arc. Then there are s original 1-arcs and the number of possible insertion points is $(2m+1-s)$. The resulting structure has one additional arc contributing to the the count of arcs and 1-arcs. Then the count of all such markings is

$$(2m+1-s)h(m,s).$$

Thus the recursion on γ-matchings with a labeled 1-arc is

$$(s+1)h(m+1,s+1) = (s+1)h(m,s+1) + (2m+1-s)h(m,s).$$

Notice that the terms in the recurrence relation resemble the coefficients of terms in the derivatives of $\boldsymbol{H}(u,e)$. Toward deriving a differential equation whose coefficients exactly match the recurrence relation at each m and s, consider the partial derivatives

$$\frac{\partial}{\partial u}\big(\boldsymbol{H}(u,e)\big) = \sum_{2m,s\geq 0} mh(m,s)u^{m-1}e^s$$
$$\frac{\partial}{\partial e}\big(\boldsymbol{H}(u,e)\big) = \sum_{2m,s\geq 0} sh(m,s)u^{m}e^{s-1}$$

Extracting coefficients gives us

$$\begin{aligned}(s+1)h(m+1,s+1) &= [u^{m+1}e^s]\frac{\partial}{\partial e}\big(\boldsymbol{H}(u,e)\big)\\ (s+1)h(m,s+1) &= [u^m e^s]\frac{\partial}{\partial e}\big(\boldsymbol{H}(u,e)\big)\\ mh(m,s) &= [u^{m-1}e^s]\frac{\partial}{\partial u}\big(\boldsymbol{H}(u,e)\big)\\ h(m,s) &= [u^m e^s]\boldsymbol{H}(u,e)\\ sh(m,s) &= [u^m e^{s-1}]\frac{\partial}{\partial e}\big(\boldsymbol{H}(u,e)\big)\end{aligned}$$

By shifting the coefficients we can extract the terms of the recurrence from the power $u^{m+1}e^s$, which in turn gives the PDE equivalent to the recurrence relation,

$$(1-u+eu)\frac{\partial}{\partial e}\big(\boldsymbol{H}(u,e)\big) = 2u^2\frac{\partial}{\partial u}\big(\boldsymbol{H}(u,e)\big) + u\boldsymbol{H}(u,e). \tag{3.4}$$

Claim:

$$\tilde{\boldsymbol{H}}(u,e) = \frac{1}{u+1-eu}\boldsymbol{H}\left(\frac{u}{(u+1-eu)^2}\right)$$

is a solution to (3.4) and $\tilde{\boldsymbol{H}}(u,e) = \boldsymbol{H}(u,e)$.

First consider the partial derivatives of $\tilde{\boldsymbol{H}}(u,e)$. For ease of reading, label $y = (u+1-eu)^{-2}$.

$$\begin{aligned}\frac{\partial}{\partial u}\big(\tilde{\boldsymbol{H}}(u,e)\big) &= (e-1)y\boldsymbol{H}(uy) + y^2(eu-u+1)\boldsymbol{H}'(uy)\\ \frac{\partial}{\partial e}\big(\tilde{\boldsymbol{H}}(u,e)\big) &= uy\boldsymbol{H}(uy) + 2(uy)^2\boldsymbol{H}'(uy)\end{aligned}$$

Plugging the potential solution into the PDE,

$$\begin{aligned}(1-u+eu)\big(uy\boldsymbol{H} + 2(uy)^2\boldsymbol{H}'\big) &= 2u^2\big((e-1)y\boldsymbol{H} + y^2(eu-u+1)\boldsymbol{H}'\big) + \frac{u\boldsymbol{H}}{u+1-eu}\\ (1-u+eu)uy\boldsymbol{H} &= 2u^2(e-1)y\boldsymbol{H} + \frac{u\boldsymbol{H}}{u+1-eu}\\ (1-u+eu)y &= 2u(e-1)y + 1.\end{aligned}$$

The final equality holds by replacing y with its original form and equating the numerators of the left and right hand sides.

To prove that $\tilde{\boldsymbol{H}}(u,e) = \boldsymbol{H}(u,e)$, we must show that $\tilde{h}(m,s) = h(m,s)$ for all $m,s \geq 0$. First note that for $m > s$, $\tilde{h}(m,s) = 0$—whenever e appears in $\tilde{\boldsymbol{H}}(u,e)$ it is in the product

eu and thus a power of e greater than the power of u is impossible. We are also granted

$$\sum_{s\geq 0} \tilde{h}(m,s) = \sum_{s\geq 0} h(m,s) \tag{3.5}$$

since $\tilde{\boldsymbol{H}}(u,1) = \boldsymbol{H}(u)$.

Consider the coefficients of $\tilde{\boldsymbol{H}}(u,e)$ as an array indexed by m and s (see array below). It is known that the array is upper triangular and that the columns of the array sum to $\sum_{s\geq 0} h(m,s)$. In terms of the recursion, any element of the array is calculated by adding the term immediately to the left and that term's neighbor above. This implies that the array can be filled in column-wise from left to right with information on the first column. Namely, if the first column ($m = 0$) of $\tilde{\boldsymbol{H}}(u,e)$ and $\boldsymbol{H}(u,e)$ match, since they follow the same recursion, the arrays will match in every entry.

For $m = 0$ the only potentially nonzero term is $\tilde{h}(0,0)$. Thus Eq. 3.5 implies $\tilde{h}(0,0) = h(0,0)$. The result follows.

$\tilde{h}(m,s)$	$m=0$	1	2	3	$\cdots$	
$s=0$	1					
1		1		$\tilde{h}(m-1,s-1)$		
2			1	$\tilde{h}(m-1,s)$		$\tilde{h}(m,s)$
3				1		
$\vdots$						$\ddots$

□

We are now in a position to derive the bivariate generating function for γ-shapes filtered by 1-arcs.

Theorem 10. *(Han et al., 2012) The bivariate generating function $\boldsymbol{P}(z,e)$ for γ-shapes that marks the number of 1-arcs by the parameter e is given by*

$$\boldsymbol{P}(z,e) = \frac{1+z}{1+2z-ez}\boldsymbol{H}\left(\frac{z(1+z)}{(1+2z-ez)^2}\right). \tag{3.6}$$

Proof. Consider a fixed shape σ with m arcs and s one arcs. Moving from a shape to a matching amounts to inflating each arc of the shape into a stack, a nested sequence of arcs.

Since inflation preserves 1-arcs, the number of γ-matchings with shape σ is

$$\boldsymbol{H}^{\sigma}(u,e) = \left(\frac{u}{1-u}\right)^{m} e^{s}.$$

The generating function for γ-matchings can be rewritten

$$\boldsymbol{H}(u,e) = \sum_{s\geq 0}\sum_{\sigma\in\mathcal{P}(s)} \boldsymbol{H}^{\sigma}(u,e) = \sum_{m\geq 0}\sum_{s\geq 0} p(m,s)\left(\frac{u}{1-u}\right)^{m} e^{s}.$$

For

$$\boldsymbol{P}(z,e) = \sum_{m\geq 0}\sum_{s\geq 0} p(m,s) z^{m} e^{s},$$

the change of variable $z = \frac{u}{1-u}$ produces the equality

$$\boldsymbol{P}(z,e) = \boldsymbol{H}\Big(\frac{z}{1+z}, e\Big) = \frac{1+z}{1+2z-ez}\boldsymbol{H}\left(\frac{z(1+z)}{(1+2z-ez)^2}\right).$$

□

Lemma 11. *(Han et al., 2012) Let σ be a fixed γ-shape with $m \geq 1$ arcs and $s \geq 0$ 1-arcs. The bivariate generating function for τ-canonical γ-structures that have shape σ is given by*

$$\boldsymbol{G}_{\tau}^{\sigma}(z) = \frac{1}{1-z}\left(\frac{z^{2\tau}}{(1-z^2)(1-z)^2 - (2z-z^2)z^{2\tau}}\right)^{m} z^{s}. \tag{3.7}$$

Proof. Recall that γ-structures do not contain 1-arcs, but possibly contain isolated vertices. To be τ-canonical is to have a minimum stack length τ. The proof follows a construction similar to γ-matching wherein we considered inflation of arcs of irreducible shadows in a particular way. Here we introduce the notion of induced arcs again, however the objects concatenated to the arc are sequences of isolated vertices and lie inside the arc. To avoid overcounting, we require that at least one of the two sequences of isolated vertices is nonempty.

Consider a fixed shadow σ with m arcs and s 1-arcs. To arrive at a τ-canonical γ-structure one inflates each arc by stacking induced arcs on top of the original arc, then inflates each of the resulting arcs to a stack, and inserts isolated vertices underneath the original arcs.

The generating function for the class of vertices $\mathcal{Z}$ and arcs $\mathcal{R}$ are $\boldsymbol{Z}(z) = z$ and $\boldsymbol{R}(z) = z^2$, and a sequence of isolated vertices is represented symbolically and in a generating function by

$$\mathcal{L} = \mathrm{SEQ}(\mathcal{Z}), \quad \boldsymbol{L}(z) = \frac{1}{1-z},$$

respectively. Thus an induced arc is represented by

$$\mathcal{N} = (\mathcal{R} \times \mathcal{L} + \mathcal{R} \times \mathcal{L} + \mathcal{R} \times \mathcal{L}^2)$$

The generating function for a nested sequence of induced arcs lying above an original arc is

$$z^2 \left(1 - \left(z^2 \frac{z}{1-z} + z^2 \frac{z}{1-z} + z^2 \Big(\frac{z}{1-z}\Big)^2\right)\right)^{-1} \tag{3.8}$$

Next, each arc in the induced stack is inflated to a τ-canonical stack. Eq. 3.8 is then transformed into

$$\frac{z^{2\tau}}{1-z^2} \left(1 - \frac{z^{2\tau}}{1-z^2}\Big(\frac{z}{1-z} + \frac{z}{1-z} + \Big(\frac{z}{1-z}\Big)^2\Big)\right)^{-1} \tag{3.9}$$

Finally, sequences of isolated vertices are inserted between any two sets of original vertices plus the two endpoints. There are $2m-1$ pairs of adjacent vertices, s of which correspond to 1-arcs. Sequences nested under 1-arcs cannot be empty to ensure no 1-arcs appear in the final structure. The corresponding generating function for these inserted sequences is

$$\left(\frac{1}{1-z}\right)^{2m+1-s} \left(\frac{z}{1-z}\right)^s = z^s \frac{1}{1-z} \left(\frac{1}{1-z}\right)^{2m} \tag{3.10}$$

Combining Eqs. 3.9 and 3.10, we arrive at the generating function for a τ-canonical γ-structures with fixed shape σ is

$$\boldsymbol{G}_\tau^\sigma(z) = z^s \frac{1}{1-z} \left(\frac{1}{1-z}\right)^{2m} \left(\frac{z^{2\tau}}{1-z^2} \left(1 - \frac{z^{2\tau}}{1-z^2}\Big(\frac{z}{1-z} + \frac{z}{1-z} + \Big(\frac{z}{1-z}\Big)^2\Big)\right)^{-1}\right)^m \tag{3.11}$$

The following algebraic manipulation gives us Eq. 3.11. In order to preserve readability, we only do computation on the two rightmost terms of the product since the first two correspond to terms in Eq. 3.7.

$$\begin{aligned}
&\left(\frac{1}{1-z}\right)^{2m} \left(\frac{z^{2\tau}}{1-z^2} \left(1 - \frac{z^{2\tau}}{1-z^2}\Big(\frac{z}{1-z} + \frac{z}{1-z} + \Big(\frac{z}{1-z}\Big)^2\Big)\right)^{-1}\right)^m \\
&\quad = \frac{\frac{z^{2\tau}}{(1-z^2)(1-z)^2}}{1 - \frac{z^{2\tau}}{1-z^2}\Big(\frac{z}{1-z} + \frac{z}{1-z} + \Big(\frac{z}{1-z}\Big)^2\Big)} \\
&\quad = \frac{z^{2\tau}}{(1-z^2)(1-z)^2 - z^{2\tau}(2z-z^2)}
\end{aligned}$$

□

Remark. Lemma 11 implies that $\boldsymbol{G}_\tau^\sigma(z)$ is only dependent on the number of arcs and 1-arcs, and thus is the same across shapes of a particular length. This property allows us to sum across all shapes in such a way that the generating function for γ-shapes reappears in the generating function for γ-structures.

Theorem 12. *(Han et al., 2012) The generating function for τ-canonical γ-structures is given by*

$$\boldsymbol{G}_\tau(z) = \frac{1}{(1-z)+u_\tau(z)z^2}\boldsymbol{H}\left(\frac{z^2u_\tau(z)}{((1-z)+u_\tau(z)z^2)^2}\right) \tag{3.12}$$

where $u_\tau(z) = \frac{z^{2(\tau-1)}}{z^{2\tau}-z^2+1}$.

