3-21-2014

Properties of Graphs Used to Model DNA Recombination

Ryan Arredondo
University of South Florida, rarredon@mail.usf.edu

Properties of Graphs Used to Model DNA Recombination

by

Ryan C. Arredondo

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts
Department of Mathematics & Statistics
College of Arts and Sciences
University of South Florida

Co-Major Professor: Nataša Jonoska, Ph.D.
Co-Major Professor: Masahiko Saito, Ph.D.
Dmytro Savchuk, Ph.D.

Date of Approval:
March 21, 2014

Keywords: Ciliates, Double occurrence words, Chord diagrams, Orientable genus of graphs, Ribbon graphs

Acknowledgments

I would like to thank my advisors, Nataša Jonoska and Masahiko Saito, as well as, Dmytro Savchuk for serving on my defense committee and providing careful review of my thesis. I would also like to thank Jonathon Burns, Egor Dolzhenko, and Timothy Yeatman for contributions that generously affected the outcome of the work presented here. Furthermore, I am indebted to Nicole Collins for her gracious support. The work presented here has been supported in part by NSF grant DMS-0900671 and NIH grant 1R01GM109459-01.

Table of Contents

List of Tables

List of Figures

Abstract

A model for DNA recombination uses 4-valent rigid vertex graphs, called assembly graphs [1]. An assembly graph, similarly to the projection of knots, can be associated with an unsigned Gauss code, or double occurrence word [2]. We define biologically motivated reductions that act on double occurrence words and, in turn, on their associated assembly graphs. For every double occurrence word w there is a sequence of reduction operations that may be applied to w so that what remains is the empty word, ϵ. Then the nesting index of a word w, denoted by $\mathrm{NI}(w)$, is defined to to be the least number of reduction operations necessary to reduce w to ϵ. The nesting index is the first property of assembly graphs that we study. We use chord diagrams as tools in our study of the nesting index. We observe two double occurrence words that correspond to the same circle graph, but that have arbitrarily large differences in nesting index values.

In 2012, Buck et al. [5] considered the cellular embeddings of assembly graphs into orientable surfaces. The genus range of an assembly graph Γ, denoted by $\mathrm{gr}(\Gamma)$, was defined to be the set of integers g where g is the genus of an orientable surface F into which Γ cellularly embeds. The genus range is the second property of assembly graphs that we study. We generalize the notion of the genus range to that of the genus spectrum, where for each $g \in \mathrm{gr}(\Gamma)$ we consider the number of orientable surfaces F obtained from Γ by a special construction, called a ribbon graph construction [5], that have genus g. By considering this more general notion we gain a better understanding of the genus range property. Lastly, we show how one can obtain the genus spectrum of a double occurrence word from the genus spectrums of its irreducible parts, i.e., its double occurrence subwords.

In the final chapter we consider constructions of double occurrence words that recognize certain values for nesting index and genus range. In general, we find that for arbitrary values of nesting index ≥ 2 and genus range, there is a double occurrence word that recognizes those values.

Chapter 1
Introduction

A vertex in a graph is *rigid* if the cyclic order of edges incident to that vertex cannot be altered without changing the overall structure of the graph. In this thesis we discuss two properties of 4-valent rigid vertex multigraphs, called *assembly graphs*, that are used to model processes of DNA recombination. The model is most prominently applied to various species of ciliates, such as *Oxytricha Nova* [1], which contain two types of DNA: one in the somatic macronucleus and one in germline micronucleus. The micronuclear DNA is made up of segments of DNA called internal eliminated sequences (*IES*s) and macronuclear destined sequences (*MDS*s), while the macronuclear DNA consists of MDS segments only. Furthermore, the order of the MDS segments in the micronuclear DNA is permuted relative to the macronuclear DNA and the IESs consist of noncoding "junk" DNA. During conjugation the IESs are excised from the micronuclear DNA and the MDSs are rearranged so that a new copy of macronuclear DNA is formed. The assembly graph model is a discrete approach to modeling these rearrangement processes. In this model the vertices of the assembly graph represent places where the DNA aligns at certain guiding sequences; the edges represent the IES and MDS segments of the micronuclear DNA. For a more thorough treatment of the assembly graph model, we refer the reader to [1].

Chapter 2 is an introduction to the required definitions and notations that will be used throughout the thesis. In particular, we define double occurrence words and assembly graphs. We describe the link between double occurrence words and simple assembly graphs, that is, assembly graphs that admit a path that visits every edge without taking 90° turns at any vertex. Next, we introduce chord diagrams and circle graphs as tools for working with double occurrence words. Afterwards we define a concatenation for double occurrence words and discuss how this relates to their associated assembly graphs.

In chapter 3 we discuss the nesting index property for assembly graphs. While this property has not previously been studied from a mathematical approach, it is motivated by several papers ([16] and [9], for example) in ciliate biology, wherein the researchers observe frequently occurring sequences in the scrambled micronucleus of certain ciliate species and relate these patterns to the evolutionary origins of the species.

These sequences correspond to double occurrence words of a particular form that we call *repeat words* and *return words*. We use these words to define two reduction operations that act on double occurrence words and in turn, on their associated assembly graphs. The first reduction operation is to remove all subwords that are repeat words and return words. The second reduction operation is to remove both occurrences of a single letter. We apply either of these reduction operations to a double occurrence word w until w is reduced to the empty word ϵ. There are some double occurrence words which can not be reduced to ϵ by only removing subwords that repeat words or return words. A word that can be reduced to ϵ by applying only the first reduction operation is called 1-*reducible*. We use the chord diagram representation of double occurrence words to give a characterization of double occurrence words that are 1-reducible. The nesting index of a double occurrence word w is the least number of reduction operations that can be applied to reduce w to ϵ. From the biological motivation, the nesting index could provide information about the evolutionary complexity of a scrambled ciliate genome. We characterize words whose nesting index is 1 and we use the characterization of 1-reducible double occurrence words to construct words with arbitrarily high nesting index. We provide examples of words whose chord diagrams have similar intersection graphs, called *circle graphs*, but arbitrarily large differences in nesting index values.

Notions in topological graph theory, such as graph embeddings and the genus range of a graph, have been extensively studied for graphs with non-rigid vertices [11]. The minimum genus of virtual knot diagrams and diagrams corresponding to signed Gauss codes is of interest in knot theory, for example, in [4] and [7]. In [5], Buck et al. considered cellular embeddings of assembly graphs into orientable surfaces that preserve the rigidity of vertices in the embedded image. The genus range of an assembly graph Γ, denoted $\mathrm{gr}(\Gamma)$, was defined to be the set of integers g such that g is the genus of some surface F into which Γ can be cellularly embedded in this manner. In Chapter 4 we study the genus range and also some more general notions. We investigate genus range of an assembly graph Γ obtained by connecting two assembly graphs Γ_1 and Γ_2. It turns out that for arbitrary Γ_1 and Γ_2 we can characterize the genus range of Γ in terms of $\mathrm{gr}(\Gamma_1)$ and $\mathrm{gr}(\Gamma_2)$ depending on certain conditions satisfied by Γ_1 and Γ_2; this is a generalization of a result in [5] where they considered Γ_1 to be the assembly graph corresponding to the double occurrence word 1212 and Γ_2 to be arbitrary. We then develop the notion of a genus spectrum for an assembly graph Γ where for each $g \in \mathrm{gr}(\Gamma)$ we associate with g the number of orientable surfaces F obtained from Γ by a special construction, called a *ribbon graph construction*, that have genus g. We consider a construction called the *connected sum* of two assembly graphs Γ_1 and Γ_2 and we characterize the genus spectrum of such a construction in terms

of the genus spectrums of Γ_1 and Γ_2. We then define the genus spectrum for a double occurrence word w by isolating a particular edge e in the assembly graph that corresponds to w and considering the number of boundary components that the edge e belongs to. We prove that repeat words and return words realize certain values for genus spectrum. The final result of this chapter gives us an explicit formula for computing the genus spectrum of a double occurrence word in terms of the genus spectrums of its irreducible parts, i.e., its double occurrence subwords.

In Chapter 5 we make comparisons between the nesting index and genus range properties. In particular, we provide examples of assembly graphs that have nesting index values ≤ 2 and arbitrary genus ranges. In contrast, we provide examples of assembly graphs with genus range $\{0\}$ and arbitrary nesting index. We use the two examples to construct assembly graphs with arbitrary genus range and arbitrary nesting index ≥ 2.

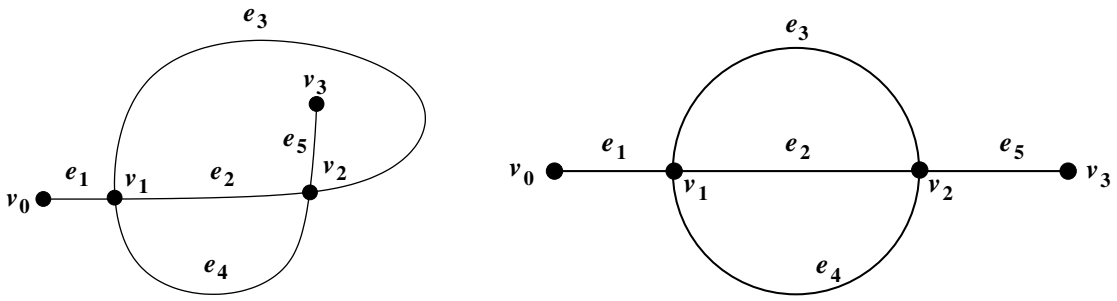
Chapter 2

Preliminaries

A *word* over an alphabet Σ is a finite sequence of elements from Σ, usually displayed as a string $w = a_1a_2\cdots a_n$ where $a_i \in \Sigma$ for $1 \leq i \leq n$. The elements of Σ are called *symbols* or *letters*. A word w with n symbols has *length* $|w| = n$. A word u is a *subword* of a word w, denoted by $u \sqsubseteq w$, if we can write $w = suv$ where u and v are also words. The word with no symbols and zero length is called the *empty word*, denoted by ϵ. Denote by Σ^* the set of all finite words over Σ including ϵ. A word $w \in \Sigma^*$ is a *double occurrence word* if for all $a \in \Sigma$, a appears in w either two times or not at all. Given two double occurrence words $w = a_1a_2\cdots a_n$ and $w' = b_1b_2\cdots b_n$, we say that w' is a *relabeling* of w if there exists a function $f : \{a_1, \ldots, a_n\} \to \{b_1, \ldots, b_n\}$ such that $w' = f(a_1)f(a_2)\cdots f(a_n)$. For the remainder of this thesis, we consider the alphabet $\Sigma = \mathbb{N}$, the set of natural numbers. This allows us to label double occurrence words in a canonical form known as ascending order. A double occurrence word w is said to be in *ascending order* if its left-most symbol is 1 and every other symbol in w is at most 1 greater than any symbol appearing to the left of it. We use w^{asc} to denote the unique relabeling of the double occurrence word w so that w^{asc} is in ascending order. For example, if $w = 94767496$, then $w^{asc} = 12343214$. If $w = a_1a_2\cdots a_n$, then the *reverse* of w is $w^R = a_n\cdots a_2a_1$. The *size* of a double occurrence word w is the number of distinct letters in w which is precisely $|w|/2$.

Let w_1 and w_2 be double occurrence words. Then

- w_1 and w_2 are *disjoint* if they have no letters in common,
- w_1 and w_2 are *equivalent*, denoted by $w_1 \sim w_2$, if one is obtained from the other by relabeling,
- w_1 and w_2 are *reverse equivalent*, denoted by $w_1 \sim_R w_2$, if $w_1 \sim w_2$ or $w_1 \sim w_2^R$,
- $w_1 = a_1a_2\cdots a_n$ is a *cyclic permutation* of w_2 if $w_2 \in \{a_1a_2\cdots a_n, a_na_1a_2\cdots a_{n-1}, \ldots, a_2\cdots a_na_1\}$,
- w_1 and w_2 are *cyclically equivalent*, denoted by $w_1 \sim_{cyc} w_2$, if w_1 is equivalent to a cyclic permutation of either w_2 or w_2^R.



(a) Simple assembly graph (b) Non-simple assembly graph

Figure 1: Examples of assembly graphs

An undirected multigraph Γ is a pair (V, E) where V is a set of points, called *vertices*, and E is a multiset of unordered pairs of elements of V called *edges*. The vertices that make up the pair $e \in E$ are called the *endpoints* of e. An edge whose endpoints are the same vertex is called a *loop*. We say that $e \in E$ is *incident* to $v \in V$ if v is an endpoint of e. The number of edges incident to $v \in V$, denoted by $\deg(v)$, is called the *degree* of v where, by convention, a loop contributes 2 to the degree of its endpoint. A cyclic ordering of a sequence $S = (a_1, a_2, \ldots, a_n)$ is an equivalence class S^{cyc} such that $S \in S^{cyc}$ and $(b_1, b_2, \ldots, b_n) \in S^{cyc}$ implies $(b_n, b_1, b_2, \ldots, b_{n-1}) \in S^{cyc}$ and $(b_n, \ldots, b_2, b_1) \in S^{cyc}$. A vertex v with $\deg(v) = n$ is said to be *rigid* if we associate with v a cyclic ordering of a fixed sequence $(e_1, e_2, \ldots, e_n)$ consisting of all edges incident to v. Then a graph is said to have rigid vertices if altering the cyclic ordering of edges around a vertex alters the overall structure of the graph. Take for example the graphs in Figure 1; from the definition of a multigraph, these graphs the same, however, if we consider the vertices of these graphs to be rigid, then we see that the cyclic order of the edges e_3 and e_5 have been permuted and, because of this, we consider these graphs to be different. For a discussion on rigid-vertex graphs in a topological context, we refer the reader to [14]. An *assembly graph* is a multigraph with rigid vertices such that each vertex has degree 1 or 4. The vertices of degree 1 in an assembly graph Γ are called the *endpoints* of Γ. Figure 1 shows two examples of assembly graphs with endpoints. Two assembly graphs are called *isomorphic* if there exists a graph isomorphism between them that preserves the cyclic order of edges associated with each rigid vertex.

If $v \in V$ is a rigid vertex of degree 4 associated with the sequence (e_0, e_1, e_2, e_3), then we say that e_0 and e_2 are *neighbors* of e_1 and e_3 with respect to v and vice-versa. In the event that one of the edges e_i for $i = 0, 1, 2, 3$ is a loop and $e_i = e_{i+1}$, we say that e_{i-1} and e_{i+2} are both neighbors and not neighbors of $e_i = e_{i+1}$, where indices are taken modulo 4. In Figure 1(a), vertex v_1 is associated with the cyclic ordering of edges (e_1, e_3, e_2, e_4), hence, e_1 has neighbors e_3 and e_4 with respect to v_1. For an assembly

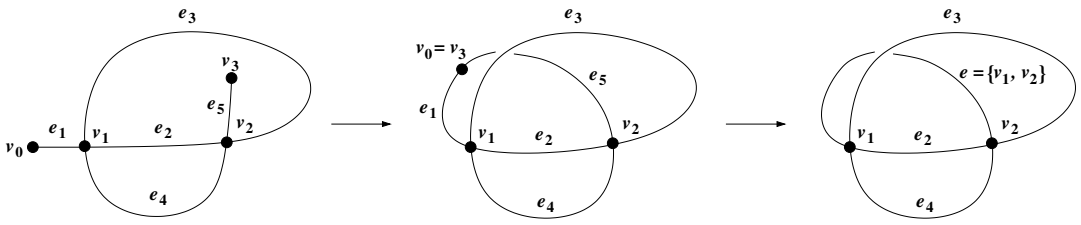


Figure 2: Closure of assembly graph from Figure 1(a)

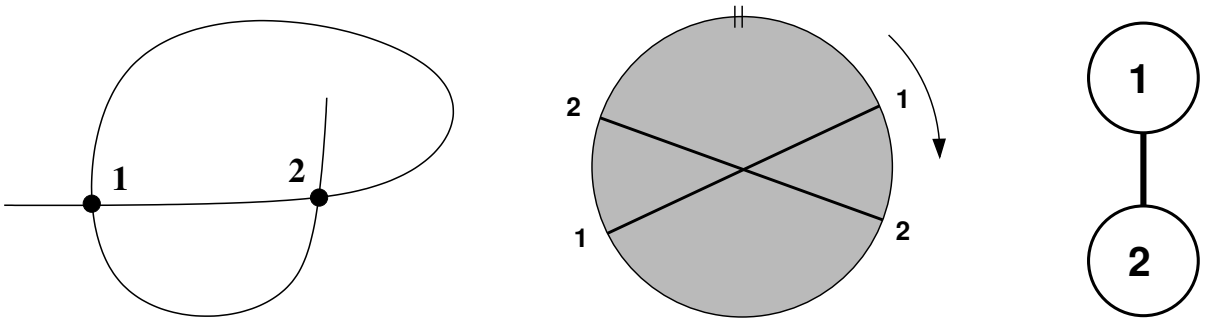
graph Γ with endpoints v_0 and v_n, a *transverse path* is a sequence $\gamma = (v_0, e_1, v_1, e_2, \ldots, e_n, v_n)$ satisfying: (1) $(v_0, ..., v_n)$ is a sequence of a subset of vertices of Γ with possible repetition of the same vertex at most twice, (2) $\{e_1, \ldots, e_n\}$ is a set of distinct edges such that e_i is incident to v_{i-1} and v_i for $i = 2, \ldots, n$, and (3) e_i is not a neighbor of e_{i-1} with respect to v_{i-1} for $i = 2, \ldots, n$. Similarly, for an assembly graph Γ without endpoints, a transverse path is a sequence $\gamma = (v_0, e_1, v_1, e_2, \ldots, v_n, e_{n+1})$ such that γ satisfies (1), (2), and (3) above, and also e_{n+1} is an edge distinct from $e_1, \ldots, e_n$ which is incident to v_n and v_0 so that e_{n+1} and e_n are not neighbors with respect to v_n and e_{n+1} and e_1 are not neighbors with respect to v_0. An assembly graph Γ is *simple* if Γ admits a transverse *Eulerian* path, that is, a transverse path that contains every edge in Γ exactly once. The assembly graph in Figure 1(a) has a transverse path with endpoints v_0 and v_3, hence, is simple. The graph Γ in Figure 1(b) has two transverse components, one without endpoints and one with endpoints v_0 and v_3, hence, is non-simple. In the remainder of this thesis an assembly graph is assumed to be simple, unless otherwise stated. Given an assembly graph Γ with endpoints v_0 and v_n, we use $\overline{\Gamma}$ to denote the *closure of* Γ, that is, the graph obtained from Γ by identifying vertices v_0 and v_n and then removing the vertex and replacing its two adjacent edges with one edge, called the *closure edge* of Γ. Figure 2 shows the process of creating the closure of the graph Γ from Figure 1(a).

We now establish the link between simple assembly graphs and double occurrence words. Note that in the transverse path of a simple assembly graph (with or without endpoints), each vertex which is not an endpoint is visited exactly twice. Thus, if $(v_0, e_1, v_1, e_2, \ldots, e_n, v_n)$ or $(e_1, v_1, e_2, \ldots, v_{n-1}, e_n)$ is the transverse path of Γ with or without endpoints, respectively, then we can associate Γ with the double occurrence word $w = v_1 v_2 \cdots v_{n-1}$. The assembly graph Γ in Figure 1(a) and its closure $\overline{\Gamma}$ in Figure 2 have transverse paths

$$(v_0, e_1, v_1, e_2, v_2, e_3, v_1, e_4, v_2, e_5, v_3) \quad \text{and} \quad (v_1, e_2, v_2, e_3, v_1, e_4, v_2, e),$$

respectively, and hence, are associated with the double occurrence word $v_1 v_2 v_1 v_2$. Mapping the letters $v_1 \mapsto 1$ and $v_2 \mapsto 2$ we may relabel the word corresponding to Γ to be 1212 in ascending order. Conversely,

one can also start with a double occurrence word w and create a corresponding assembly graph $\Gamma(w)$; for each distinct letter a in w, we designate a vertex v_a in $\Gamma(w)$ so that if a letter b is adjacent to a in w we construct an edge between vertices v_a and v_b in $\Gamma(w)$. For the vertices v_a and v_b that correspond to the first and last letters a and b in w, respectively, we add two edges: one that is incident to v_a and an initial endpoint v_i, and one that is incident to v_b and a terminal endpoint v_f. Then every vertex in $\Gamma(w)$ has degree 1 or 4, and if we cyclically order the edges around the vertices of $\Gamma(w)$ so that $\Gamma(w)$ is simple and can be associated with the word w, then we induce a rigidity of the vertices in $\Gamma(w)$. We use $\overline{\Gamma}(w)$ to denote $\overline{\Gamma(w)}$, i.e., the closure of $\Gamma(w)$.