Proof. We are given $\boldsymbol{G}_\tau^\sigma(z)$ in Lemma 11. Summing across all possible shapes, we arrive at the following form of $\boldsymbol{G}_\tau(z)$.

$$\begin{aligned}
\boldsymbol{G}_\tau(z) &= \sum_{s\geq 0}\sum_{\sigma\in\mathcal{P}(s)} \boldsymbol{G}_\tau^\sigma(z) \\
&= \frac{1}{1-z}\sum_{m\geq 0}\sum_{s=0}^{m} p(m,s)\left(\frac{z^{2\tau}}{(1-z^2)(1-z)^2-(2z-z^2)z^{2\tau}}\right)^m z^s \\
&= \frac{1}{1-z}\boldsymbol{P}\left(\frac{z^{2\tau}}{(1-z^2)(1-z)^2-z^{2\tau}(2z-z^2)}, z\right)
\end{aligned}$$

From Theorem 10, $\boldsymbol{P}(z,e)$ can be written in terms of $\boldsymbol{H}(z)$. Making this substitution gives us Eq. 3.12. □

We are now in a position to derive asymptotics for γ-structures. Going forward we only consider $\gamma = 1$.

3.2 Asymptotics of γ-structures

When an explicit functional form is given for a generating function, it is possible to use techniques from complex analysis to come up with a singular expansion that is then used to extract coefficient asymptotics. However, it is often the case that an explicit functional form for a generating function is unknown. This is the case for γ-structures where finding an explicit form amounts to solving a polynomial of degree 10.

The good news is that an implicit functional form is sufficient for extracting coefficient asymptotics. The following lemma provides an implicit functional form for $\boldsymbol{H}(u)$.

Lemma 13. *There exists a polynomial $P(u,X) \in R[X]$ such that $P(u, \boldsymbol{H}(u)) = 0$.*

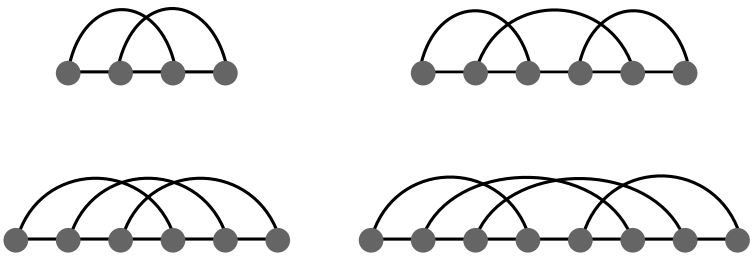


Figure 3.3: Irreducible shadows of genus $g = 1$.

Proof. Recall Eq. 3.2 from the previous section,

$$\boldsymbol{H}(u) - u\boldsymbol{H}(u)^2 - \sum_{g=1}^{\gamma} \boldsymbol{I}_g\Big(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\Big)\Big) = 1.$$

For $\gamma = 1$, the generating function for irreducible shadows is $\boldsymbol{I}_1 = z^2 + 2z^3 + z^4$ (see Figure. 3.3). Then the functional form of $\boldsymbol{H}(u)$ for $\gamma = 1$ is given by

$$\boldsymbol{H}(u) - \boldsymbol{H}(u)^2 - \Big(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\Big)^2 + 2\Big(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\Big)^3 + \Big(\frac{u\boldsymbol{H}(u)^2}{1-u\boldsymbol{H}(u)^2}\Big)^4 = 1 \tag{3.13}$$

By manipulating Eq. 3.13 so that the LHS is a polynomial in u and $\boldsymbol{H}(u)$ set equal to 0, we can extract the polynomial $P(u, X)$ such that $P(u, \boldsymbol{H}(u)) = 0$. To arrive at $P(u, \boldsymbol{H}(u))$ one multiplies each side of Eq. 3.13 by $(1 - u\boldsymbol{H}(u)^2)^4$.

$$P(u, X) = -1 + X + 3X^2u - 4X^3u - 3X^4u^2 + 6X^5u^2 - 2X^6u^3 - 4X^7u^3 + 3X^8u^4 + X^9u^4 - X^{10}u^5 \tag{3.14}$$

□

It is not necessary to find the roots of $P(u, X)$ explicitly as seen in Section 2.2. Instead we use the implicit-function schema. To derive the asymptotics for $\mathbb{P}(\mathbb{B}_n = n - k)$ we need the following singular expansions and corresponding transfers.

Remark. Han et al. [10] includes a theorem stating that the dominant singularity μ_H of $\boldsymbol{H}(u)$ is unique and a root of the resultant,

$$\Delta(u) = \boldsymbol{R}\left(P(u, X), D_X(P(u, X)), X\right).$$

We use this fact to compute μ_H explicitly using Mathematica in order to employ the implicit function schema.

Lemma 14. *The singular expansions for $\boldsymbol{H}(u)$ is given by*

$$\boldsymbol{H}(u) = \delta_0 + \delta_1(\mu_H - u)^{\frac{1}{2}} + O(\mu_H - u), \quad z \to \mu_H \tag{3.15}$$

where $\delta_0 = \boldsymbol{H}(\mu_H)$, $\delta_1 = \mu_H^{-\frac{1}{2}}\sqrt{\frac{2\mu_H P_u(\mu_H, \boldsymbol{H}(\mu_H))}{P_{XX}(\mu_H, \boldsymbol{H}(\mu_H))}}$. *In addition, the asymptotics of the coefficients of* $\boldsymbol{H}(u)$ *are*

$$[u^m]\boldsymbol{H}(u) = \tilde{c}m^{-\frac{3}{2}}\mu_H^{-m}(1 + O(m^{-1})). \tag{3.16}$$

where $\tilde{c} = \delta_1 \mu_H^{\frac{1}{2}}\Gamma(-\frac{1}{2})^{-1}$.

Proof. Consider the bivariate function $P(u, X)$ given by Eq. 3.13. This function is a polynomial and consequently analytic at $(\mu_H, \boldsymbol{H}(\mu_H))$. By checking that the conditions of Theorem 4 hold for $P(u, X)$, we are given the asymptotic expansion with the square root factor. The four conditions are

$$P(\mu_H, \boldsymbol{H}(\mu)) = 0, \quad P_u(\mu_H, \boldsymbol{H}(\mu_H)) \neq 0, \quad P_X(\mu_H, \boldsymbol{H}(\mu_H)) = 0, \quad P_{XX}(\mu_H, \boldsymbol{H}(\mu_H)) \neq 0.$$

We check these conditions by explicitly computing the relevant partial derivatives and plugging in μ_H and $\boldsymbol{H}(\mu_H)$ to each of them. Again the asymptotics follow from Theorem 2.

□

Lemma 15. *The singular expansions for* $\boldsymbol{G}_\tau(z)$ *is given by*

$$\boldsymbol{G}_\tau(z) = \theta_0 + \theta_1(\mu - z)^{\frac{1}{2}} + \theta_2(\mu - z) + O((\mu - z)^{\frac{3}{2}}), \quad z \to \mu \tag{3.17}$$

where $\theta_0 = \boldsymbol{G}_\tau(\mu)$. *In addition, the asymptotics of the coefficients of* $\boldsymbol{G}_\tau(z)$ *are*

$$[z^n]\boldsymbol{G}_\tau(z) = cn^{-\frac{3}{2}}\mu^{-n}(1 + O(n^{-1})). \tag{3.18}$$

where $c = -\theta_1 \mu^{\frac{1}{2}}\Gamma(-\frac{1}{2})^{-1}$.

Proof. Recall from Theorem 12 the expression for $\boldsymbol{G}_\tau(z)$ in terms of $\boldsymbol{H}(u)$ given by Eq. 3.12,

$$\boldsymbol{G}_\tau(z) = \frac{1}{(1 - z) + u_\tau(z)z^2}\boldsymbol{H}\left(\frac{z^2 u_\tau(z)}{((1 - z) + u_\tau(z)z^2)^2}\right)$$

where $u_\tau(z) = \frac{z^{2(\tau-1)}}{z^{2\tau} - z^2 + 1}$.

Let the inner function composed with $\boldsymbol{H}(u)$ be labeled $\theta(z)$. First note that $\boldsymbol{H}(\theta(z))$ is analytic at the origin with non-negative coefficients. By Pringsheim's Theorem it must have a dominant singularity that lies on the positive real axis.

The candidates for its dominant singularities are the dominant singularities of $\theta(z)$ and the solutions to $\theta(z) = \mu_H$. It is simple to check that a dominant singularity of $\theta(z)$ does not lie on the positive real line. Therefore, it must be the case that a solution to $\theta(z) = \mu_H$ smallest in modulus must fall on the positive real line. By inspection, the positive real-valued solution

μ to $\theta(z) = \mu_H$ smallest in modulus is unique. Thus $\boldsymbol{H}(\theta(z))$ falls under the supercritical composition schema.

The singular expansion for $\boldsymbol{H}(\theta(z))$ is computed as the singular expansion of $\boldsymbol{H}(u)$ at $z = \mu_H$ with the Taylor expansion of $\theta(z)$ at $z = \mu$. Recall from Lemma 14 the singular expansion of $\boldsymbol{H}(u)$,

$$\boldsymbol{H}(u) = \delta_0 + \delta_1(\mu_H - u)^{\frac{1}{2}} + O(\mu_H - u).$$

Then the composition with the Taylor series of $\theta(z)$ is

$$\begin{aligned}\boldsymbol{H}(\theta(z)) &= \delta_0 + \delta_1(\mu_H - (\mu_H + \theta'(\mu)(z-\mu) + O(z-\mu)))^{\frac{1}{2}} + O(\mu - z)\\ &= \delta_0 + \delta_1\theta'(\mu)(\mu - z)^{\frac{1}{2}} + O(\mu - z)\end{aligned}$$

To arrive at the singular expansion of $\boldsymbol{G}_\tau(z)$, we note that the singularities of the factor $\frac{1}{(1-z)+u_\tau(z)z^2}$ are a subset of the singularities of $\theta(z)$ which we have proven to be larger in modulus than μ. Thus the dominant singularity of $\boldsymbol{G}_\tau(z)$ is also μ and its singular expansion $\boldsymbol{G}_\tau(z)$ at $z = \mu$ is the product of the Taylor expansion of $\frac{1}{(1-z)+u_\tau(z)z^2}$ and the singular expansion of $\boldsymbol{H}(\theta(z))$. The result follows. □

3.3 The Longest Block

The main result of this section is a precise statement regarding the expected length of the longest irreducible component (block) of γ-structures for $\gamma = 1$ as the sequence length gets arbitrarily large. The method used to derive the expected length parallels the work of Li and Reidys in [15] for secondary structure. To this end, we define the random variable $\mathbb{B}_n$ representing the length of the longest block in a structure of length n.

The average length of the longest rainbow is given by the expectation of the random variable $\mathbb{B}_n$, labeled $\mathbb{E}[\mathbb{B}_n]$. The lemmas and theorems that follows provide the derivation of the expectation $\mathbb{E}[\mathbb{B}_n]$ and variance $\mathbb{V}[\mathbb{B}_n]$, found to be on the order of $n - O(n^{\frac{1}{2}})$ and $O(n^{\frac{3}{2}})$ respectively. In particular, we show that the expected length of the longest rainbow is $n - \alpha n^{\frac{1}{2}}(1 + o(1))$ with standard deviation $\sqrt{\beta}n^{\frac{3}{4}}(1 + o(1))$.

The combinatorial class $\mathcal{G}_{\tau,n}$ is the set of all τ-canonical γ-structures of length n now considered as a sample space. We take the uniform probability function to form the finite probability space $(\mathcal{G}_{\tau,n}, \mathbb{P})$ such that $\mathbb{P}(S) = \frac{1}{\mathrm{card}(\mathcal{G}_{\tau,n})}$ for any $S \in \mathcal{G}_{\tau,n}$.

Consider the general definition of expectation given the discrete random variable $\mathbb{B}_n$:

$$\mathbb{E}[\mathbb{B}_n] = \sum_{\omega \in \Omega} \mathbb{B}_n(\omega)\mathbb{P}(\omega)$$

where ω is an event in the sample space Ω. We choose to partition the sample space into disjoint events $\mathbb{B}_n = n - k$ for $1 \leq k \leq n$. As such, $\mathbb{E}[\mathbb{B}_n]$ can be rewritten

$$\mathbb{E}[\mathbb{B}_n] = \sum_{k=1}^{n} (n-k)\mathbb{P}(\mathbb{B}_n = n - k)$$

where $\mathbb{P}(\mathbb{B}_n = n - k)$ is the probability that a γ-structure of length n has longest block of length $n - k$ since $\bigcup_{1 \leq k \leq n}(\mathbb{B}_n = n - k) = \mathcal{G}_{\tau,n}$. The following derivation of $\mathbb{P}(\mathbb{B}_n = n - k)$ in terms of generating function coefficients allows us to compute the expectation and variance through means of transfer theorems and singular expansions of the generating functions derived in the previous sections.