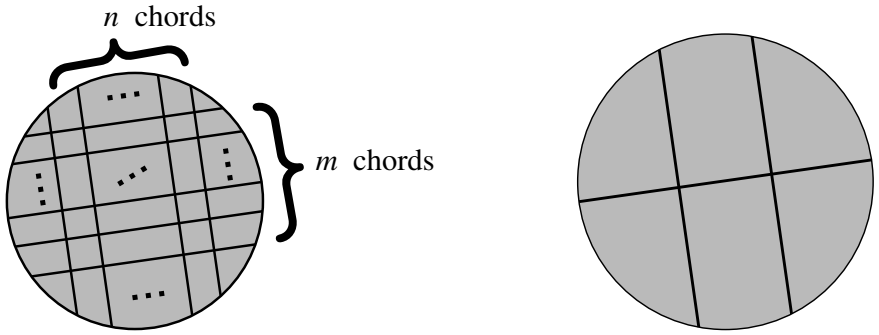


(a) Assembly graph of 1212 (b) Chord diagram of 1212 (c) Circle graph of 1212

Figure 3: Representations of the the double occurrence word 1212

A *chord diagram* is a pictorial representation of a double occurrence word w obtained by arranging the $2n$ letters of w around the circumference of a circle and then for each letter, joining the two occurrences of the same letter by a chord of the circle. Figure 3(b) shows the chord diagram for the double occurrence word 1212. A chord diagram $\mathcal{C}'$ is said to be a *sub-chord diagram* of a chord diagram $\mathcal{C}$ if the chords of $\mathcal{C}'$ make up some subset of the chords of $\mathcal{C}$. Note that two double occurrence words may correspond to chord diagrams which differ only by the labeling of chords, for instance, the chord diagrams for 123231 and 121233. Occasionally a chord diagram is given a *base point* and an *orientation* to emphasize the word that corresponds to that chord diagram. The basepoint of the chord diagram in Figure 3(b) is indicated by two dashes on the boundary of the circle and its orientation is indicated by the clock-wise directed arrow outside of the circle. The circle graph G of a chord diagram $\mathcal{C}$ is the intersection graph of the chords in $\mathcal{C}$, that is, G is the graph whose vertex set is in correspondence with the set of chords in $\mathcal{C}$ such that two vertices in G are joined by an edge if and only if their corresponding chords in $\mathcal{C}$ intersect. Figure 3(c) shows the circle

graph for the double occurrence word 1212. For integers $1 \leq m \leq n$, we use $\mathcal{C}_{m\times n}$ to denote the chord diagram of $m+n$ chords that is depicted in Figure 4(a). The chord diagram $\mathcal{C}_{1\times 2}$ in Figure 4(b) will be used in Chapter 3 to characterize words that are 1-reducible.



(a) Chord diagram $\mathcal{C}_{m\times n}$ (b) Chord diagram $\mathcal{C}_{1\times 2}$

Figure 4: Special chord diagrams

We will often use the notion of concatenating double occurrence words and the analogous notion of connecting two assembly graphs. Let w_1 and w_2 be double occurrence words. Then we use $w_1 * w_2$ to denote the concatenation $w_1 w_2'$ where w_2' is the relabeling of w_2 so that $w_1 w_2'$ is also a double occurrence word; $w_1 * w_2$ is called the *double occurrence word concatenation* of w_1 and w_2. When the context is clear we may omit the "$*$", for instance, $\Gamma(w_1 w_2)$ will always mean $\Gamma(w_1 * w_2)$. Let w^n denote the double occurrence word concatenation $w*w*\cdots*w$ of n copies of w. To define the analogous notion for assembly graphs, let us fix edges e_1 and e_2 in assembly graphs without endpoints Γ_1 and Γ_2, respectively, and prescribe orientations to Γ_1 and Γ_2. Then we construct the graph Γ obtained by *connecting* Γ_1 and Γ_2 *through* edges e_1 and e_2 by the following procedure as depicted in Figure 5: (i) cut edges e_1 and e_2 introducing initial endpoints v_{i_1}, v_{i_2} and terminal endpoints v_{f_1}, v_{f_2} to Γ_1 and Γ_2 according on their orientations, respectively, (ii) identify the terminal endpoints of each graph with the initial endpoints of the other graph, (iii) and replace edges incident to v_{i_2} (resp. v_{f_2}) with a single edge e_1' (resp. e_2') so that the resulting graph Γ is an assembly graph without endpoints. Another way to think of this procedure: if we let w_1 and w_2 be the double occurrence words associated with the oriented graphs Γ_1 and Γ_2 after introducing endpoints in step (i), then Γ is the same as $\overline{\Gamma}(w_1 w_2)$, or, the assembly graph obtained by connecting $\overline{\Gamma}(w_1)$ and $\overline{\Gamma}(w_2)$ through the closure edges of $\Gamma(w_1)$ and $\Gamma(w_2)$. Note that e_2' is the closure edge for $\Gamma(w_1 w_2)$. We will refer back to these observations in Chapters 4 and 5.

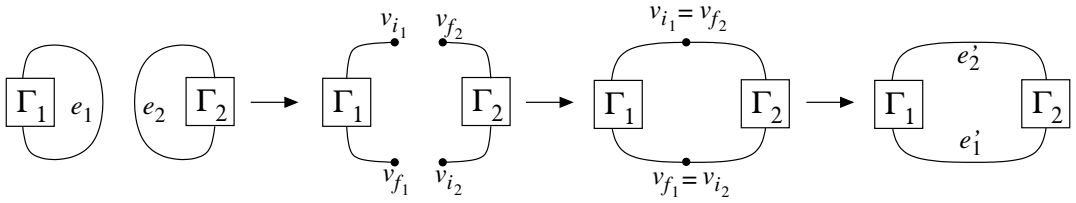


Figure 5: Procedure for connecting two assembly graphs through the edges e_1 and e_2.

Chapter 3

Nesting Index

In this chapter we discuss a property of assembly graphs, called the "nesting index" of an assembly graph. A majority of the material covered in this chapter was accepted to appear in the journal *Congressus Numerantium* as part of the proceedings to the 44th Southeastern International Conference on Combinatorics, Graph Theory, and Computing. Aside from the addition of Lemma 3.2 and some figures to assist with the proof of Theorem 3.1, only minor changes have been made to the present version from the original [3].

3.1 Reduction notation

We first fix some notation for the reduction of double occurrence words.

DEFINITION 3.1 If $w = w_1 v w_2$ where w and v are both double occurrence words, then $w - v = w_1 w_2$ is called the *subword removal* of v from w.

DEFINITION 3.2 If $D = \{v_1, v_2, \ldots, v_n\}$ is a set containing disjoint double occurrence subwords of w, and ϕ is a permutation of $\{1, \ldots, n\}$, then we use $w -_\phi D$ to mean $((\cdots((w - v_{\phi(1)}) - v_{\phi(2)})\cdots) - v_{\phi(n)})$.

REMARK 3.1 If D is a set of disjoint double occurrence subwords of w and ϕ and ϕ' are two permutations of $\{1, \ldots, n\}$, then $w -_\phi D = w -_{\phi'} D$ and hence, we simply write $w - D$.

DEFINITION 3.3 If $w = w_1 a w_2 a w_3$ is a double occurrence word and $a \in \Sigma$, then $w - a = w_1 w_2 w_3$ is called the *letter removal* of a from w.

EXAMPLE 3.1 Let $w = 1123234554$. Then

1. $w - 4554 = 112323$,
2. $w - \{11, 4554\} = ((w - 4554) - 11) = 2323$, and
3. $w - 3 = 11224554$.

3.2 Biological motivation

Several sources ([12], [16], and [9], for example) have observed frequently occurring sequences in the scrambled micronuclear genome of certain ciliate species. The sources propose theories that relate the nesting of these sequences in micronuclear DNA to the evolutionary complexity of the species. Potentially, the more nested the sequences are, the more mutated, or evolved, the ciliate species may be. In the present section we introduce double occurrence words of a particular form, called repeat words and return words, to match the observed sequences and we use these words to introduce the notion of a nesting index for double occurrence words. From a biological perspective the nesting index could act as a measurement of the evolutionary complexity of a scrambled ciliate genome.

There is also the belief [13] that during conjugation, wherein the micronuclear genome undergoes processes of rearrangement to create a new copy of the macronuclear genome, the parts of the micronuclear genome corresponding to the frequently occurring sequences (repeat words and return words) become aligned before other parts of the genome. Then from this perspective the nesting index would provide insight into the number of steps in the rearrangement process of the micronuclear genome.

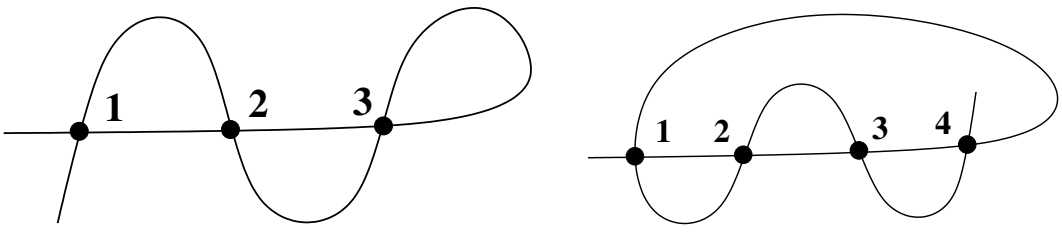
3.3 Double occurrence word reductions and nesting index

DEFINITION 3.4 A *return word* is a word of the form

$$a_1 a_2 \cdots a_n a_n \cdots a_2 a_1, \qquad a_i \in \Sigma \text{ for all } i, \text{ and } a_i \neq a_j \text{ for } i \neq j.$$

A *repeat word* is a word of the form

$$a_1 a_2 \cdots a_n a_1 a_2 \cdots a_n, \qquad a_i \in \Sigma \text{ for all } i, \text{ and } a_i \neq a_j \text{ for } i \neq j.$$



(a) 123321 is a return word (b) 12341234 is a repeat word

Figure 6: Assembly graphs of a repeat word and a return word

REMARK 3.2 All repeat words and return words are double occurrence words.

DEFINITION 3.5 Let $\mathcal{R}$ denote the set of all repeat words and return words and let w be a double occurrence word. Then a word u said to be a *maximal subword* of w with respect to $\mathcal{R}$ if $u \sqsubseteq w$, $u \in \mathcal{R}$, and $u \sqsubseteq v \sqsubseteq w$ implies $v \notin \mathcal{R}$ or $u = v$.