The probability that a given γ-structure of length n has longest block of length $n - k$ is the ratio of the number of structures with longest block length $n - k$ to the total number of structures, assuming structures are uniformly sampled. In terms of coefficients of generating functions, this ratio is the coefficient of some unknown generating function over $[z^n]\boldsymbol{G}_\tau(z)$. The generating function for γ-structures with longest block $n-k$ proves to be too difficult to compute directly as such a task requires a complex sequence of inclusion/exclusion arguments that do not lend themselves well to singularity analysis. As such, we note the following. If $\boldsymbol{G}_{\leq m}(z)$ is the generating function for γ-structures whose blocks are of length less at most to m, then the generating function for structures with longest block length exactly m is

$$\boldsymbol{G}_{\leq m}(z) - \boldsymbol{G}_{\leq m-1}(z).$$

Therefore,

$$\mathbb{P}(\mathbb{B}_n = n - k) = \frac{[z^n](\boldsymbol{G}_{\leq n-k}(z) - \boldsymbol{G}_{\leq n-k-1}(z))}{[z^n]\boldsymbol{G}_\tau(z)}. \tag{3.19}$$

It suffices to show how $\boldsymbol{G}_{\leq m}(z)$ can be expressed in terms of known generating functions. In the previous section we constructed the generating function $\boldsymbol{G}_\tau(z)$ from inflating γ-matchings. Since we were able to derive a closed form for $\boldsymbol{H}(u)$, we were also granted a closed form for $\boldsymbol{G}_\tau(z)$ and thus computed the asymptotics of its coefficients. However, it is now advantageous to consider γ-structures as sequences of blocks. Let $\boldsymbol{T}(z)$ be the generating function for these irreducible components of γ-structures. Then the generating function $\boldsymbol{G}_\tau(z)$ can be realized as

$$\boldsymbol{G}_\tau(z) = \sum_{i \geq 0} (\boldsymbol{T}(z))^i = \frac{1}{1 - \boldsymbol{T}(z)}.$$

Toward deriving an expression for $\mathbb{P}(\mathbb{B}_n = n-k)$ in terms of coefficients of known generating functions, let

$$\boldsymbol{T}_{\leq m}(z) = \sum_{n=0}^{m} t_n z^n$$

be the truncated series representing the generating function for blocks of length at most m. Then the generating function for γ-structures with block length *at most* m is given by

$$\boldsymbol{G}_{\leq m}(z) = \sum_{i\geq 0} (\boldsymbol{T}_{\leq m}(z))^i = \frac{1}{1-\boldsymbol{T}_{\leq m}(z)}.$$

From here we are able to realize $[z^n](\boldsymbol{G}_{\leq n-k}(z) - \boldsymbol{G}_{\leq n-k-1}(z))$ as the product of coefficients of generating functions with known coefficient asymptotics. The following Lemma gives this alternative form.

Lemma 16. *For* $k \leq \frac{n}{2} - 1$*,*

$$\mathbb{P}(\mathbb{B}_n = n-k) = \frac{[z^k]\Phi'(\boldsymbol{T}(z))[z^{n-k}]\boldsymbol{T}(z)}{[z^n]\boldsymbol{G}_\tau(z)} \tag{3.20}$$

where $\Phi(z) = \frac{1}{1-z}$*.*

Proof. To arrive at Eq. 3.20, first recall Eq. 3.19,

$$\mathbb{P}(\mathbb{B}_n = n-k) = \frac{[z^n](\boldsymbol{G}_{\leq n-k}(z) - \boldsymbol{G}_{\leq n-k-1}(z))}{[z^n]\boldsymbol{G}_\tau(z)}.$$

Consider the Taylor expansions of $\boldsymbol{G}_{\leq n-k}(z)$ and $\boldsymbol{G}_{\leq n-k-1}(z)$ as the compositions $\Phi(\boldsymbol{T}_{\leq n-k}(z))$ and $\Phi(\boldsymbol{T}_{\leq n-k-1}(z))$ respectively, centered at $\boldsymbol{T}(z)$. We observe that both series terminate at the second term.

$$\begin{aligned}
\boldsymbol{G}_{\leq n-k}(z) &= \Phi(\boldsymbol{T}_{\leq n-k}(z)) = \sum_{i\geq 0} \frac{\Phi^{(i)}(\boldsymbol{T}(z))}{i!}(\boldsymbol{T}_{\leq n-k}(z) - \boldsymbol{T}(z))^i, \\
\boldsymbol{G}_{\leq n-k-1}(z) &= \Phi(\boldsymbol{T}_{\leq n-k-1}(z)) = \sum_{i\geq 0} \frac{\Phi^{(i)}(\boldsymbol{T}(z))}{i!}(\boldsymbol{T}_{\leq n-k-1}(z) - \boldsymbol{T}(z))^i
\end{aligned} \tag{3.21}$$

For $k \leq \frac{n}{2} - 1$, the lowest power in the difference

$$\boldsymbol{T}_{\leq n-k}(z) - \boldsymbol{T}(z) = -\sum_{j>n-k} t_j z^j.$$

is $\frac{n}{2} + 1$. When it is raised to a power $i \geq 2$ the degree of the series is necessarily $\geq n$. As such, for $i \geq 2$,

$$[z^n](\boldsymbol{T}_{\leq n-k}(z) - \boldsymbol{T}(z))^i = 0.$$

The same argument holds for $[z^n](\boldsymbol{T}_{\leq n-k}(z) - \boldsymbol{T}(z))^i$ since the lowest power in the sum is $\frac{n}{2}$. As a result,

$$[z^n]\boldsymbol{G}_{\leq n-k}(z) = [z^n]\Big(\Phi(\boldsymbol{T}(z)) + \Phi'(\boldsymbol{T}(z))(\boldsymbol{T}_{\leq n-k}(z) - \boldsymbol{T}(z))\Big)$$
$$[z^n]\boldsymbol{G}_{\leq n-k-1}(z) = [z^n]\Big(\Phi(\boldsymbol{T}(z)) + \Phi'(\boldsymbol{T}(z))(\boldsymbol{T}_{\leq n-k-1}(z) - \boldsymbol{T}(z))\Big)$$

Plugging into Eq. 3.20 and simplifying the difference yields the desired form.

$$\begin{aligned}\frac{[z^n](\boldsymbol{G}_{\leq n-k}(z) - \boldsymbol{G}_{\leq n-k-1}(z))}{[z^n]\boldsymbol{G}_\tau(z)} &= \frac{[z^n]\Phi'(\boldsymbol{T}(z))(\boldsymbol{T}_{\leq n-k}(z) - \boldsymbol{T}_{\leq n-k-1}(z))}{[z^n]\boldsymbol{G}_\tau(z)} \\ &= \frac{[z^n]\Phi'(\boldsymbol{T}(z))t_{n-k}z^{n-k}}{[z^n]\boldsymbol{G}_\tau(z)}\end{aligned}$$

The product $z^{n-k}\Phi'(\boldsymbol{T}(z))$ shifts the coefficients of $\Phi'(\boldsymbol{T}(z))$ $n-k$ positions to the left from the coefficient of z^i to the coefficient of $i+n-k$. Thus the coefficient of z^n in $\Phi'(\boldsymbol{T}(z))z^{n-k}$ is the coefficient of z^k in $\Phi'(\boldsymbol{T}(z))$. Therefore,

$$\frac{[z^n]\Phi'(\boldsymbol{T}(z))t_{n-k}z^{n-k}}{[z^n]\boldsymbol{G}_\tau(z)} = \frac{[z^k]\Phi'(\boldsymbol{T}(z))[z^{n-k}]\boldsymbol{T}(z)}{[z^n]\boldsymbol{G}_\tau(z)}$$

This completes the proof. □

Remark. Although the 2 equations given by 3.21 are presented as Taylor series of a composition, they can also be realized combinatorially as counting the same objects. By definition, $\boldsymbol{G}_{\leq n-k}(z)$ counts the number of γ-structures with block length at most $n-k$. To see that the RHS counts these structures as well, consider $\frac{\Phi^{(i)}(\boldsymbol{T}(z))}{i!}$ as counting the way to mark the $(i+1)$ positions (unordered) where a block of length greater than $n-k$ can be inserted. $\boldsymbol{T}_{\leq n-k}(z) - \boldsymbol{T}(z)$ counts the possible structures that do not meet the maximum block length criteria.

As such, the i-th term in the series counts the number of ways to insert i blocks of length greater than $n-k$ into a sequence of block followed by the alternating coefficient $(-1)^i$. The term corresponding to $i=0$ is exactly $\boldsymbol{G}_\tau(z)$. Thus the full series is realized as the inclusion/exclusion principle applied to removing the set of structures containing at least one block of length greater than $n-k$ from the set of all possible structures.

Furthermore, The choice to restrict consideration of $\mathbb{P}(\mathbb{B}_n = n-k)$ to cases $k \leq \frac{n}{2} - 1$ stems from the advantageous nature of the sums in 3.21; the choice of k affects the number of nonzero coefficients in the sum. An expansion with nonzero terms from only $\boldsymbol{T}(z)$ and $\Phi'(\boldsymbol{T}(z))$ allows us to limit the amount of computation necessary later down the road. Given the coefficient asymptotics of $\boldsymbol{T}(z)$ and $\Phi'(\boldsymbol{T}(z))$, we are afforded an asymptotic approximation of $\mathbb{P}(\mathbb{B}_n = n-k)$.

Lemma 17. *The singular expansions for $\boldsymbol{T}(z)$ and $\Phi'(\boldsymbol{T}(z))$ are the following*

$$\boldsymbol{T}(z) = 1 - \frac{1}{\theta_0} + \frac{\theta_1}{\theta_0^2}(\mu - z)^{\frac{1}{2}} + \theta_3(\mu - z) + O((\mu - z)^{\frac{3}{2}}), \tag{3.22}$$

$$\Phi'(\boldsymbol{T}(z)) = \theta_0^2 + 2\theta_0\theta_1(\mu - z)^{\frac{1}{2}} + \theta_4(\mu - z) + O((\mu - z)^{\frac{3}{2}}), \tag{3.23}$$

as $z \to \mu$, where μ is the dominant singularity of $\boldsymbol{G}_\tau(z)$, $\theta_0 = \boldsymbol{G}_\tau(\mu)$, and all other θ_is are positive constants.

In addition, the asymptotics of the coefficients of $\boldsymbol{T}(z)$ and $\Phi'(\boldsymbol{T}(z))$ are given by

$$[z^n]\boldsymbol{T}(z) = \frac{c}{\theta_0^2} n^{-\frac{3}{2}} \mu^{-n}(1 + O(n^{-1}))$$
$$[z^n]\Phi'(\boldsymbol{T}(z)) = 2\theta_0 c n^{-\frac{3}{2}} \mu^{-n}(1 + O(n^{-1})).$$

where $c = -\theta_1 \mu^{\frac{1}{2}} \Gamma(-\frac{1}{2})^{-1}$.

Proof. By construction,

$$\boldsymbol{T}(z) = \frac{\boldsymbol{G}_\tau(z) - 1}{\boldsymbol{G}_\tau(z)}.$$

One can view $\boldsymbol{T}(z)$ as the composition $(\phi \circ \boldsymbol{G}_\tau)(z)$ where $\phi = 1 - 1/z$. Then the candidates for the dominant singularity of $\boldsymbol{T}(z)$ are the dominant singularity of $\boldsymbol{G}_\tau(z)$ and the solution to $\boldsymbol{G}_\tau(z) = 0$. Since $\boldsymbol{G}_\tau(0) = 1$ and $\boldsymbol{G}_\tau(z)$ restricted to the positive real line is an increasing function, $\boldsymbol{G}_\tau(z) > 0$. Thus the dominant singularity of $\boldsymbol{T}(z)$ is the dominant singularity of $\boldsymbol{G}_\tau(z)$, μ. Since $\phi(z)$ is analytic near $\boldsymbol{G}_\tau(\mu)$ This implies that the singular expansion of $\boldsymbol{T}(z)$ is the Taylor expansion of $\phi(x)$ at θ_0 composed with the singular expansion of $\boldsymbol{G}_\tau(z)$. Namely,

$$\begin{aligned} 1 - \frac{1}{x} &= 1 - \frac{1}{\theta_0} + \frac{x - \theta_0}{\theta_0^2} + O((x - \theta_0)^2), \\ 1 - \frac{1}{\boldsymbol{G}_\tau(z)} &= 1 - \frac{1}{\theta_0} + \frac{1}{\theta_0^2}(\theta_1(\mu - z)^{\frac{1}{2}} + \theta_2(\mu - z)) + O((\mu - z)^{\frac{3}{2}}) \\ &= 1 - \frac{1}{\theta_0} + \frac{\theta_1}{\theta_0^2}(\mu - z)^{\frac{1}{2}} + \theta_3(\mu - z) + O((\mu - z)^{\frac{3}{2}}). \end{aligned}$$

It follows immediately from Theorem 2 that the transfer to coefficients is

$$[z^n]\boldsymbol{T}(z) = \frac{c}{\theta_0^2} n^{-\frac{3}{2}} \mu^{-n}(1 + O(n^{-1})).$$

Now consider $\Phi'(\boldsymbol{T}(z)) = \boldsymbol{G}_\tau(z)^2$. Clearly the singularity of $\Phi'(\boldsymbol{T}(z))$ is μ since x^2 is entire. The singular expansion of $\boldsymbol{G}_\tau(z)^2$ is the singular expansion of $\boldsymbol{G}_\tau(z)$ squared. The result follows. □

We are now in a position to compute $\mathbb{E}[\mathbb{B}_n]$. The following manipulation results in a form that allows us to make use of the coefficient asymptotics previously computed.

$$\begin{aligned}\mathbb{E}[\mathbb{B}_n] &= \sum_{i=1}^{n}(n-k)\mathbb{P}(\mathbb{B}_n = n-k)\\ &= \sum_{k=1}^{n} n\mathbb{P}(\mathbb{B}_n = n-k) - \sum_{k=1}^{n} k\mathbb{P}(\mathbb{B}_n = n-k)\\ &= n - \sum_{k=1}^{\frac{n}{2}-1} k\mathbb{P}(\mathbb{B}_n = n-k) - \sum_{k=\frac{n}{2}}^{n} k\mathbb{P}(\mathbb{B}_n = n-k) \qquad (3.24)\end{aligned}$$

The final lines of this manipulation take enough trickery they are afforded the lemmas that follow. Theorem 20 states the asymptotics of $\mathbb{E}[\mathbb{B}_n]$ as well as $\mathbb{V}[\mathbb{B}_n]$.

Lemma 18. *For $n \to \infty$,*

$$\sum_{k=1}^{\frac{n}{2}-1} k\mathbb{P}(\mathbb{B}_n = n-k) = \alpha n^{\frac{1}{2}}(1+o(1)) \qquad (3.25)$$

where $\alpha = \frac{4c}{\theta_0}$.

Proof. Since $k \leq \frac{n}{2} - 1$, Lemma 16 applies to $\mathbb{P}(\mathbb{B}_n = n-k)$ in Eq. 3.25. Furthermore, the coefficient asymptotics from Lemmas 15 and 17 bring us the desired approximation.

$$\begin{aligned}\mathbb{P}(\mathbb{B}_n = n-k) &= \frac{[z^k]\Phi'(\boldsymbol{T}(z))[z^{n-k}]\boldsymbol{T}(z)}{[z^n]\boldsymbol{G}_\tau(z)}\\ &= \frac{(2\theta_0 c k^{-\frac{3}{2}}\mu^{-k})(c(n-k)^{-\frac{3}{2}}\mu^{-(n-k)})(1+O(k^{-1}))(1+O((n-k)^{-1}))}{\theta_0^2 c n^{-\frac{3}{2}}\mu^{-n}(1+O(n^{-1}))}\\ &= \frac{2c}{\theta_0}\Big(1-\frac{k}{n}\Big)^{-\frac{3}{2}}k^{-\frac{3}{2}}(1+O(k^{-1}))(1+O(n^{-1})), \quad n,k\to\infty.\end{aligned}$$

The third equality holds because

$$\frac{(1+O((n-k)^{-1}))}{(1+O(n^{-1}))} = \frac{(1+O(n^{-1}))}{(1+O(n^{-1}))} = (1+O(n^{-1})).$$

In order for the asymptotics of $[z^k]\Phi'(\boldsymbol{T}(z))$ to hold, k must approach infinity. For $1 \leq k \leq \frac{n}{2}-1$, this is not guaranteed. However, it is when k is bounded below by n as well. To reconcile this, consider the sum in two pieces.

$$\sum_{1\leq k\leq \frac{n}{2}-1} k\mathbb{P}(\mathbb{B}_n = n-k) = \sum_{1\leq k< n^{\frac{1}{8}}} k\mathbb{P}(\mathbb{B}_n = n-k) + \sum_{n^{\frac{1}{8}}\leq k\leq \frac{n}{2}-1} k\mathbb{P}(\mathbb{B}_n = n-k).$$

For $1 \leq k < n^{\frac{1}{8}}$, an approximation of the sum is derived without employing the coefficient approximations. Since the probability function is bounded by 1,

$$\sum_k k\mathbb{P}(\mathbb{B}_n = n-k) \leq \sum_k k = \frac{n^{\frac{1}{8}}(n^{\frac{1}{8}}+1)}{2} = o(n^{\frac{1}{2}}).$$

Note that the sum could have been separated at any place such that the final inequality resulted in a term of order strictly less than $1/2$.

Now consider $n^{\frac{1}{8}} \leq k \leq \frac{n}{2} - 1$. Since k is bounded above and below by powers of n, we can take both k and n toward infinity and substitute the asymptotics of $\mathbb{P}(\mathbb{B}_n = n-k)$ into the second sum. For $n, k \to \infty$,

$$\begin{aligned}\sum_k k\mathbb{P}(\mathbb{B}_n = n-k) &= \frac{2c}{\theta_0}\sum_k k\Big(1-\frac{k}{n}\Big)^{-\frac{3}{2}}k^{-\frac{3}{2}}(1+O(k^{-1}))(1+O(n^{-1})) \\ &= \frac{2c}{\theta_0}n^{\frac{1}{2}}\sum_k \frac{1}{n}\Big(1-\frac{k}{n}\Big)^{-\frac{3}{2}}\Big(\frac{k}{n}\Big)^{-\frac{1}{2}}(1+O(k^{-1}))(1+O(n^{-1})). \qquad (3.26)\end{aligned}$$

The sum in final line of the equality can be realized as a Riemann sum which converges to a Riemann integral. Recall the formula,

$$\int_a^b f(x)\,dx = \lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} f(a+\frac{b-a}{n}k).$$

For $a = 0$, $b = \frac{1}{2}$, $f(x) = (1-x)^{-\frac{3}{2}}x^{-\frac{1}{2}}$, the terms of the sum in the expression of $\mathbb{P}(\mathbb{B}_n = n-k)$ match that of the corresponding Riemann sum. The difference of the Riemann sum and the sum in line 3.26 is

$$\begin{aligned}\sum_{k=1}^{n^{\frac{1}{8}}}\frac{1}{n}\Big(1-\frac{k}{n}\Big)^{-\frac{3}{2}}\Big(\frac{k}{n}\Big)^{-\frac{1}{2}} + \frac{4}{n} &\leq \sum_{k=1}^{n^{\frac{1}{8}}}(1-n^{-\frac{7}{8}})^{-\frac{3}{2}}n^{-\frac{1}{2}} + \frac{4}{n} \\ &= n^{\frac{1}{8}}(1-n^{-\frac{7}{8}})^{-\frac{3}{2}}n^{-\frac{1}{2}} + \frac{4}{n} \\ &= o(1)\end{aligned}$$

It follows that

$$\sum_{n^{\frac{1}{8}}\leq k\leq \frac{n}{2}-1}\frac{1}{n}\Big(1-\frac{k}{n}\Big)^{-\frac{3}{2}}\Big(\frac{k}{n}\Big)^{-\frac{1}{2}} = \int_0^{\frac{1}{2}}(1-x)^{-\frac{3}{2}}x^{-\frac{1}{2}}\,dx(1+o(1)) = 2(1+o(1))$$

We can eliminate dependence on k in the sum in line 3.26 by bounding $(1 + O(k^{-1}))$ above and below by $(1 + o(1))$. Substituting the above asymptotics into the original sum,

$$\begin{aligned}\sum_k \mathbb{P}(\mathbb{B}_n = n-k) &= \frac{2c}{\theta_0} n^{\frac{1}{2}} \sum_k \frac{1}{n}\Big(1-\frac{k}{n}\Big)^{-\frac{3}{2}}\Big(\frac{k}{n}\Big)^{-\frac{1}{2}}(1+O(k^{-1}))(1+O(n^{-1})) \\ &= \frac{4c}{\theta_0} n^{\frac{1}{2}}(1+o(1))(1+O(n^{-1})) \\ &= \frac{4c}{\theta_0} n^{\frac{1}{2}}(1+o(1))\end{aligned}$$

Setting $\alpha = \dfrac{4c}{\theta_0}$ yields the desired result. □

Lemma 19. $\displaystyle\sum_{\frac{n}{2}-1<k\le n} k\mathbb{P}(\mathbb{B}_n = n-k) = o(n^{\frac{1}{2}})$ *as* $n \to \infty$.

Proof. To invoke Lemma 16 once again, we make use of basic properties of probability functions. Namely,

$$\sum_{\frac{n}{2}-1<k\le n} \mathbb{P}(\mathbb{B}_n = n-k) = 1 - \sum_{1\le k\le \frac{n}{2}-1} \mathbb{P}(\mathbb{B}_n = n-k).$$

Again the sum is broken up to yield advantageous asymptotics.

$$\begin{aligned}&\sum_{\frac{n}{2}-1<k\le n} k\mathbb{P}(\mathbb{B}_n = n-k) \\ &\qquad \le n\Big(1 - \sum_{1\le k\le \frac{n}{2}-1} \mathbb{P}(\mathbb{B}_n = n-k)\Big) \\ &\qquad = n\Big(1 - \sum_{1\le k< n^{\frac{2}{5}}} \mathbb{P}(\mathbb{B}_n = n-k) - \sum_{n^{\frac{2}{5}}\le k\le \frac{n}{2}-1} \mathbb{P}(\mathbb{B}_n = n-k)\Big).\end{aligned}$$

For $1 \le k < n^{\frac{2}{5}}$, the coefficient asymptotics of $\boldsymbol{T}(z)$ and $\boldsymbol{G}_\tau(z)$ can be substituted into $\mathbb{P}(\mathbb{B}_n = n-k)$. The resulting form is

$$\begin{aligned}\mathbb{P}(\mathbb{B}_n = n-k) &= \frac{[x^k]\Phi'(\boldsymbol{T}(z))[x^{n-k}]\boldsymbol{T}(z)}{[z^n]\boldsymbol{G}_\tau(z)} \\ &= \frac{[x^k]\Phi'(\boldsymbol{T}(z))\mu^k}{\theta_0^2}\left(1-\frac{k}{n}\right)^{-\frac{3}{2}}(1+O(n^{-1})) \\ &= \frac{[x^k]\Phi'(\boldsymbol{T}(z))\mu^k}{\theta_0^2}(1+O(n^{-\frac{3}{5}}))(1+O(n^{-1})) \\ &= \frac{[x^k]\Phi'(\boldsymbol{T}(z))\mu^k}{\theta_0^2}(1+o(n^{-\frac{1}{2}})) \qquad (3.27)\end{aligned}$$

The third inequality follows from applying the upper bound on k,

$$\left(1-\frac{k}{n}\right)^{-\frac{3}{2}} < (1-n^{-\frac{3}{5}})^{-\frac{3}{2}} = (1+O(n^{-\frac{3}{5}})).$$

The choice of approximation $(1+o(n^{-\frac{1}{2}}))$ becomes useful when the asymptotics are combined in the final line of the proof.

To simplify notation, let $b_k = [z^k]\Phi'(\boldsymbol{T}(z))$.

$$\begin{aligned}\sum_{1\le k<n^{\frac{2}{5}}} \mathbb{P}(\mathbb{B}_n = n-k) &= \sum_{1\le k<n^{\frac{2}{5}}} \frac{b_k\mu^k}{\theta_0^2}(1+o(n^{-\frac{1}{2}})) \\ &= \Big(\frac{1}{\theta_0^2}\sum_{k\ge 1} b_k\mu^k - \frac{1}{\theta_0^2}\sum_{k\ge n^{\frac{2}{5}}} b_k\mu^k\Big)(1+o(n^{-\frac{1}{2}})) \\ &= \Big(1-\frac{1}{\theta_0^2}\sum_{k\ge n^{\frac{2}{5}}} b_k\mu^k\Big)(1+o(n^{-\frac{1}{2}}))\end{aligned}$$

Recall that $\theta_0^2 = \Phi'(\boldsymbol{T}(\mu))$. The first sum in the final equality is the generating function $\Phi'(\boldsymbol{T}(z))$. Evaluating $\Phi'(\boldsymbol{T}(z))$ at μ yields $\Phi'(\boldsymbol{T}(\mu)) = (\theta_0^2)^{-1}$. Thus

$$\frac{1}{\theta_0^2}\sum_{k\ge 1} b_k\mu^k = 1.$$

Since k is now bounded below by a power of n in the final form of the original sum, we are in a position to apply the asymptotics of $[z^k]\Phi'(\boldsymbol{T}(z))$.

$$\begin{aligned}&\Big(1-\frac{1}{\theta_0^2}\sum_{k\ge n^{\frac{2}{5}}} b_k\mu^k\Big)(1+o(n^{-\frac{1}{2}})) \\ &\quad= \Big(1-\frac{2c}{\theta_0}\sum_{k\ge n^{\frac{2}{5}}} k^{-\frac{3}{2}}(1+O(k^{-1}))\Big)(1+o(n^{-\frac{1}{2}})) \\ &\quad= \Big(1-\frac{\alpha}{2}\sum_{k\ge n^{\frac{2}{5}}} k^{-\frac{3}{2}} + O\big(\sum_{k\ge n^{\frac{2}{5}}} k^{-\frac{5}{2}}\big)\Big)(1+o(n^{-\frac{1}{2}}))\end{aligned}$$