When we wish to distinguish between repeat words and return words we sometimes say a maximal return word of w to mean a return word that is a maximal subword of w with respect to $\mathcal{R}$ and similarly for a maximal repeat word of w. Note that the word aa for some $a \in \Sigma$ may be a maximal subword with respect to $\mathcal{R}$ which is both a repeat word and a return word. In the remainder of the thesis, a maximal subword of a word w will mean a maximal subword with respect to $\mathcal{R}$.

EXAMPLE 3.2 Let $w = 1233214545$. Then 123321, 2332, 33, and 4545, are all subwords of w which are repeat or return words. 2332 and 33 are not maximal subwords because they are subwords of the return word 123321. On the other hand, 123321 and 4545 are maximal subwords of w.

REMARK 3.3 If s is a repeat word or a return word and we write $s = uv$ where u and v are both non-empty, then neither u nor v is a double occurrence word.

Note that if S is a set of double occurrence subwords of w, and the words in S are not pairwise disjoint, then $w - S$ may not be defined as it is for disjoint subwords in Definition 3.2. The following lemma and corollary show that if $\mathcal{M}_w$ is the set of maximal subwords of a double occurrence word w, then $\mathcal{M}_w$ is a set of disjoint subwords of w, hence, $w - \mathcal{M}_w$ is defined.

LEMMA 3.1 *Let w be a double occurrence word with subwords s_1 and s_2, such that $s_1 \in \mathcal{R}$ and $s_2 \in \mathcal{R}$. If $s_1 \not\sqsubseteq s_2$ and $s_2 \not\sqsubseteq s_1$, then s_1 and s_2 are disjoint words.*

Proof. Recall that two words w_1 and w_2 are disjoint if they share no letters in common. Assume to the contrary that s_1 and s_2 have at least one letter a in common. First, consider the case that there exists a subword separating s_1 and s_2, that is $w = u_1 s_1 u_2 s_2 u_3$. However, since s_1 and s_2 are double occurrence words (Remark 3.2), the letter a appears in w at least 4 times which contradicts the assumption that w is a double occurrence word. Note that the outcome is the same if we let any combination of u_1, u_2 and u_3 be empty words.

Then suppose the subwords s_1 and s_2 have an overlap, meaning that without loss of generality we can write $s_1 = v_1u$ and $s_2 = uv_2$. Since $s_1 \not\sqsubseteq s_2$ and $s_2 \not\sqsubseteq s_1$, it follows that v_1 and v_2 are non-empty. However, u can not be a double occurrence word (Remark 3.3). Then there exists a letter a in u such that a has only one occurrence in u. However, since s_1 and s_2 are double occurrence words (Remark 3.2), then a has at least 3 occurrences in w. This contradicts the fact that w is a double occurrence word. □

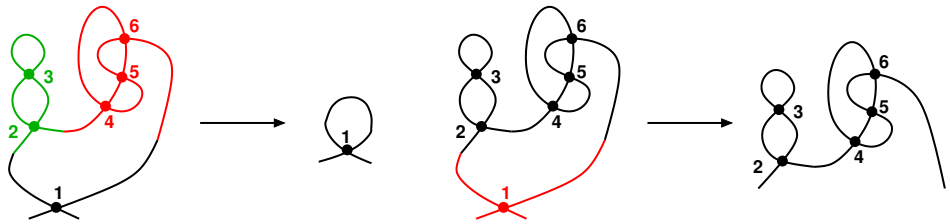
Directly from Definition 3.5 we obtain the following corollary.

COROLLARY 3.1 *If u_1 and u_2 are distinct maximal subwords of a double occurrence word w, then u_1 and u_2 are disjoint words.*

Using the notion of maximal subwords we define two reduction operations on double occurrence words.

DEFINITION 3.6 Let w be a double occurrence word and let $\mathcal{M}_w$ be the set of all maximal subwords of w with respect to $\mathcal{R}$. Then we say w' is obtained from w by *reduction operation 1* if $w' = w - \mathcal{M}_w$ or w' is obtained from w by *reduction operation 2* if for some $a \in \Sigma$, $w' = w - a$.

Figure 7 gives an example of each reduction operation applied to the word 123324564561.



(a) 11 obtained from 123324564561 (b) 2332456456 obtained from 123324564561

Figure 7: Examples of reduction operations 1 (left) and 2 (right)

DEFINITION 3.7 A *reduction* of w is a sequence of words $(u_0, u_1, \ldots, u_n)$ in which (1) $u_0 = w$, (2) for $0 \leq k < n$, u_{k+1} is obtained from u_k by application of one of the reduction operations, and (3) $u_n = \epsilon$.

Note that every double occurrence word has at least one reduction (in any case we can remove a letter from u_i to obtain a possible u_{i+1}), and most double occurrence words, in fact, have many distinct reductions.

EXAMPLE 3.3 Consider $w = 1234554231$. Applying reduction operation 1 to w gives $w_1 = 123231$. A second application of the reduction operation to w_1 gives 11, and so a third application gives ϵ. Then

$R_1 = (1234554231, 123231, 11, \epsilon)$ is a reduction of w. For a second example, if we apply reduction operation 2 to w by removing the letter 3, we get $w_1' = 12455421$. Since w_1' is a return word, an application of reduction operation 1 to w_1' gives ϵ. Then $R_2 = (1234554231, 12455421, \epsilon)$ is also a reduction of w.

DEFINITION 3.8 A double occurrence word w is called 1-*reducible* if there exists a reduction $(u_0, u_1, \ldots, u_n)$ of w such that for all $0 \leq i < n$, u_{i+1} is obtained from u_i by application of reduction operation 1.

In the previous example we saw that $w = 1234554231$ is 1-reducible by reduction R_1. In the following section we give a characterization of words which are 1-reducible.

DEFINITION 3.9 $\mathrm{NI}(w) := \min\{n : (u_0, u_1, \ldots, u_n)$ is a reduction of $w\}$ is the *nesting index* of the double occurrence word w.

Note that a word w with $\mathrm{NI}(w) = 1$ is necessarily 1-reducible. Indeed, either $|w| > 2$ and reduction operation 2 could not have been used to reduce w in one step, or $w = aa$ for some $a \in \Sigma$ which is also reduced to ϵ by applying reduction operation 1. The following lemma characterizes double occurrence words w with $\mathrm{NI}(w) = 1$.

LEMMA 3.2 *Let w be a double occurrence word. Then $\mathrm{NI}(w) = 1$ if and only if w is a concatenation of repeat words and return words.*

Proof. Suppose $\mathrm{NI}(w) = 1$ and let $\mathcal{M}_w$ be the set of all maximal subwords of w. Then, by the remark made above, $w - \mathcal{M}_w = \epsilon$ and since the words of $\mathcal{M}_w$ are maximal, none of them are subwords of another word in $\mathcal{M}_w$. Thus, we can build up w by starting with ϵ and concatenating the words in $\mathcal{M}_w$.

Conversely, if w is a concatenation of repeat words and return words, then by the definition of double occurrence word concatenation, none of them can be a subword of another, and hence, they are all maximal. Then applying reduction operation 1 to w results in the empty word and thus, $\mathrm{NI}(w) = 1$. □

In [2] it is shown that two assembly graphs Γ_1 and Γ_2 with endpoints are isomorphic if and only if the double occurrence words of Γ_1 and Γ_2 are reverse equivalent. Note that if w_1 and w_2 are reverse equivalent, then every repeat (return) word in w_1 appears as a repeat (return) word in w_2. Then there is a one-to-one correspondence between reductions of w_1 and reductions of w_2, hence, $\mathrm{NI}(w_1) = \mathrm{NI}(w_2)$. It follows that the nesting index is an invariant of isomorphic assembly graphs with endpoints.

Let us again consider the reductions R_1 and R_2 in Example 3.3. Note that the second word in R_1 is obtained from w by removing a subword of length 4. In R_2 the second word is obtained from w by a letter

removal. Although we removed less from w in the beginning for R_2, the number of reduction operations needed to reduce w to the empty word was less than in R_1. This example shows that a greedy algorithm based on the number of letters that can be removed would be incorrect for the computation of the nesting index. The current algorithm[1] to compute the nesting index is only slightly better than brute force. It is unknown whether there exists a more efficient algorithm to compute the nesting index of a double occurrence word.

Using our nesting index program we obtained counts on the number of double occurrence words (up to equivalency) with a given size and nesting index, as presented in Table 1. For words of size ≤ 9 the counts are given for all nesting index values. For words of size 10, 11, and 12, the number of words is quite large and so the computation for all nesting index values would be somewhat time consuming. However, the following lemma allows us to more easily compute the number of words of size 10, 11, and 12 and nesting index values 8, 9, and 10.

LEMMA 3.3 *If w and w' are double occurrence words such that $w' = w - a$ for some letter $a \in \Sigma$, then* $\mathrm{NI}(w) \leq \mathrm{NI}(w') + 1$. *In other words, by adding a letter to a double occurrence word, the nesting index is increased by at most one.*

Proof. If $\mathrm{NI}(w') = n$, let $(u_0, u_1, \ldots, u_n)$ be a reduction of w'. Then $(w, u_0, u_1, \ldots, u_n)$ is a reduction of w in which $u_0 = w' = w - a$. Thus, $\mathrm{NI}(w) \leq n + 1 = \mathrm{NI}(w') + 1$. □

In Chapter 6, we use Table 1 to formulate Conjecture 1 on the minimum number of letters needed to construct a word with nesting index $n \in \mathbb{N}$.

3.4 A study on the nesting index

Chord diagrams and circle graphs are useful tools in the study of double occurrence words, for example in [10]. In the present section we use chord diagrams and circle graphs as tools to study the nesting index of double occurrence words. The main result will be a characterization of double occurrence words that are 1-reducible. This characterization allows us to show that for arbitrary $n \geq 0$ there exists a word with nesting index n.