In order to eliminate dependence on k, we apply the well known asymptotics of the Hurwitz-Zeta function defined as

$$\zeta(s,n) = \sum_{i\ge 0}(n+i)^{-s},$$

with the expansion $\zeta(s,n) = \frac{n^{1-s}}{s-1}(1+O(n^{-1}))$.

$$\sum_{k\geq n^{\frac{2}{5}}} k^{-\frac{3}{2}} = \sum_{k\geq 0}(k+n^{\frac{2}{5}})^{-\frac{3}{2}} = 2n^{-\frac{1}{5}}(1+O(n^{-\frac{2}{5}})),$$
$$\sum_{k\geq n^{\frac{2}{5}}} k^{-\frac{5}{2}} = \sum_{k\geq 0}(k+n^{\frac{2}{5}})^{-\frac{5}{2}} = \frac{2}{3}n^{-\frac{3}{5}}(1+O(n^{-\frac{2}{5}})).$$

Substituting the asymptotics into the final expression,

$$\begin{aligned}
&\sum_{1\leq k<n^{\frac{2}{5}}} \mathbb{P}(\mathbb{B}_n = n-k) \\
&\qquad = \Big(1-\frac{\alpha}{2}\sum_{k\geq n^{\frac{2}{5}}} k^{-\frac{3}{2}} + O\big(\sum_{k\geq n^{\frac{2}{5}}} k^{-\frac{5}{2}}\big)\Big)(1+o(n^{-\frac{1}{2}})) \\
&\qquad = \Big(1-\alpha n^{-\frac{1}{5}}(1+O(n^{-\frac{2}{5}})) + O(n^{-\frac{3}{5}})\Big)(1+o(n^{-\frac{1}{2}})) \\
&\qquad = 1-\alpha n^{-\frac{1}{5}} + o(n^{-\frac{1}{2}}).
\end{aligned}$$

Now consider $n^{\frac{2}{5}} \leq k \leq \frac{n}{2}-1$. Since k is bounded below by n, we can let k, $n\to\infty$ and substitute the coefficient asymptotics of $\mathbb{P}(\mathbb{B}_n = n-k)$,

$$\begin{aligned}
&\sum_{n^{\frac{2}{5}}\leq k\leq \frac{n}{2}-1} \mathbb{P}(\mathbb{B}_n = n-k) \\
&\qquad = \frac{2c}{\theta_0}\sum_{n^{\frac{2}{5}}\leq k\leq \frac{n}{2}-1} \Big(1-\frac{k}{n}\Big)^{-\frac{3}{2}} k^{-\frac{3}{2}}(1+O(k^{-1}))(1+O(n^{-1})).
\end{aligned}$$

Recall in Lemma 18 that a sum similar to the one above was realized as an Riemann integral. However, the sum above is lacking an additional term k. Without this term the associated integral does not converge. As such, we instead approximate the sum using the Euler-Maclaurin summation formula given below.

$$\sum_{a\leq k<b} f(k) = \int_a^b f(x)\,dx + B_1 f(x)\Big|_a^b + \frac{B_2}{2}f'(x)\Big|_a^b + R_2,$$

where B_i's are the Bernoulli numbers and R_2 is on the order $O\big(\int_a^b |f''(x)|\,dx\big)$. For more detail on the derivation of this formula, see Ch. 9 in [9].

Let $a = n^{\frac{2}{5}}$, $b = \frac{n}{2}$, $f(k) = \left(k - \frac{k^2}{n}\right)^{-\frac{3}{2}}$. Then

$$\begin{aligned}
&\sum_{n^{\frac{2}{5}} \le k < \frac{n}{2}} \left(k - \frac{k^2}{n}\right)^{-\frac{3}{2}} \\
&\qquad = \int_{n^{\frac{2}{5}}}^{\frac{n}{2}} \left(x - \frac{x^2}{n}\right)^{-\frac{3}{2}} dx + \frac{1}{2}\left(x - \frac{x^2}{n}\right)^{-\frac{3}{2}}\Big|_{n^{\frac{2}{5}}}^{\frac{n}{2}} + \frac{1}{12} f'(x)\Big|_{n^{\frac{2}{5}}}^{\frac{n}{2}} + O(n^{-1}) \\
&\qquad = 2n^{-\frac{1}{5}}(1 + O(n^{-\frac{3}{5}})) + O(n^{-\frac{3}{5}}) + O(n^{-1}) \\
&\qquad = 2n^{-\frac{1}{5}}(1 + O(n^{-\frac{2}{5}}))
\end{aligned}$$

Approximating each term in the second equality gives us the third equality. The choice of asymptotic approximation is based on later computation. It will become clear that the choice of power of n in the final inequality yields the desired result. As for the error term, we substitute the appropriate upper bounds on k to get

$$\begin{aligned}
\sum_{n^{\frac{2}{5}} \le k \le \frac{n}{2} - 1} \left(k - \frac{k^2}{n}\right)^{-\frac{3}{2}} O(k^{-1}) &\le \sum_{n^{\frac{2}{5}} \le k \le \frac{n}{2} - 1} (n^{\frac{2}{5}} - 1)^{-\frac{3}{2}} O(n^{-\frac{2}{5}}) \\
&= \sum_{n^{\frac{2}{5}} \le k \le \frac{n}{2} - 1} O(n^{-1}) \\
&= (\frac{n}{2} - 1 - n^{\frac{2}{5}}) O(n^{-1}) \\
&= O(n^{-\frac{3}{5}}).
\end{aligned}$$

Combining the above computation,

$$\begin{aligned}
&\sum_{n^{\frac{2}{5}} \le k \le \frac{n}{2} - 1} \mathbb{P}(\mathbb{B}_n = n - k) \\
&\qquad = \frac{2c}{\theta_0} \sum_{n^{\frac{2}{5}} \le k \le \frac{n}{2} - 1} \left(1 - \frac{k}{n}\right)^{-\frac{3}{2}} k^{-\frac{3}{2}} (1 + O(k^{-1}))(1 + O(n^{-1})) \\
&\qquad = \frac{\alpha}{2}\left(2n^{-\frac{1}{5}}(1 + O(n^{-\frac{2}{5}})) + O(n^{-\frac{3}{5}})\right)(1 + O(n^{-1})) \\
&\qquad = \alpha n^{-\frac{1}{5}} + o(n^{-\frac{1}{2}}).
\end{aligned}$$

We are now in a position to plug all of the above asymptotics into the initial sum.

$$\begin{aligned}
&\sum_{\frac{n}{2}-1<k\leq n} k\mathbb{P}(\mathbb{B}_n = n-k) \\
&\qquad \leq n\big(1 - \sum_{1\leq k<n^{\frac{2}{5}}} \mathbb{P}(\mathbb{B}_n = n-k) - \sum_{n^{\frac{2}{5}}\leq k\leq \frac{n}{2}-1} \mathbb{P}(\mathbb{B}_n = n-k)\big) \\
&\qquad = n\big(1 - (1-\alpha n^{-\frac{1}{5}} + o(n^{-\frac{1}{2}})) - (\alpha n^{-\frac{1}{5}} + o(n^{-\frac{1}{2}}))\\
&\qquad = n(o(n^{-\frac{1}{2}}) - o(n^{-\frac{1}{2}}))\big) \\
&\qquad = o(n^{\frac{1}{2}}).
\end{aligned}$$

□

Theorem 20. *The mean and variance of the random variable $\mathbb{X}_n$ are the following.*

$$\mathbb{E}[\mathbb{B}_n] = n - \alpha n^{\frac{1}{2}}(1+o(1)) \tag{3.28}$$

$$\mathbb{V}[\mathbb{B}_n] = \beta n^{\frac{3}{2}}(1+o(1)) \tag{3.29}$$

where $\alpha = \frac{4c}{\theta_0}$, $\beta = (1-\frac{\pi}{4})\alpha$.

Proof. Consider the expectation $\mathbb{E}[\mathbb{B}_n]$ of the length of the longest block as n grows large.

$$\begin{aligned}
\mathbb{E}[\mathbb{B}_n] &= \sum_{k=1}^{n}(n-k)\mathbb{P}(\mathbb{B}_n = n-k) \\
&= n - \sum_{k=1}^{n} k\mathbb{P}(\mathbb{B}_n = n-k) \\
&= n - \sum_{k=1}^{\frac{n}{2}-1} k\mathbb{P}(\mathbb{B}_n = n-k) - \sum_{k=\frac{n}{2}}^{n} k\mathbb{P}(\mathbb{B}_n = n-k) \\
&= n - \alpha n^{\frac{1}{2}}(1+o(1)) - o(n^{\frac{1}{2}}) && \text{(Lemmas 18, 19)} \\
&= n - \alpha n^{\frac{1}{2}}(1+o(1))
\end{aligned}$$

Now consider the variance $\mathbb{V}[\mathbb{B}_n]$ of the length of the longest block as n grows large.

$$\begin{aligned}
\mathbb{V}[\mathbb{B}_n] &= \sum_{k=1}^{n}(n-k)^2\mathbb{P}(\mathbb{B}_n = n-k) - \mathbb{E}[\mathbb{B}_n]^2 \\
&= n^2 - 2n\sum_{k=1}^{n} k\mathbb{P}(\mathbb{B}_n = n-k) + \sum_{k=1}^{n} k^2\mathbb{P}(\mathbb{B}_n = n-k) - \mathbb{E}[\mathbb{B}_n]^2 \\
&= \sum_{k=1}^{n} k^2\mathbb{P}(\mathbb{B}_n = n-k) + n^2 - 2n\big(an^{\frac{1}{2}}(1+o(1))\big) - \big(n - an^{\frac{1}{2}}(1+o(1))\big)^2 \\
&= \sum_{k=1}^{n} k^2\mathbb{P}(\mathbb{B}_n = n-k) - o(n^{\frac{3}{2}}) \\
&= \beta n^{\frac{3}{2}}(1+o(1))
\end{aligned}$$

where $\beta = (1-\frac{\pi}{4})\alpha$. The last equality follows from the below claims that parallel the computation of expectation in Lemmas 18 and 19.

Claim 1:

$$\sum_{1\le k\le n^{\frac{1}{8}}} k^2\mathbb{P}(\mathbb{B}_n = n-k) = o(n^{\frac{1}{2}}).$$

Proof. Claim 1 follows from bounding $\mathbb{P}(\mathbb{B}_n = n-k)$ by 1 and employing the well known formula for a sum of squares.

$$\sum_{k=1}^{n^{\frac{1}{8}}} k^2\mathbb{P}(\mathbb{B}_n = n-k) \le \sum_{k=1}^{n^{\frac{1}{8}}} k^2 = \frac{n^{\frac{1}{8}}(n^{\frac{1}{8}}+1)(2n^{\frac{1}{8}}+1)}{6} = o(n^{\frac{1}{2}}).$$

Claim 2:

$$\sum_{n^{\frac{1}{8}}\le k\le \frac{n}{2}-1} k^2\mathbb{P}(\mathbb{B}_n = n-k) = \beta n^{\frac{3}{2}}(1+o(1)).$$

Proof. Consider $n^{\frac{1}{8}} \le k \le \frac{n}{2} - 1$. Since k is bounded below by n, letting k and n tend to

infinity allows us to substitute the asymptotics for $\mathbb{P}(\mathbb{B}_n = n-k)$ into the desired sum,

$$\begin{aligned}
&\sum_k k^2 \mathbb{P}(\mathbb{B}_n = n-k) \\
&\quad = \frac{2c}{\theta_0} \sum_k k^2 \Big(1-\frac{k}{n}\Big)^{-\frac{3}{2}} k^{-\frac{3}{2}} (1+O(k^{-1}))(1+O(n^{-1})) \\
&\quad = \frac{\alpha}{2} n^{\frac{3}{2}} \sum_k \frac{1}{n} \Big(1-\frac{k}{n}\Big)^{-\frac{3}{2}} \Big(\frac{k}{n}\Big)^{\frac{1}{2}} (1+O(k^{-1}))(1+O(n^{-1})) \\
&\quad = \frac{\alpha}{2} \Big(2-\frac{\pi}{2}\Big) n^{\frac{3}{2}} (1+o(1)).
\end{aligned}$$

The final equality again follows from the previous sum realized as the tail end of a Riemann sum. For $a = 0$, $b = \frac{1}{2}$, $f(x) = (1-x)^{-\frac{3}{2}} x^{\frac{1}{2}}$, the terms of the sum in the expression of $\mathbb{P}(\mathbb{B}_n = n-k)$ match that of the corresponding Riemann sum. The difference of the Riemann sum and the sum in $\mathbb{P}(\mathbb{B}_n = n-k)$ is

$$\begin{aligned}
\sum_{k=1}^{n^{\frac{1}{8}}} \frac{1}{n} \Big(1-\frac{k}{n}\Big)^{-\frac{3}{2}} \Big(\frac{k}{n}\Big)^{\frac{1}{2}} + 2n^{-1} &\le \sum_{k=1}^{n^{\frac{1}{8}}} n^{-1} \big(1-n^{-\frac{7}{8}}\big)^{-\frac{3}{2}} \big(n^{-\frac{7}{8}}\big)^{\frac{1}{2}} + 2n^{-1} \\
&= (n^{\frac{7}{8}}-1)^{-\frac{3}{2}} \\
&= n^{-\frac{21}{16}} (1-n^{-\frac{7}{8}})^{-\frac{3}{2}} + 2n^{-1} \\
&= o(1),
\end{aligned}$$

and the Riemann integral evaluates to

$$\int_0^{\frac{1}{2}} (1-x)^{-\frac{3}{2}} x^{\frac{1}{2}} \, dx = 2 - \frac{\pi}{2}.$$

It follows that

$$\sum_{n^{\frac{1}{8}} \le k \le \frac{n}{2}-1} \frac{1}{n} \Big(1-\frac{k}{n}\Big)^{-\frac{3}{2}} \Big(\frac{k}{n}\Big)^{\frac{1}{2}} = \int_0^{\frac{1}{2}} (1-x)^{-\frac{3}{2}} x^{\frac{1}{2}} \, dx (1+o(1)) = (2-\frac{\pi}{2})(1+o(1))$$

For $\beta = (1-\frac{\pi}{2})\alpha$ the result follows.