[1] Implemented in C code, readily available for download at http://knot.math.usf.edu/software/NI/NestIndex.zip

Table 1: Number of double occurrence words with a given size and nesting index

Size	Nesting Index									
	1	2	3	4	5	6	7	8	9	10
1	1	0	0	0	0	0	0	0	0	0
2	3	0	0	0	0	0	0	0	0	0
3	7	8	0	0	0	0	0	0	0	0
4	17	78	10	0	0	0	0	0	0	0
5	41	424	479	1	0	0	0	0	0	0
6	99	1915	6248	2133	0	0	0	0	0	0
7	239	7914	50247	69879	6856	0	0	0	0	0
8	577	31370	328810	1004642	648065	13561	0	0	0	0
9	1393	122530	1927900	10125920	17081040	5187788	12854	0	0	0
10	–	–	–	–	–	–	–	2019	0	0
11	–	–	–	–	–	–	–	–	4	0
12	–	–	–	–	–	–	–	–	–	0

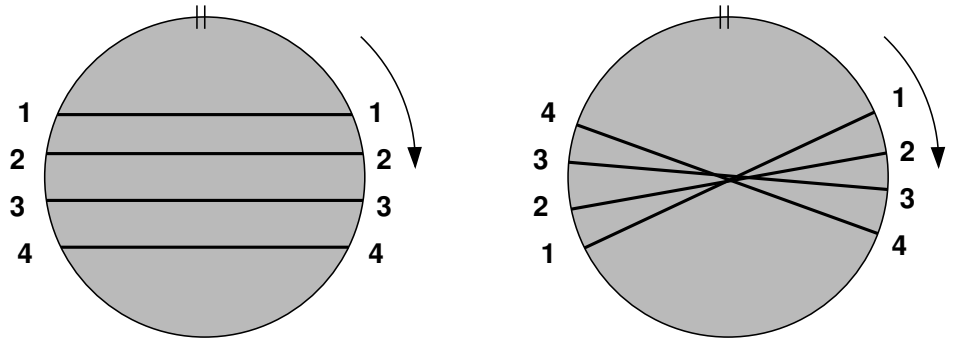
3.4.1 Nesting index and chord diagrams

Recall that a chord diagram of double occurrence word w is a circle $\mathcal{C}$ where the letters of w are placed around the circumference of $\mathcal{C}$ and for each distinct letter a in w a chord of $\mathcal{C}$ is drawn from the first occurrence of a to the second occurrence.

EXAMPLE 3.4 Figure 8(a) and Figure 8(b) are chord diagram representations of the return word 12344321 and repeat word 12341234, respectively.

REMARK 3.4 In the chord diagram of any return word no pair of chords intersects. In the chord diagram of any repeat word every pair of chords intersects.

REMARK 3.5 If w is a double occurrence word that corresponds to a chord diagram $\mathcal{C}$ and $u \sqsubseteq w$ is also a double occurrence word, then the chords in $\mathcal{C}$ associated with u have no intersection with the chords in $\mathcal{C}$ that correspond to the symbols in $w - u$.



(a) Chord diagram for the return word 1234432

(b) Chord diagram for the repeat word 12341234

Figure 8: Chord diagram representations of a repeat word and a return word

THEOREM 3.1 *Let w be a double occurrence word. Then w is 1-reducible if and only if the chord diagram of w does not contain the chord diagram $\mathcal{C}_{1\times 2}$ (Figure 9) as a sub-chord diagram.*

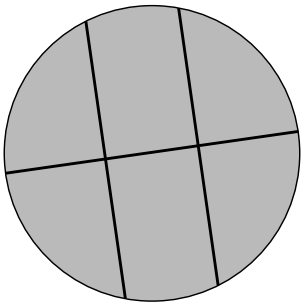


Figure 9: Chord diagram $\mathcal{C}_{1\times 2}$ associated with the double occurrence words 121323, 123213, and 123132

Proof. The proof follows by induction on the size of w. One can easily verify that all double occurrence words of size 1 and 2 are 1-reducible and their chord diagrams have less than three chords, hence, do not contain $\mathcal{C}_{1\times 2}$ as a sub-chord diagram.

Now suppose the theorem holds for w of size k where $3 \leq k < n$. For the final part of the proof we treat the right and left implications separately.

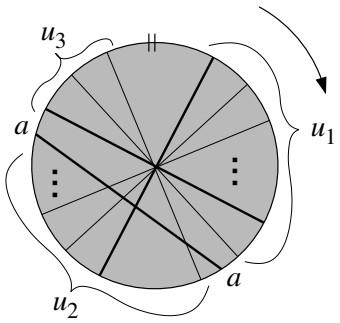
($\Rightarrow$): Let w be of size n and suppose w is 1-reducible. Let $\mathcal{M}_w$ be the set of maximal subwords of w. Then $w' = w - \mathcal{M}_w$ is 1-reducible, hence, by induction hypothesis, the chord diagram of w' does not contain $\mathcal{C}_{1\times 2}$ as a sub-chord diagram. By Remark 3.5, the chords in $\mathcal{C}$ associated with the words in $\mathcal{M}_w$, have no intersection with the chords in $\mathcal{C}$ associated with w'. Then if $\mathcal{C}_{1\times 2}$ is a sub-chord diagram of $\mathcal{C}$, $\mathcal{C}_{1\times 2}$ must be a sub-chord diagram of the chords in $\mathcal{C}$ associated with the words in $\mathcal{M}_w$. However, since a pair of chords

associated with letters in two distinct double occurrence words in $\mathcal{M}_w$ cannot intersect (Remark 3.5) and there is a chord in $\mathcal{C}_{1\times 2}$ that intersects the other two chords, then the chords of $\mathcal{C}_{1\times 2}$ can not be associated with more than one word in $\mathcal{M}_w$, and hence, $\mathcal{C}_{1\times 2}$ is a sub-chord diagram of the chords associated with a single word $u \in \mathcal{M}_w$. But this cannot be the case by Remark 3.4. Thus, $\mathcal{C}$ does not contain $\mathcal{C}_{1\times 2}$ as a sub-chord diagram and so the right implication is proved.

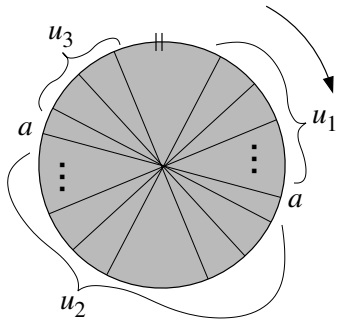
($\Leftarrow$): Let w be a word of size n and suppose $\mathcal{C}$ does not contain $\mathcal{C}_{1\times 2}$ as a sub-chord diagram. Let $a \in \Sigma$ and let $\mathcal{C}'$ denote the chord diagram of $w' = w - a$. Since $\mathcal{C}'$ does not contain $\mathcal{C}_{1\times 2}$ as a sub-chord diagram, it follows by induction hypothesis that w' is 1-reducible. Let $\mathcal{M}_{w'}$ denote the set of maximal subwords of w'.

We claim that w has a maximal subword. If for some $u \in \mathcal{M}_{w'}$, u is a subword of w, then we are done. Since a has only two occurrences in w, it follows that if $|\mathcal{M}_{w'}| \geq 3$, then there exists $u \in \mathcal{M}_{w'}$ such that $u \sqsubseteq w$ and we are done. Assume $|\mathcal{M}_{w'}| \leq 2$. If $\mathcal{M}_{w'} = \{u, v\}$ and u and v are not subwords of w, then we can write $u = u_1u_2$ and $v = v_1v_2$ such that u_1au_2 and v_1av_2 are subwords of w. Since u and v are not subwords of w, we have that u_1, u_2, v_1, and v_2 are non-empty. Since u_1, u_2, v_1, and v_2 are non-empty, it follows that they cannot be double occurrence words (Remark 3.3), hence, the chord for a intersects a chord from u and a chord from v. Since the chords from u and v do not intersect by Remark 3.5, it follows that $\mathcal{C}_{1\times 2}$ is a sub-chord diagram of $\mathcal{C}$ which is a contradiction. Lastly, we consider $\mathcal{M}_{w'} = \{u\}$ in which u is not a subword of w. Let us write $u = u_1u_2u_3$ so that $u' = u_1au_2au_3$ is a subword of w. If u_2 is empty, then aa is a subword of w which is maximal or contained in a maximal subword of w. Assume u_2 is non-empty. If u is a repeat word, then the chord for a must intersect all chords from u, else, $\mathcal{C}_{1\times 2}$ is a sub-chord diagram of $\mathcal{C}$ (Figure 10(a)). Since all of the chords of u' intersect, then the word is a maximal repeat word in w (Figure 10(b)). Now assume $u = a_1a_2\cdots a_na_n\cdots a_2a_1$ is a return word. Then the chord of a can intersect at most one chord from u, else, $\mathcal{C}_{1\times 2}$ is a sub-chord diagram of $\mathcal{C}$ (Figure 11(a)). Suppose a intersects a chord, say with label a_i. If $i = n$, then aa_naa_n or a_naa_na is a maximal repeat word in w. If $i \neq n$, then $a_{i+1}a_{i+2}\cdots a_na_n\cdots a_{i+2}a_{i+1}$ is a maximal return word in w. Otherwise, assume a intersects no chords from u. Then u' is a maximal return word of w (Figure 11(b)).

By the above claim, we can apply reduction operation 1 to w to obtain a word w' of size $< n$. Since $\mathcal{C}$ does not contain $\mathcal{C}_{1\times 2}$ as a sub-chord diagram, the chord diagram of w' also does not contain $\mathcal{C}_{1\times 2}$. By induction hypothesis, w' is 1-reducible. Thus, w is 1-reducible. □

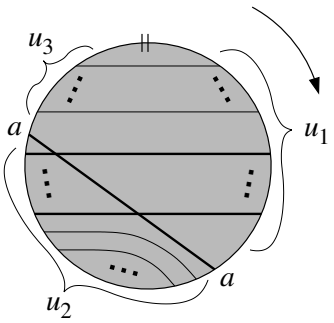


(a) If a does not intersect all chords from u, then $\mathcal{C}_{1\times 2}$ (bold) is a sub-chord diagram of $\mathcal{C}$

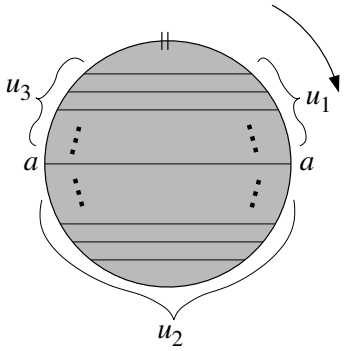


(b) If a intersects all chords from u, then $u' = u_1au_2au_3$ is a maximal repeat word in w.

Figure 10: If u is a repeat word.



(a) If a intersects more than one chord from u, then $\mathcal{C}_{1\times 2}$ (bold) is a sub-chord diagram of $\mathcal{C}$.



(b) If a does not intersect any chords from u, then $u' = u_1au_2au_3$ is a maximal return word in w.

Figure 11: If u is a return word.

The preceding theorem tells us that if $\mathcal{C}_{1\times 2}$ is a sub-chord diagram of $\mathcal{C}$ which corresponds to a double occurrence word w, then in any reduction of w at some point we are forced to apply reduction operation 2. What it does not tell us is how many times we must apply reduction operation 2. The following lemma and theorem aim to do just that.

LEMMA 3.4 *Let w be a double occurrence word with chord diagram $\mathcal{C}$ and let w' be the word obtained from w by application of reduction operation 1 with chord diagram $\mathcal{C}'$. If $\mathcal{C}_{1\times 2}$ is a sub-chord diagram of $\mathcal{C}$ where b is a chord in $\mathcal{C}_{1\times 2}$, then b is also a chord in $\mathcal{C}'$.*

Proof. Assume to the contrary that b is not a chord in $\mathcal{C}'$. Then b must belong to some maximal subword u of w. Since b is a chord in $\mathcal{C}_{1\times 2}$, b either intersects the other two chords in $\mathcal{C}_{1\times 2}$, or b intersects another chord in $\mathcal{C}_{1\times 2}$ which intersects the third chord in $\mathcal{C}_{1\times 2}$. Then by Remark 3.5, since u is a double occurrence word, we have that the three letters that correspond to the chords in $\mathcal{C}_{1\times 2}$ are letters in u, hence, $\mathcal{C}_{1\times 2}$ is a sub-chord diagram of the chords that correspond to u. However, since u is a repeat word or a return word, then by Remark 3.4, this cannot be the case. This gives a contradiction. □

THEOREM 3.2 *Let w be a double occurrence word with corresponding chord diagram $\mathcal{C}$ and let $2 \leq m \leq n$ be integers. If $\mathcal{C}$ contains the chord diagram $\mathcal{C}_{m\times n}$ (Figure 12) as a sub-chord diagram, then* $\mathrm{NI}(w) \geq m+1$.

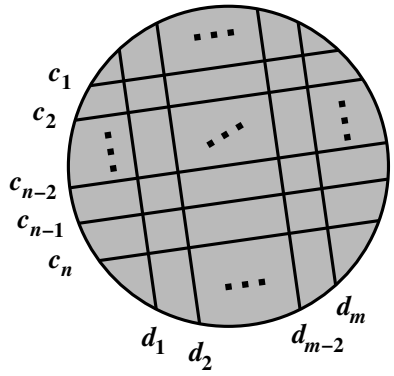


Figure 12: Chord diagram $\mathcal{C}_{m\times n}$

Proof. Note that each chord in $\mathcal{C}_{m\times n}$ is a chord in some $\mathcal{C}_{1\times 2}$ as a sub-chord diagram of $\mathcal{C}_{m\times n}$, hence, as a sub-chord diagram of $\mathcal{C}$. Then by Lemma 3.4, if we apply reduction operation 1 some number of times to w to obtain w', then $\mathcal{C}_{m\times n}$ remains a sub-chord diagram of the chord diagram of w'. Then we must apply reduction operation 2 to remove any letter from w corresponding to some chord in $\mathcal{C}_{m\times n}$. Further, note that if we remove a chord from $\mathcal{C}_{m\times n}$ by removing the corresponding letter with reduction operation 2, then every chord in the resulting chord diagram $\mathcal{C}'_{m\times n}$ is also a chord in some $\mathcal{C}_{1\times 2}$ as a sub-chord diagram of $\mathcal{C}'_{m\times n}$. Hence, by Lemma 3.4, we are required to apply reduction operation 2 again. This necessity of applying reduction operation 2 continues until one of the following occurs.

(i) The letters that correspond to the chords $c_1, \ldots, c_n$ have all been removed by n applications of reduction operation 2,

(ii) the letters that correspond to the chords $d_1 \ldots, d_m$ have all been removed by m applications of reduction operation 2, or

(iii) the letters that correspond to $m-1$ of the chords d_i and $n-1$ of the chords c_j have all been removed by $m+n-2$ applications of reduction operation 2.

Since $m \leq n \leq m+n-2$, it follows that we must apply reduction operation 2 a minimum of m times for any reduction of w. This gives $\mathrm{NI}(w) \geq m$. Now since there are still chords left over from $\mathcal{C}_{m\times n}$, we see that w has not yet been reduced to the empty word and so at least one additional reduction operation is necessary to complete a reduction of w. Thus, $\mathrm{NI}(w) \geq m+1$. □

COROLLARY 3.2 *For all $n \in \mathbb{N}$, there exists a double occurrence word w with $\mathrm{NI}(w) = n$.*

Proof. We have $\mathrm{NI}(11) = 1$, $\mathrm{NI}(123231) = 2$ and for $n \geq 3$, we can take w to be a double occurrence word corresponding to the chord diagram $\mathcal{C}_{(n-1)\times(n-1)}$ so that, by Theorem 3.2, $\mathrm{NI}(w) = n$. □

We now introduce some notions to rephrase the characterization of 1-reducible double occurrence words in terms of its subwords.

DEFINITION 3.10 If $w = a_1 a_2 \cdots a_n$ and $u = a_{i_1} a_{i_2} \cdots a_{i_k}$ such that $i_1, i_2, \ldots, i_k \in \{1, 2, \ldots, n\}$ and $i_1 < i_2 < \cdots < i_k$, then we say that u is a *sparse subword* of w. If w' is a double occurrence word and there exists a sparse subword u of w such that $w' = u^{asc}$, then we say that w' is *inherent in* w.

COROLLARY 3.3 *Let w be a double occurrence word. Then w is 1-reducible if and only if neither 123213, 123132, nor 121323 is inherent in w.*

Proof. Since the words 123213, 123132, and 121323 correspond to the chord diagram $\mathcal{C}_{1\times 2}$ in Figure 9, it follows that one of the words is inherent in w if and only if $\mathcal{C}_{1\times 2}$ is a sub-chord diagram of the chord diagram for w. Then by Theorem 3.1, the result follows. □

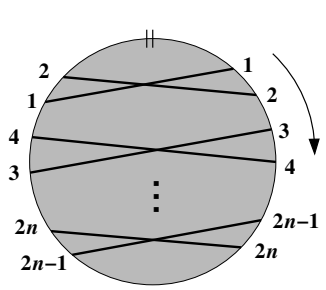
3.4.2 Nesting index and circle graphs

In the previous subsection we found some interesting relationships between the nesting index of a word and the chord diagram of that word. This prompts the question whether any relationships can be found between the nesting index of a double occurrence word and its circle graph. The following observations, although not a resounding "no" to the question, do show that the nesting index is not an invariant of circle graphs.

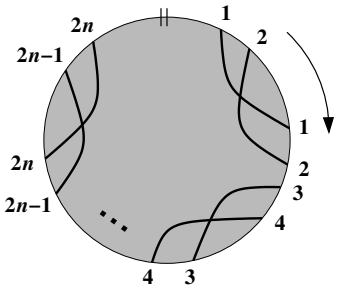
Let us consider the words w_1 and w_2 of size $2n$ that have the following form

$$w_1 = 1234\cdots(2n-1)(2n)(2n-1)(2n)\cdots 3421,$$
$$w_2 = 12123434\cdots(2n-1)(2n)(2n-1)(2n).$$

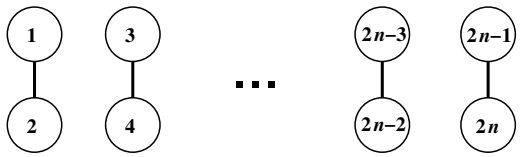
One can easily verify that for arbitrary $n \geq 1$, we have $\mathrm{NI}(w_1) = n$ and $\mathrm{NI}(w_2) = 1$. Also, Figure 13 shows that the two words correspond to the same circle graph. Then for arbitrary $n \geq 1$, w_1 and w_2 are words of size $2n$ that correspond to the same circle graph and whose nesting index values differ by $n-1$.



(a) Chord diagram of w_1



(b) Chord diagram of w_2



(c) Circle graph of w_1 and w_2

Figure 13: Two words that correspond to the same circle graph with arbitrarily large differences in nesting index values

Chapter 4

Genus Range and Genus Spectrum

In this chapter we discuss the genus range property of assembly graphs and we generalize a result in [5] on how the genus range is affected by connecting two assembly graphs. We then consider a more general property called the genus spectrum of an assembly graph. Lastly, we discuss the genus spectrum in even more generality for double occurrence words.

4.1 Orientable genus range for assembly graphs

For this chapter an assembly graph will be assumed to be without endpoints, unless otherwise stated. We will primarily be concerned with the genus of surfaces into which assembly graphs are cellularly embedded.

DEFINITION 4.1 An *embedding of an assembly graph* Γ into a surface is an embedding such that the cyclic order of the edges around each vertex in Γ agrees with cyclic order of the embedded images of those edges. Such an embedding is called *cellular* if each component of the complement of the graph in the surface is an open disk.

DEFINITION 4.2 The *genus range* of an assembly graph Γ, denoted by $\mathrm{gr}(\Gamma)$, is the defined to be the set of all integers g such that F is a surface of genus g into which Γ cellularly embeds.

In [5] one of the main problems was to characterize the sets of integers that were realized as the genus range of some assembly graph on a given number of vertices. The authors in [5] showed that the genus range for a given assembly graph is always a set of consecutive integers. As such, we will often represent the genus range by $[m, n] = \{m, m + 1, \ldots, n\}$ where $0 \leq m \leq n$ are integers.

The computation of the genera of surfaces into which an assembly graph cellularly embeds relies heavily on a construction by Scott Carter [8] which we will call a ribbon graph.

DEFINITION 4.3 A *ribbon graph* is a surface into which an assembly graph Γ cellularly embeds and is obtained in the following way: associate a square for each vertex v in Γ so that the edges incident to v coincide with the coordinate axes of the square; further, for each edge e in Γ, if e is incident to vertices u and u', then we join the sides of the squares of u and u' that correspond to e with a band. Figure 14 depicts the process of constructing a ribbon graph for $\overline{\Gamma}(1212)$.

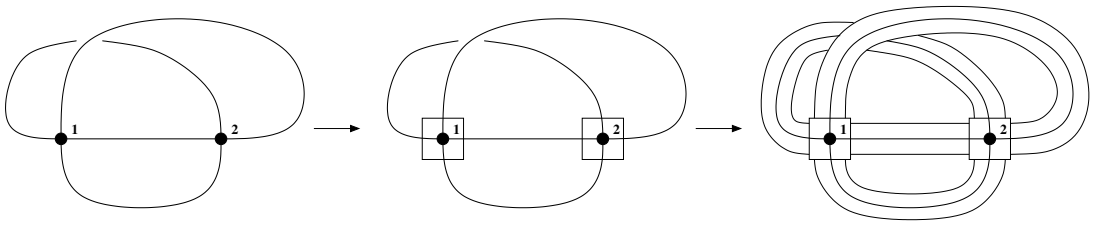


Figure 14: Ribbon graph construction for 1212

On the one hand the ribbon graph construction is a compact orientable surface with boundary and as such, for a given ribbon graph F, we have a formula relating its Euler characteristic $\chi(F)$, its genus $g(F)$ and its number of boundary components $b(F)$: $\chi(F) = 2 - 2g(F) - b(F)$. On the other hand, the ribbon graph F is homotopy equivalent to an assembly graph Γ which as a 1-complex with n vertices and $2n$ edges has Euler characteristic $\chi(F) = \chi(\Gamma) = n - 2n = -n$. These observations give the following formula for evaluating the genus of ribbon graphs which we state as a remark so that we may refer back to it throughout the chapter.