Claim 3:

$$\sum_{\frac{n}{2}-1 < k \le n} k^2 \mathbb{P}(\mathbb{B}_n = n-k) = o(n^{\frac{3}{2}}).$$

Proof.

$$\begin{aligned}\sum_{\frac{n}{2}-1<k\leq n} k^2\mathbb{P}(\mathbb{B}_n = n-k) &\leq n^2 \sum_{\frac{n}{2}-1<k\leq n} \mathbb{P}(\mathbb{B}_n = n-k) \\ &= n^2\Big(1 - \sum_{1<k\leq n^{\frac{2}{5}}} \mathbb{P}(\mathbb{B}_n = n-k) + \sum_{n^{\frac{2}{5}}<k\leq \frac{n}{2}-1} \mathbb{P}(\mathbb{B}_n = n-k)\Big) \\ &= n^2(1-(1-an^{-\frac{1}{5}}+o(n^{-\frac{1}{2}}))-(an^{-\frac{1}{5}}+o(n^{-\frac{1}{2}}))) \\ &= n^2 o(n^{-\frac{1}{2}}) \\ &= o(n^{\frac{3}{2}})\end{aligned}$$

The second equality is computed in the proof of Lemma 19. □

Remark. This powerful result implies that the distribution of $\mathbb{B}_n$ becomes increasingly more concentrated as n becomes large. Namely, $\mathbb{E}[\mathbb{B}_n]$ is on the order larger than the standard deviation, $\sqrt{\mathbb{V}[\mathbb{B}_n]} = O(n^{\frac{3}{4}})$. Because $\mathbb{E}[\mathbb{B}_n] = n - O(n^{\frac{1}{2}}) > \frac{n}{2}$, we can conclude that there exists a unique longest block. To finalize the discussion on the longest block, Theorem 21 gives the distribution of the difference in the length of the sequence and the length of the longest block.

Theorem 21. *For any $t > \frac{3}{4}$,*

$$\lim_{n\to\infty} \mathbb{P}(n - \mathbb{B}_n \geq \Omega(n^t)) = 0 \tag{3.30}$$

and for any $k = o(n)$

$$\lim_{n\to\infty} \mathbb{P}(n - \mathbb{B}_n = k) = \frac{[z^k]\Phi'(\boldsymbol{T}(z))\mu^k}{\theta_0^2}. \tag{3.31}$$

Proof. We apply Chebyshev's Inequality to $\mathbb{E}[\mathbb{B}_n]$ and $\mathbb{V}[\mathbb{B}_n]$ given by Theorem 20 to arrive at Eq. 3.30. By Chebyshev's Inequality,

$$\mathbb{P}(|\mathbb{B}_n - \mathbb{E}[\mathbb{B}_n]| \geq m) \leq \frac{\mathbb{V}[\mathbb{B}_n]}{m^2}.$$

Set $m = \Omega(n^t)$ for some $t > \frac{3}{4}$. Taking the limit of both sides of the inequality allows us to substitute the asymptotics provided in Theorem 20.

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(|\mathbb{B}_n - \mathbb{E}[\mathbb{B}_n]| \geq m) &\leq \lim_{n\to\infty} \frac{\mathbb{V}[\mathbb{B}_n]}{m^2} \\ \lim_{n\to\infty} \mathbb{P}(n - an^{\frac{1}{2}} - \mathbb{B}_n \geq \Omega(n^t)) &\leq \lim_{n\to\infty} \frac{O(n^{\frac{3}{2}})}{\Omega(n^t)^2} \\ \lim_{n\to\infty} \mathbb{P}(n - \mathbb{B}_n \geq \Omega(n^t) + an^{\frac{1}{2}}) &= 0 \\ \lim_{n\to\infty} \mathbb{P}(n - \mathbb{B}_n \geq \Omega(n^t)) &= 0\end{aligned}$$

The third line follows from $\Omega(n^t)^2 > \Omega(n^{\frac{3}{2}})$. Since probability functions are strictly positive, the inequality becomes an equality.

Now consider $k = o(n)$. Eq. 3.20,

$$\mathbb{P}(\mathbb{B}_n = n - k) = \frac{[z^k]\Phi'(\boldsymbol{T}(z))[z^{n-k}]\boldsymbol{T}(z)}{[z^n]\boldsymbol{G}_\tau(z)}$$

can be applied here since $k \leq \frac{n}{2} - 1$. Eq. 3.31 follows immediately from plugging in the coefficient asymptotics of $\boldsymbol{T}(z)$ and $\boldsymbol{G}_\tau(z)$. □

Remark. Theorem 21 implies that the distribution of $n - \mathbb{B}_n$ a.a.s. converges to a discrete limit law. Eq. 3.30 says that, as n gets large, the probability that the difference between the 5'-3' distance of a γ-structure and the length of its longest block is on the order greater than $n^{\frac{3}{4}}$ is 0.

It is worth noting the special case $k = n^{\frac{1}{2}}$. Eq. 3.31 gives the limiting probability as

$$\lim_{n\to\infty} \mathbb{P}(n - \mathbb{B}_n = k) = \frac{[z^k]\Phi'(\boldsymbol{T}(z))\mu^k}{\theta_0^2}.$$

Since in this case $k \to \infty$, we can apply the coefficient asymptotics of $\Phi'(\boldsymbol{T}(z))$ to the RHS of the equation.

$$\lim_{n\to\infty} \mathbb{P}(n - \mathbb{B}_n = n^{\frac{1}{2}}) = \frac{2n^{-\frac{3}{4}}}{\theta_0}$$

As n gets large, $\mathbb{P}(n - \mathbb{B}_n = n^{\frac{1}{2}})$ approaches 0. Contrast this with the expectation of $n - \mathbb{B}_n$,

$$\mathbb{E}[n - \mathbb{B}_n] = n - \mathbb{E}[\mathbb{B}_n] = \alpha n^{\frac{1}{2}}.$$

At first glance it appears contradictory that the probability that the difference in length of structure and length of longest block achieves the expected value be 0. Refer back to the definition of expectation,

$$\mathbb{E}[\mathbb{B}_n] = \sum_{k\geq 0}(n - k)\mathbb{P}(\mathbb{B}_n = n - k).$$

Although $\mathbb{P}(\mathbb{B}_n = n - k)$ approaches 0 when k is $O(n^{\frac{1}{2}})$, the contribution from $(n - k)$ dominates and inevitably skews the expectation toward infinity. One naturally asks, where on the length spectrum is the highest probability observed? It turns out that with high probability, the tail end of the spectrum is observed. Corollary 21.1 provides information on blocks other than the longest one (these blocks are referred to as short).

Corollary 21.1. *For all $\epsilon > 0$ there exists an integer $t(\epsilon)$ such that*

$$\lim_{n\to\infty} \mathbb{P}(\mathbb{B}_n \geq n - t(\epsilon)) \geq 1 - \epsilon.$$

Proof. Singularity analysis shows that $\phi'(\boldsymbol{T}(z))$ converges at μ. Following from an alternative definition of convergence, for every $\epsilon > 0$ there exists an integer $t(\epsilon)$ such that for $k > t(\epsilon)$,

$$\sum_{k>t(\epsilon)} b_k \mu^k < \epsilon,$$

where $b_k = [z^k]\phi'(\boldsymbol{T}(z))$. Additionally, for $d = \phi'(\boldsymbol{T}(\mu))^{-1}$

$$\sum_{k>t(\epsilon)} d b_k \mu^k < \epsilon.$$

By Theorem 21 the below list of algebraic manipulations hold.

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(\mathbb{B}_n \geq n - t(\epsilon)) &= \lim_{n\to\infty} \sum_{k\leq t(\epsilon)} \mathbb{P}(\mathbb{B}_n = n - k) \\
&= \sum_{k\leq t(\epsilon)} d b_k \mu^k \\
&= \sum_{k\geq 1} d b_k \mu^k - \sum_{k>t(\epsilon)} d b_k \mu^k \\
&= d\phi'(\boldsymbol{T}(\mu)) - \epsilon \\
&= 1 - \epsilon.
\end{aligned}$$

□

The discussion on short rainbows is continued to the next section.

3.4 The Spectrum of Blocks

As previously discussed, the field of analytic combinatorics also provides a framework to study properties of parameters of graphs such as: How many substructures of a particular kind appear in a random graph? In our case, we ask how many blocks of chosen length k appear in uniformly sampled γ-structures of length n. The appearance of a second parameter, the number of blocks of a given length, shifts the nature of the problem slightly. Instead of an ordinary generating function carrying information on one parameter (length), we employ a bivariate generating function, $\boldsymbol{G}_{\tau,k}(z,u)$ carrying information on two parameters. The main result of this section is a statement regarding the limiting distribution of small blocks in γ-structures.

For fixed k, define the bivariate generating function

$$\boldsymbol{G}_{\tau,k}(z,u) = \sum_{n,b} g_{\tau,k}(n,b) z^n u^b$$

where $g_{\tau,k}(n,b)$ is the count of γ-structure of length n containing b blocks of length k.

Define the random variable, $\mathbb{W}_k$, to be the function from a given γ-structure to the count of its blocks of length k. Namely, the set $\mathbb{W}_{k,n} = b$ is the set of γ-structures of length n with b blocks of length k.

The asymptotic behavior of $\mathbb{W}_{k,n}$ is, as before, intrinsically linked to the behavior of $\boldsymbol{G}_{\tau,k}(z,u)$ near its singularities. The study of this behavior parallels nicely the OGF case due to theorems relating to uniform convergence and continuity expounded upon in Section 2. As such, we provide asymptotics of the probability generating function of $\mathbb{W}_{k,n}$ in order to determine its limiting distribution.

Lemma 22. *The bivariate generating function for the number of τ-canonical γ-structures with minimum arc-length λ filtered by rainbows of length k is given by*

$$\boldsymbol{G}_{\tau,k}(z,u) = \frac{1}{1 - r(z,u)} \tag{3.32}$$

where $r(z,u) = \boldsymbol{T}(z) + (u-1)t_k z^k$.

Proof. As before, γ-structures can be constructed by concatenating irreducible block components, represented by the power series

$$\boldsymbol{G}_{\tau}(z) = \frac{1}{1 - \boldsymbol{T}(z)} = \sum_{i \geq 0} (\boldsymbol{T}(z))^i.$$

The coefficient corresponding to the n-th power of $\boldsymbol{G}_{\tau}(z)$ is computed

$$[z^n]\boldsymbol{G}_{\tau}(z) = \sum_{i_1 + \ldots + i_m = n} t_{i_1} \cdots t_{i_m}.$$

The coefficient corresponding to the n-th power of $\sum_{i \geq 0}(r(z,u))^i$ differs by $[z^n]\boldsymbol{G}_{\tau}(z)$ such that when t_k appears in the product, it is accompanied by u. The power of u in the product of coefficients corresponds to the number of appearances of a block of length k, as desired. Thus

$$\boldsymbol{G}_{\tau,k}(z,u) = \sum_{i \geq 0} (r(z,u))^i.$$

It suffices to show that $(1 - r(z,u))$ has an inverse in $\mathbb{Q}[[z]]$ since, by construction, $\boldsymbol{G}_{\tau,k}(z,u)(1 - r(z,u)) = 1$. A power series has an inverse in $\mathbb{Q}[[z]]$ if and only if its constant term is nonzero. Then $1 - r(z,u)$ has an inverse if the constant term of $\boldsymbol{T}(z)$ is different from 1. Since we assume there is no empty block, this is the case.