REMARK 4.1 Let Γ be an assembly graph on n vertices and let F be a ribbon graph constructed from Γ as described in Definition 4.3. Then letting $g(F)$ and $b(F)$ denote the genus and number of boundary components of F, respectively, we have $g(F) = \frac{1}{2}(n - b(F) + 2)$.

By convention, when constructing an assembly graph from a double occurrence word w, at the first occurrence of a given letter in w we draw the edges corresponding to this part of the transverse path from west to east through the vertex. At the second occurrence of a given letter in w, we have a choice to draw the corresponding edges of the transverse path from north to south through the vertex or south to north through the vertex. Since the cyclic ordering of the edges incident to the vertex does not depend on this choice, the two graphs obtained from making different choices at the vertex are isomorphic as assembly graphs. What may change, however, is the resulting ribbon graph construction of the assembly graph.

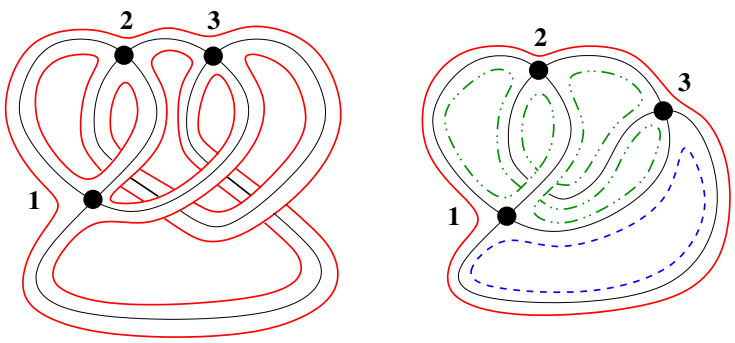


Figure 15: Different ribbon graphs of $\overline{\Gamma}(121323)$ obtained by different choices of entering the vertex 3 for the second time

DEFINITION 4.4 The operation of changing the ribbon graph at a vertex v from Figure 16(a) to Figure 16(b) or vice-versa is called a *connection change* at v.

In Figure 16(a), observe that the arrows on the boundary components on the opposite sides of each edge go in opposite directions, indicating that the ribbon graph is orientable. Note that changing the connection at the vertex v does not change the orientability of the surface.

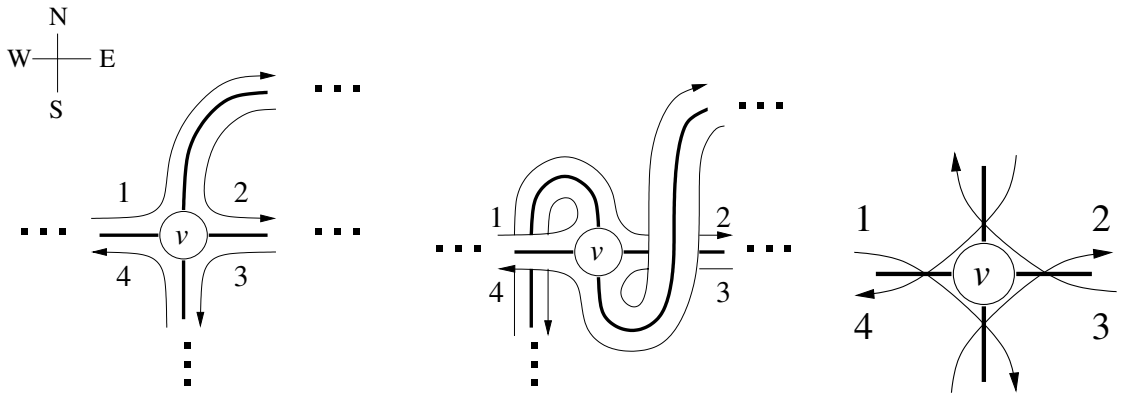
REMARK 4.2 For an assembly graph Γ on n vertices with ribbon graph F, one obtains the genus range of Γ by computing the number of boundary components for each ribbon graph F' obtained from F by changing connections at vertices of Γ. Then there are 2^n possible ribbon graphs that can be constructed for Γ.

DEFINITION 4.5 Let F be a ribbon graph of an assembly graph Γ and let e be an edge in Γ. Then e is said to be *traced by the boundary component* δ in F if the boundary of the ribbon that contains e is a portion of δ. Note that every edge in a ribbon graph is traced by either one or two (distinct) boundary components. As an example, consider the ribbon graphs in Figure 15. In the ribbon graph on the left, every edge is traced by a single boundary; in the ribbon graph on the right, both edges between the vertices 1 and 3 are traced by two distinct boundary components.

From the proof of Lemma 3.2 in [5] we may deduce the following.

LEMMA 4.1 *Let Γ be an assembly graph. For a given edge e in Γ, there exists a ribbon graph F of Γ such that e is traced by two distinct boundary components.*

We will use the following remark in several results on the genus range and its generalizations.



(a) Ribbon graph at vertex v where edges corresponding to second occurrence of v go north to south through v

(b) Ribbon graph at vertex v where edges corresponding to second occurrence of v go south to north through v

(c) Schematic drawing of how the boundary components of the ribbon graph change from (a) to (b)

Figure 16: Changing the connection at a vertex v

REMARK 4.3 Let Γ_1 and Γ_2 be assembly graphs and let Γ be the assembly graph obtained by connecting Γ_1 and Γ_2 through edges e_1 and e_2 as depicted in Figure 17 with some chosen orientations of Γ_1 and Γ_2. Note that the connections at the vertices in Γ_1 and Γ_2 determine their respective ribbon graph constructions F_1 and F_2. Then connecting Γ_1 and Γ_2 through edges e_1 and e_2 without changing the connection at any of their vertices, we produce unique connections at the vertices in Γ and hence, determine the ribbon graph construction F of Γ. Most importantly, all ribbon graphs F of Γ are realized by connecting Γ_1 and Γ_2 through edges e_1 and e_2 by considering different possible connections at the vertices of Γ_1 and Γ_2.

The following theorem is a generalization of Lemma 2.8 in [5] wherein they considered the assembly graph Γ' obtained by connecting an assembly graph Γ with the graph $\overline{\Gamma}(1212)$. Recall that the definition of connecting assembly graphs Γ_1 and Γ_2 relies on choosing orientations of Γ_1 and Γ_2. However, note that the proof of the Theorem 4.1 holds regardless of how we choose orientations of Γ_1 and Γ_2 and hence, we may consider orientations to be chosen arbitrarily.

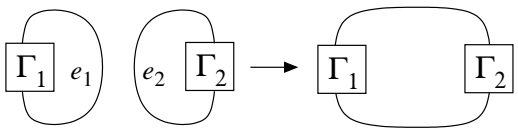


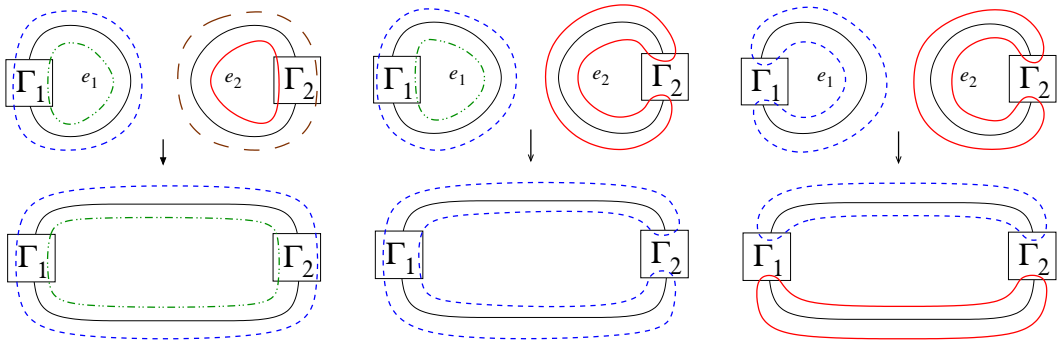
Figure 17: Connecting the graphs Γ_1 and Γ_2 through edges e_1 and e_2

THEOREM 4.1 *Let Γ_1 and Γ_2 be assembly graphs and let Γ be the graph obtained by connecting Γ_1 and Γ_2 through edges e_1 and e_2 as depicted in Figure 17. Suppose $\mathrm{gr}(\Gamma_i) = [m_i, n_i]$ for $i = 1, 2$.*

(i) If there exists $i \in \{1, 2\}$ such that for all ribbon graphs of Γ_i the edge e_i is traced by two distinct boundary components, then $\mathrm{gr}(\Gamma) = [m_1 + m_2, n_1 + n_2]$.

(ii) Otherwise, $\mathrm{gr}(G) = [m_1 + m_2 - k, n_1 + n_2 - \ell]$ for some $k, \ell \in \{0, 1\}$.

Proof. Let v_1, v_2, and v denote the number of vertices of Γ_1, Γ_2, and Γ, respectively, and note that $v = v_1 + v_2$. Figure 18 shows some of the possibilities of the boundary components tracing e_1 and e_2 (top) and the resulting ribbon graph F of Γ after connecting Γ_1 and Γ_2 (bottom). The only possible situation not depicted in Figure 18 is e_2 traced by two distinct boundary components and e_1 traced by one boundary component, however, this is symmetric to the situation in Figure 18(b) and we shall not consider this case. From Remark 4.3, by considering only these ribbon graphs that result from connecting Γ_1 and Γ_2 we realize all of $\mathrm{gr}(\Gamma)$.



(a) Boundary curves tracing e_1 and e_2 belong to distinct boundary components

(b) Boundary curves tracing e_1 belong to distinct boundary components and boundary components tracing e_2 belong to the same boundary component

(c) Boundary curves tracing e_1 and e_2 belong to the same boundary component

Figure 18: Boundary components before and after connecting graphs Γ_1 and Γ_2

Let $g_1 \in \mathrm{gr}(\Gamma_1)$ and $g_2 \in \mathrm{gr}(\Gamma_2)$. Then there exists ribbon graphs F_1 and F_2 of Γ_1 and Γ_2, respectively, such that $g_i = g(F_i)$ for $i = 1, 2$.