□

For fixed n and k, the probability of the appearance of a γ-structure of length n with b rainbows of length k is the ratio of the number of such structures to the total number of structures.

$$\mathbb{P}(\mathbb{W}_{k,n} = b) = \frac{g_{\tau,k}(n,b)}{g_\tau(n)} = \frac{[z^n u^b]\boldsymbol{G}_{\tau,k}(z,u)}{[z^n]\boldsymbol{G}_\tau(z)}.$$

The probability generating function for $\mathbb{W}_{k,n}$ is

$$\sum_{b\geq 0} \mathbb{P}(\mathbb{W}_{k,n} = b)u^b = \frac{[z^n u^b]\boldsymbol{G}_{\tau,k}(z,u)}{[z^n]\boldsymbol{G}_\tau(z)} u^b.$$

The coefficient asymptotics of $\boldsymbol{G}_\tau(z)$ are given by

$$[z^n]\boldsymbol{G}_{\tau,k}(z,u) = cn^{-\frac{3}{2}}\mu^{-n}(1+o(1)), \quad n\to\infty,$$

where $c = \theta_1 \mu^{\frac{1}{2}}\Gamma(-\frac{1}{2})^{-1}$.

To derive the asymptotics of $[z^n u^b]\boldsymbol{G}_{\tau,k}(z,u)$, we consider $\boldsymbol{G}_{\tau,k}(z,u)$ as a function of z parameterized by u in order to employ the perturbation analysis given by Flajolet and Sedgewick (refer to Section 2.2). It suffices to prove that $\boldsymbol{G}_{\tau,k}(z,u)$ has a uniform asymptotic expansion with respect to u and to specify how the coefficient, power, and singularity of the asymptotics of $\boldsymbol{G}_\tau(z)$ change by perturbing u.

Lemma 23. *$r(z,u)$ has a uniform asymptotic expansion with respect to $u \in 1 \pm o(1)$.*

Proof. It suffices to show that $(u-1)t_k z^k$ has a uniform asymptotic expansion as $\boldsymbol{T}(z)$ does not depend on u.

The Taylor expansion of $(u-1)t_k z^k$ at $z=\mu$ is

$$(u-1)t_k z^k = (u-1)t_k\big(\mu^k + k\mu^{k-1}(\mu - z) + \ldots + O((\mu-z)^m)\big). \tag{3.33}$$

For given $m<k$, Eq. 3.33 is a uniform expansion if there exists L (independent of u) and a fixed neighborhood $N(\mu)$ such that for all $u\in N_\epsilon(1),\ z\in N(\mu)$,

$$|(u-1)t_k k(k-1)\cdots(k-m)\mu^{k-m}(\mu-z)^m| \leq L|(\mu-z)^m|.$$

Let $L = \epsilon t_k k(k-1)\cdots(k-m)\mu^{k-m}$ where ϵ is the radius of the neighborhood around 1 bounding u. The result follows. □

Theorem 24. *The dominant singularity of $\boldsymbol{G}_{\tau,k}(z,u)$ is the dominant singularity μ of $\boldsymbol{G}_\tau(z)$. The singular expansion for $\boldsymbol{G}_{\tau,k}(z,u)$ is given by*

$$\boldsymbol{G}_{\tau,k}(z,u) = \Phi(r(\mu,u)) + \Phi'(r(\mu,u))\frac{\theta_1}{\theta_0^2}(\mu-z)^{\frac{1}{2}}(1+o(1)) \tag{3.34}$$

and the asymptotics of the coefficients of $\boldsymbol{G}_{\tau,k}(z,u)$ *are given by*

$$[z^n]\boldsymbol{G}_{\tau,k}(z,u) = \frac{\Phi'(r(\mu,u))}{\theta_0^2} c n^{-\frac{3}{2}} \mu^{-n}(1+o(1)) \tag{3.35}$$

where $c = \theta_1 \mu^{\frac{1}{2}} \Gamma(-\frac{1}{2})^{-1}$.

Proof. Consider $\boldsymbol{G}_{\tau,k}(z,u) = \Phi(r(z,u))$ for fixed u such that $u \in 1 \pm o(1)$. The singularities of $\boldsymbol{G}_{\tau,k}(z,u)$ are the singularities of $r(z,u)$ and the solution to $r(z,u) = 1$.

μ, also the dominant singularity of $\boldsymbol{T}(z)$, is the dominant singularity of $r(z,u)$ since $(u-1)t_k\mu^k$ is constant and thus entire. If $r(z,u) < 1$ for $u \in 1 \pm o(1)$, then the dominant singularity of $\boldsymbol{G}_{\tau,k}(z,u)$ is also μ. By inspection, $r(z,u) < 1$ for $u \in 1 \pm o(1)$, $|z| \leq \mu$.

Consequently, $\boldsymbol{G}_{\tau,k}(z,u)$ falls under the subcritical paradigm. As such, the singular expansion of $\boldsymbol{G}_{\tau,k}(z,u)$ is computed as the composition of the Taylor expansion of $\Phi(z)$ at $z = r(\mu,u)$ and the singular expansion of $r(z,u)$ at $z = \mu$.

The singular expansion of $r(z,u)$ at $z = \mu$ is the sum of the singular expansion of $\boldsymbol{T}(z)$ at $z = \mu$ and the Taylor expansion of $(u-1)t_k z^k$ at $z = \mu$.

$$\begin{aligned} r(z,u) &= 1 - \frac{1}{\theta_0} + \frac{\theta_1}{\theta_0^2}(\mu - z)^{\frac{1}{2}}(1+o(1)) + (u-1)t_k\mu^k + O(\mu - z) \\ &= 1 - \frac{1}{\theta_0} + (u-1)t_k\mu^k + \frac{\theta_1}{\theta_0^2}(\mu - z)^{\frac{1}{2}}(1+o(1)) \\ &= r(\mu,u) + \frac{\theta_1}{\theta_0^2}(\mu - z)^{\frac{1}{2}}(1+o(1)) \end{aligned}$$

The Taylor expansion of $\Phi(z)$ at $z = r(\mu,u)$ is

$$\Phi(z) = \Phi(r(\mu,u)) + \Phi'(r(\mu,u))(z - r(\mu,u)) + O((z - r(\mu,u))^2)$$

Combining the two expansions gives us Eq. 3.34.

$$\begin{aligned} \Phi(r(z,u)) &= \Phi(r(\mu,u)) + \Phi'(r(\mu,u))(r(\mu,u) + \frac{\theta_1}{\theta_0^2}(\mu - z)^{\frac{1}{2}}(1+o(1)) - r(\mu,u)) \\ &= \Phi(r(\mu,u)) + \Phi'(r(\mu,u))\frac{\theta_1}{\theta_0^2}(\mu - z)^{\frac{1}{2}}(1+o(1)) \end{aligned}$$

Eq. 3.35 follows immediately from Theorem 2.

□

Theorem 24 provides the asymptotics used in the main result of this section.

Theorem 25. *For fixed k, the limit law of $\mathbb{W}_{k,n}$ is a negative binomial distribution $NB(2,t)$ where*

$$t = \frac{t_k \mu^k}{1 - \boldsymbol{T}(\mu) + t_k \mu^k},$$

and $t_k = [z^k]\boldsymbol{T}(z)$. Consequently, the expectation and variance for $\mathbb{W}_{k,n}$ are given by

$$\mathbb{E}[\mathbb{W}_{k,n}] = \frac{2t_k \mu^k}{1 - \boldsymbol{T}(\mu)}, \quad \mathbb{V}[\mathbb{W}_{k,n}] = \frac{2t}{(1-t)^2}.$$

Proof. To prove that for fixed k, the limit law of $\mathbb{W}_{k,n}$ is a negative binomial distribution $NB(2,t)$ for t defined above, it suffices to show that the probability generating function of the limit distribution can be expressed as

$$p_k(u) = \left(\frac{1-t}{1-tu}\right)^2.$$

Plugging in the asymptotics derived in Theorem 24 to the probability generating function derived at the beginning of this section, we arrive at the desired form.

$$\begin{aligned}
p_k(u) &= \lim_{n\to\infty} \sum_b \mathbb{P}(\mathbb{W}_{k,n} = b) u^b \\
&= \lim_{n\to\infty} \frac{[z^n u^b]\boldsymbol{G}_{\tau,k}(z,u)}{[z^n]\boldsymbol{G}_\tau(z)} u^b \\
&= \lim_{n\to\infty} \frac{[z^n]\boldsymbol{G}_{\tau,k}(z,u)}{[z^n]\boldsymbol{G}_\tau(z)} \\
&= \lim_{n\to\infty} \frac{\frac{\Phi'(r(\mu,u))}{\theta_0^2} c n^{-\frac{3}{2}} \mu^{-n}(1+o(1))}{c n^{-\frac{3}{2}} \mu^{-n}(1+o(1))} \\
&= \frac{\Phi'(r(\mu,u))}{\theta_0^2} \\
&= \left(\frac{1 - \boldsymbol{T}(\mu)}{1 - \boldsymbol{T}(\mu) - (u-1)t_k\mu_k}\right)^2
\end{aligned}$$

The final equality follows from noting $\theta_0^2 = \Phi'(\boldsymbol{T}(\mu))$ and $\Phi'(z) = \frac{1}{(1-z)^2}$. Solving for t in the equation

$$\frac{1 - \boldsymbol{T}(\mu)}{1 - \boldsymbol{T}(\mu) - (u-1)t_k\mu_k} = \frac{1-t}{1-tu}$$

yields the desired expression for t. The mean and variance are immediate form properties of the negative binomial distribution $NB(2,t)$.

□

This completes the analysis of the block length spectrum for γ-structures. We have provided the expectation and variance for the unique longest block, and obtained the limiting distribution of the distribution of blocks of (finite) length k. We put these results in context with other subsets of RNA structures in the discussion.

Chapter 4

Discussion

In Section 3 we prove that the expected length of the longest block for γ-structures is $\mathbb{E}[\mathbb{B}_n] = n - O(n^{\frac{1}{2}})$ with standard deviation $\mathbb{V}[\mathbb{B}_n]^{\frac{1}{2}} = O(n^{\frac{3}{4}})$ by treating the structures first as combinatorial objects as then as objects in a probability space. As such, we are able to derive asymptotic appximations of generating function coefficients carrying information on the number of γ-structures of a given length. These generating function coefficients arise in the definitions of probability and expectation of the length of the longest block. Furthermore, we show in Corollary 21.1 that with high probability, the second longest block has finite length. Namely, for choice of an integer $t(\epsilon)$, the probability that the length of the longest block has length $n - t(\epsilon)$ is close to 1. Thus the second longest rainbow has length bounded by $t(\epsilon)$ with high probability.

Recall the initial question prompting this research: How does the length of the longest rainbow change as the complexity of the structures increases? As mentioned previously, the analysis on the expected block length of a γ-structure for $\gamma = 1$ parallels the analysis of rainbow length for secondary structure in [15] and the results have deep similarities. Lemma 1 in [15] gives the expectation of the length of the longest rainbow as $n - an^{\frac{1}{2}}(1 + o(1))$ and with standard deviation $\sqrt{b}n^{\frac{3}{4}}(1 + o(1))$ where a and b differ from α and β in Theorem 20 of this paper.

What leads on that these results should be so similar? It turns out that secondary structures can be realized as γ-structures such that $\gamma = 0$. Recall that γ-structures are decomposed into blocks with maximal irreducible shadow of genus at most γ. When $\gamma > 0$, these shadows include crossing arcs. However, for $\gamma = 0$ this irreducible shadow is a single rainbow.

In this way secondary structures are decomposed into irreducible components that are identified by a maximal rainbow. Furthermore, an irreducible structure of length m has longest

rainbow of length $m-1$. Thus the expected length of the longest rainbow is the expected length of the longest irreducible component.

Table 4.1 gives the coefficients α_γ in the expectation of the longest block $n - \alpha_\gamma n^{\frac{1}{2}}(1+o(1))$ for γ-structures for varying γ.

Table 4.1: Growth coefficients α_γ for the expectation of the longest block in γ-structures, $0 \leq \gamma \leq 3$

γ	0	1	2	3
α_γ	2.804	1.416	0.964	0.734

Appendix A

Analyticity

The following list of definitions and theorems provides the background necessary to perform singularity analysis. As such, all information can be found throughout [6].

Definition 10. A *region* Ω is an open, connected subset of the complex plane.

Definition 11. A function $f(z)$ defined over a region Ω is *analytic* at a point $z_0 \in \Omega$ if, for z in some open disc centered at z_0 and contained in Ω, it is representable by a convergent power series expansion

$$f(z) = \sum_{i \geq 0} f_i (z - z_0)^i.$$

A function is analytic in a region Ω iff it is analytic at every point of Ω.

Similarly, a converging power series is analytic inside its radius of convergence. Where a function ceases to be analytic, or where a power series fails to converge is an essential part of the analysis of Section 3.

Definition 12. The point $z = a$ is called a *singularity* of the complex function $f(z)$ if f is not analytic at a, but every neighborhood $N_\epsilon(a)$ around a contains at least one point at which $f(z)$ is analytic. A *dominant singularity* is a singularity of smallest modulus, or rather closest to the origin. Note that the dominant singularity need not be unique.

Theorem 26. *A function $f(z)$ analytic at the origin, whose expansion at the origin has a finite radius of convergence R, necessarily has a singularity on the boundary of its disc of convergence $|z| = R$.*

Theorem 26 relates analyticity to the location of a function's singularities in a powerful way. Namely, a converging power series must have at least one singularity on the boundary of its disc of convergence.

Definition 13. Given two numbers ϕ, R with $R > 1$ and $0 < \phi < \frac{\pi}{2}$, the open domain $\Delta(\phi, R)$ is defined as

$$\Delta(\phi, R) = \{z \mid |z| < r,\ z \neq 1,\ |\arg(z-1)| > \phi\}.$$

A domain is a Δ-domain at 1 if it is a $\Delta(\phi, R)$ for some R and ϕ. A function is Δ-analytic if it is analytic in some Δ-domain.

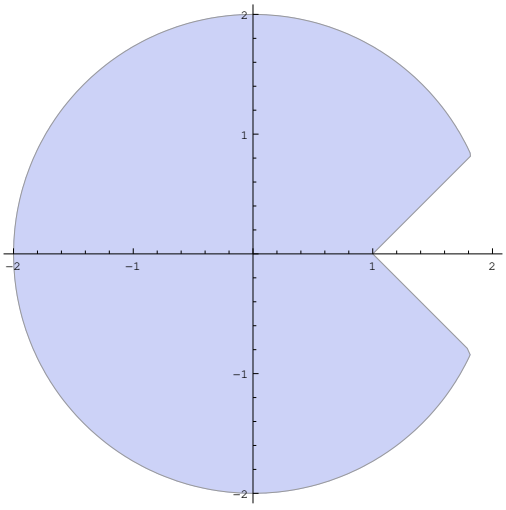


Figure A.1: An example of a Δ-domain, $\Delta(\pi/4, 2)$.

A.1 Composition Schema

This is a generalization of the composition schema given on pg. 411 of [6]. Consider the generating function $h(z) = f(g(z))$. f, g, and h each have a radius of converge labeled ρ_f, ρ_g, and ρ_h respectively. Similarly, set $\tau_f = f(\rho_f)$, $\tau_g = g(\rho_g)$, and $\tau_h = h(\rho_h)$.

A singular expansion of a function is its power series expansion centered at its radius of convergence. For a generating function with positive coefficients, the radius of convergence is also its dominant singularity, however we need not restrict our attention to the composition of generating functions. To this end, we describe two cases observed in the analysis of this paper used to determine the singular expansion of h.

1. *subcritical case*: for all $r \in \mathbb{C}$ such that $g(r) = \rho_f$, $|\rho_g| < |r|$. Namely, as z increases from the origin in concentric circles, ρ_g appears on the boundary of the disc before r and thus triggers a singularity of $f(g(z))$. Therefore, $\rho_h = \rho_g$ and $\tau_h = f(\rho_g)$. Around ρ_g, f is analytic. Then the singular expansion of h is the composition of the Taylor expansion of f at τ_g with the singular expansion of g at ρ_g.

2. *supercritical case*: there exists $r \in \mathbb{C}$ such that $g(r) = \rho_f$ and $|r| < |\rho_g|$. Namely, as z increases from the origin in concentric circles, r appears on the boundary of the disc before ρ_g and thus triggers a singularity of $f(g(z))$. Therefore, $\rho_h = r$ and $\tau_h = f(r)$. Around r, g is analytic, so the singular of h is the composition of the singular expansion of f at ρ_f with the Taylor expansion of g at r.

In the case where f and g have non-negative coefficients, the radii of convergence are the dominant singularities, which from Pringsheim's Theorem are known to be positive and real valued. Thus one can restrict the attention of r to the positive real line as h will also satisfy the conditions of Pringsheim's Theorem.

A.2 Analytic Transfer

Theorem 27. ***(Flajolet, Sedgewick)*** *Let α be an arbitrary complex number in $\mathbb{C}\backslash\mathbb{Z}_{\leq 0}$. The coefficient of z^n in*

$$f(z) = (1-z)^{-\alpha}$$

admits for large n a complete asymptotic expansion in descending powers of n,

$$[z^n]f(z) \sim \frac{n^{\alpha-1}}{\Gamma(\alpha)}(1 + O(n^{-1})).$$

Proof. Here we provide a sketch of the proof. The full details can be found in [5].

Our interest is in approximating $[z^n]f(z)$ as n grows large. To do so, consider Cauchy's coefficient formula

$$[z^n]f(z) = \frac{1}{2\pi i}\int_{\mathcal{C}} f(z)\frac{dz}{z^{n+1}},$$

a consequence of Cauchy's Integral Formula where $\mathcal{C}$ is a contour that encircles the origin such that $f(z)$ is analytic in the disc bounded by $\mathcal{C}$. For $f(z) = (1-z)^{-\alpha}$, any radius less than 1 suffices. The contour can be deformed without changing the value of the integral so long as the region preserves analyticity. By specifying an advantageous contour and performing a change of variable, one can approximate the above integral by the Γ-function. In doing so we arrive at the asymptotics.

The contour is most easily specified by a series of deformations. Begin with $\mathcal{C}_0$, a circle of radius $\frac{1}{2}$ centered at $z = 0$. Deform $\mathcal{C}_0$ into $\mathcal{C}_1$, a circle with larger radius $R > 1$, and a notch $\mathcal{H}(n)$ that wraps around the half line $\mathbb{R}_{\geq 1}$, defined piecewise as

$$\mathcal{H}(n) = \mathcal{H}^{-}(n) \cup \mathcal{H}^{+}(n) \cup \mathcal{H}^{\circ}(n)$$

with

$$\begin{cases} \mathcal{H}^-(n) = \{z = \omega - \frac{i}{n},\ 1 \le \omega \le R\} \\ \mathcal{H}^+(n) = \{z = \omega + \frac{i}{n},\ 1 \le \omega \le R\} \\ \mathcal{H}^\circ(n) = \{z = \omega - \frac{e^{i\varphi}}{n},\ \varphi \in [\frac{-\pi}{2}, \frac{\pi}{2}]\} \end{cases}$$

Note that as R tends to infinity,

$$\frac{(1-z)^{-\alpha}}{z^{n+1}} = O(R^{-n}).$$

Namely, $(1-z)^{-\alpha}$ grows in modulus as $O(R)$ for α fixed. This can seen by representing $(1-z)^{-\alpha}$ as $(Re^{i\theta})^{-\alpha}$, however the details are omitted here.

Now consider the change of variable $z = 1 + \frac{t}{n}$. The integrand is transformed into

$$\begin{aligned} \frac{1}{2\pi i}\int_{\mathcal{C}_1+\mathcal{H}(n)} (1-z)^{-\alpha}\frac{dz}{z^{n+1}} &= \frac{1}{2\pi i}\int_{\mathcal{C}_1+\mathcal{H}} \left(\frac{-t}{n}\right)^{-\alpha}\left(1+\frac{t}{n}\right)^{-n-1}\frac{1}{n}dt \\ &= \frac{n^{\alpha-1}}{2\pi i}\int_{\mathcal{H}} (-t)^{-\alpha}\left(1+\frac{t}{n}\right)^{-n-1} dt + O(R^{-n}) \end{aligned}$$

where

$$\mathcal{H} = \begin{cases} \mathcal{H}^- = \{z = \omega - i,\ \omega \ge 0\} \\ \mathcal{H}^+ = \{z = \omega + i,\ \omega \ge 0\} \\ \mathcal{H}^\circ = \{z = -e^{i\varphi},\ \varphi \in [\frac{-\pi}{2}, \frac{\pi}{2}]\} \end{cases}$$

by letting R tend to infinity. The notch $\mathcal{H}$ is commonly referred to as a Hankel contour used to represent the Γ-function. $\mathcal{H}$ starts at positive infinity, winds around the origin at a distance 1 in the positive direction from the real axis, then approaches infinity again along the line $z = -i$.

Recall the asymptotic form of the exponential function

$$e^{-t} = \lim_{n\to\infty}\left(1+\frac{t}{n}\right)^{-n}.$$

Then

$$\lim_{n\to\infty}\left(1+\frac{t}{n}\right)^{-n-1} = e^{-t}\lim_{n\to\infty}\left(1+\frac{t}{n}\right)^{-1} = e^{-t}(1+O(n^{-1})).$$

Now the integrand can be rewritten,

$$\frac{n^{\alpha-1}}{2\pi i}\int_{\mathcal{H}} (-t)^{-\alpha}\left(1+\frac{t}{n}\right)^{-n-1} dt = \frac{n^{\alpha-1}}{2\pi i}\int_{\mathcal{H}} (-t)^{-\alpha}e^{-t}dt\ (1+O(n^{-1})).$$

Hankel's Formula for the Γ-function relates the integral to $\Gamma(s)^{-1}$,

$$\Gamma(s)^{-1} = \frac{-1}{2\pi i} \int_{\mathcal{H}} (-t)^{-s} e^{-t} dt.$$

Lastly,

$$[z^n](1-z)^{-\alpha} = \frac{n^{\alpha-1}}{\Gamma(\alpha)}(1 + O(n^{-1})), \quad n \to \infty.$$

□

Bibliography

[1] Anderson, J. E., Huang, F. W., Penner, R. C., Reidys, C. M. (2012) Topology of RNA-RNA Interaction Structures. Journal of Computational Biology. 19, 928-943.

[2] Chen, W. Y., Han, H. S., Reidys, C. M. (2009) Random k-noncrossing RNA structures. Proceedings of the National Academy of Sciences. 106(52) 22061-22066.

[3] Crick, F. H. C. (1958) On protein synthesis. Symp. Society of Experimental Biology. 12, 138–163.

[4] Ellis-Monaghan, J. and Moffatt, I. (2013) *Graphs on Surfaces: Dualities, Polynomials, and Knots*. Springer-Verlag New York.

[5] Flajolet, P. and Odlyzko, A (1990) Singularity Analysis of Generating Functions. Journal of Discrete Mathematics 3(2), 216-240.

[6] Flajolet, P. and Sedgewick, R. (2009). *Analytic Combinatorics*. Cambridge University Press.

[7] Fu, B. M., Han, H. S., Reidys, C. M. (2015) On RNA-RNA interaction structures of fixed topological genus. Mathematical Biosciences. 262, 88-104.

[8] Gorodkin, J. and Ruzzo, W. L. (Eds.). (2014). RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. New York, NY: Springer.

[9] Graham, R., Knuth, D., Patashnik, O. (1994) *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley.

[10] Han, H. S., Li, T. X., Reidys, C. M. (2014) Combinatorics of γ-structures. Journal of Computational Biology, (21), 591-608.

[11] Han, H. S., Reidys, C. (2012) The 5'-3' Distance of RNA Secondary Structures. Journal of Computational Biology. 19, 867-878.

[12] Huang, F. W., Nebel, M. E., Reidys, C. M. (2013) Shapes of topological RNA structures. Mathematical Biosciences 245, 216-225.

[13] Huang, F. W. D. and Reidys, C. M. (2012) On the combinatorics of sparsification. Algorithm. Mol. Biol. 7(28).

[14] Lander, E. S. (2011) Initial impact of the sequencing of the human genome. Nature. 470, 187–197.

[15] Li, T. X., Reidys, C. M. (2018). The rainbow-spectrum of RNA secondary structures. Bulletin of Mathematical Biology.

[16] Mercer, T. R., Dinger, M. E., Mattick, J. S. (2009) Long non-coding RNAs: insights into function. Nature Reviews Genetics 10. 155-159.

[17] Möhl, M., Salari, R., Will, S., Backofen, R., Sahinalp, S. C. (2010) Sparsification of RNA structure prediction including pseudoknots. Alg. for Mol. Biology. 5(39).

[18] Nebel, M. E. and Weinberg, F. (2012) Algebraic and combinatorial properties of common RNA pseudoknot classes with applications. J. Comput. Biol. 19, 1134-1150.

[19] Penner, R. C. and Waterman, M. S. (1993) Spaces of RNA secondary structures. Adv. Math. 101(31).

[20] Reidys, C. M. (2011) *Combinatorial Computational Biology of RNA*. Springer.

[21] Thomassen, C. (1989) The graph genus problem is NP-complete. J. Algorithms. 10(4) 568-576.

[22] Tinoco, I. and Bustamante, C. (1999) How RNA Folds. Journal of Molecular Biology. 293, 271–281.

[23] Von, M., Vernizzi, G., Zee, A., Orland, H. (2008) Topological Classification of RNA Structures. Journal of Molecular Biology. 379(4), 900-911.

[24] Watson, J. (1965). *Molecular Biology of the Gene*, Volume 1.

[25] Yoffe, A. M., Prinsen, P., Gelbart, W. M., Ben-Shaul, A. (2011) The ends of a large RNA molecule are necessarily close. Nucleic Acids Research. 39(1) 292-299.