(i) Without loss of generality, we may assume that e_1 in Γ_1 is traced by two distinct boundary components in every ribbon graph of Γ_1. Then depending on the ribbon graph F_2 our situation is that of Figure 18(a)

or Figure 18(b). In both situations, we have $b(F) = b(F_1) + b(F_2) - 2$ and thus, by Remark 4.1,

$$g(F) = \frac{1}{2}(v - b(F) + 2) = \frac{1}{2}(v_1 + v_2 - (b(F_1) + b(F_2) - 2) + 2)$$
$$= \frac{1}{2}(v_1 - b(F_1) + 2) + \frac{1}{2}(v_2 - b(F_2) + 2) = g_1 + g_2.$$

This implies $\mathrm{gr}(\Gamma) = [m_1 + m_2, n_1 + n_2]$.

(ii) By Lemma 4.1, e_1 and e_2 are not traced by a single boundary component in all ribbon graphs of Γ_1 and Γ_2, respectively. Then all three situations in Figure 18 are possible. The situations in Figure 18(a) and Figure 18(b) were considered in (i). For the situation in Figure 18(c), we have $b(F) = b(F_1) + b(F_2)$ and by Remark 4.1 we have

$$g(F) = \frac{1}{2}(v - b(F) + 2) = \frac{1}{2}(v_1 + v_2 - (b(F_1) + b(F_2)) + 2)$$
$$= \frac{1}{2}(v_1 - b(F_1) + 2) + \frac{1}{2}(v_2 - b(F_2) + 2) - 1 = g_1 + g_2 - 1.$$

It follows that $\mathrm{gr}(\Gamma) = [m_1 + m_2 - k, m_1 + m_2 - \ell]$ for some $k, \ell \in \{0, 1\}$. Moreover, note that $k = 1$ if and only if for each $i \in \{1, 2\}$, there exists a ribbon graph F_i of Γ_i such that $g(F_i) = \min(\mathrm{gr}(\Gamma_i))$ and e_i is traced by a single boundary component in F_i. Similarly, $\ell = 0$ if and only if there exists $i \in \{1, 2\}$ such that F_i is a ribbon graph of Γ_i satisfying $g(F_i) = \max(\mathrm{gr}(\Gamma))$ and e_i is traced by two distinct boundary components in F_i.

□

COROLLARY 4.1 *Connecting two assembly graphs as in Figure 17 does not decrease the size of the genus range; that is, in terms of Theorem 4.1,* $|\mathrm{gr}(\Gamma)| \geq |\mathrm{gr}(\Gamma_i)|$ *for* $i = 1, 2$.

4.2 Genus spectrum for assembly graphs

Now we generalize the genus range by introducing a property of assembly graphs called the "genus spectrum."

DEFINITION 4.6 The *genus frequency* of an assembly graph Γ at $g \in \mathbb{N}$, denoted by $\mathrm{gf}(\Gamma, g)$ is the number of possible ribbon graph constructions F of Γ where $g(F) = g$. The *genus spectrum* of Γ, denoted by $\mathrm{gs}(\Gamma)$, is the set of pairs $(g, \mathrm{gf}(\Gamma, g))$ for all $g \in \mathrm{gr}(\Gamma)$.

By Remark 4.2, the number of possible ribbon graph constructions of an assembly graph Γ on n vertices is 2^n. This implies the following.

PROPOSITION 4.1 Let Γ be an assembly graph on n vertices. Then

$$\sum_{g \in \mathrm{gr}(\Gamma)} \mathrm{gf}(\Gamma, g) = 2^n.$$

The authors in [5] introduced a "cross sum" for assembly graphs and showed what effect the cross sum had on the genus range (Lemma 2.6). Here we prove an analogous result for the genus spectrum. The proof here is roughly the same as in Lemma 2.6 in [5] with some additional arguments to generalize the result for the genus spectrum.

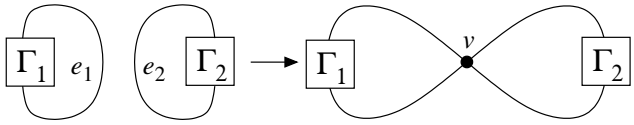


Figure 19: Cross sum of Γ_1 and Γ_2

DEFINITION 4.7 Let Γ_1 and Γ_2 be assembly graphs with edges e_1 and e_2, respectively. Then an assembly graph Γ is said to be obtained from Γ_1 and Γ_2 by *cross sum* through edges e_1 and e_2 if it is formed by connecting the two graphs to the figure-eight graph as we see in Figure 19. The vertex v is called the *figure-eight vertex*.

LEMMA 4.2 *Let Γ_1 and Γ_2 be assembly graphs and let Γ be the graph obtained from Γ_1 and Γ_2 by cross sum. Then*

$$\mathrm{gf}(G, g) = \sum_{g = g_1 + g_2} 2 \cdot \mathrm{gf}(\Gamma_1, g_1) \cdot \mathrm{gf}(\Gamma_2, g_2).$$

Proof. Let v, v_1, and v_2 denote the number of vertices of Γ, Γ_1, and Γ_2, respectively. Then $v = v_1 + v_2 + 1$.

Let F_1 and F_2 be ribbon graphs of Γ_1 and Γ_2, respectively, where $g(F_1) = g_1$ and $g(F_2) = g_2$. Figures 20(a)-(c) depict possibilities for the number of boundary components tracing e_1 and e_2 in F_1 and F_2, respectively, omitting the case that is symmetric to Figure 20(b), just as in the proof of Theorem 4.1. Then constructing the cross sum Γ without changing connections at any of the vertices in Γ_1 or Γ_2 will give a distinct connection at the vertices in Γ that determines some ribbon graph F of Γ. Figures 20(d)-(f) depict

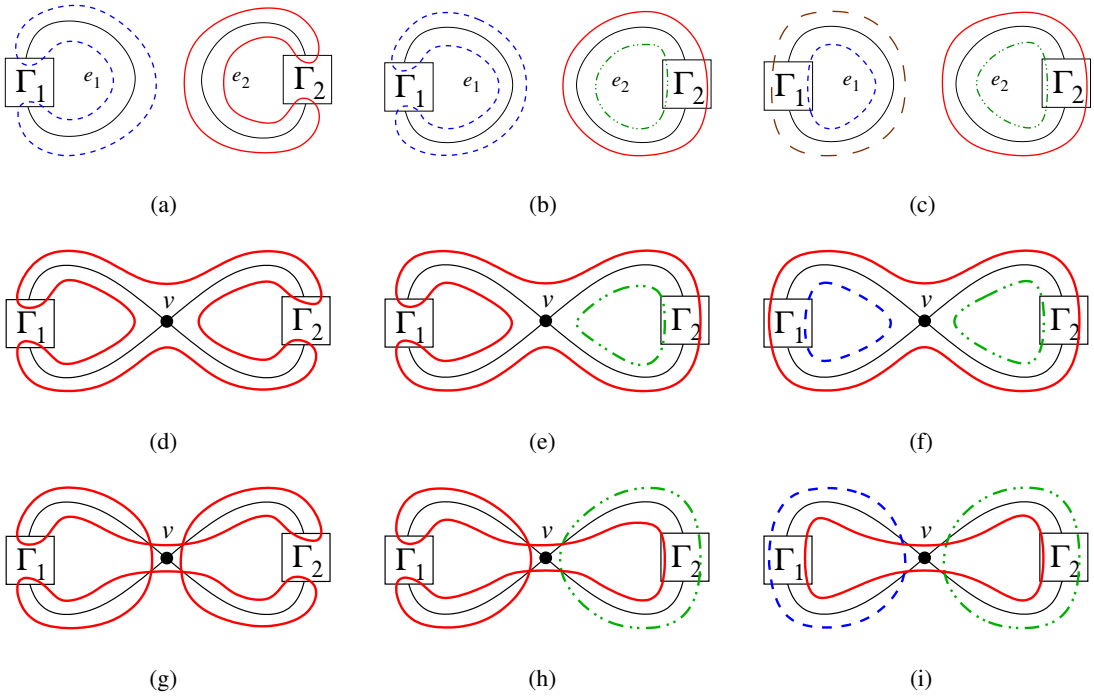


Figure 20: Possible ribbon graphs for Γ.

all possibilities for boundary components around the figure-eight vertex v and Figures 20(g)-(i) show the boundary components corresponding to ribbon graphs in Figures 20(a)-(c), respectively, after changing the connection at v. In any case, we have $b(F) = b(F_1) + b(F_2) - 1$. Thus, by Remark 4.1,

$$g(F) = \frac{1}{2}(v - b(F) + 2) = \frac{1}{2}((v_1 + v_2 + 1) - (b(F_1) + b(F_2) - 1) + 2) = g(F_1) + g(F_2).$$

Now for any pair (g_1, g_2) satisfying $g_1 \in \mathrm{gr}(\Gamma_1)$, $g_2 \in \mathrm{gr}(\Gamma_2)$, and $g_1 + g_2 = g$, note that there are $2 \cdot \mathrm{gf}(\Gamma_1, g_1) \cdot \mathrm{gf}(\Gamma_2, g_2)$ possible connections of Γ which determine a ribbon graph F of Γ with $g(F) = g$. Indeed, we count by rule of product: there are $\mathrm{gf}(\Gamma_1, g_1)$ connections of Γ_1 that give a ribbon graph F_1 with $g(F_1) = g_1$, there are $\mathrm{gf}(\Gamma_2, g_2)$ connections of Γ_2 that give a ribbon graph F_2 with $g(F_2) = g_2$, and, there are two possible connections at the figure-eight vertex v. Since each F_1 and F_2 produces a distinct ribbon graph F of the cross-sum Γ with $g(F) = g$, the claim follows. Now by summing over all pairs (g_1, g_2) satisfying $g_1 + g_2 = g$, we obtain our result. □

As a corollary we get Lemma 2.6 from [5].

COROLLARY 4.2 *Let Γ_1 and Γ_2 be assembly graphs. If Γ is obtained from Γ_1 and Γ_2 by cross sum, then* $\mathrm{gr}(\Gamma) = \{g_1 + g_2 : g_1 \in \mathrm{gr}(\Gamma_1), g_2 \in \mathrm{gr}(\Gamma_2)\}$.

Proof. Let $g \in \mathrm{gr}(\Gamma)$. Then $\mathrm{gf}(\Gamma, g) \neq 0$, hence, by Lemma 4.2 there exists g_1, g_2 such that $g_1 + g_2 = g$, $\mathrm{gf}(\Gamma_1, g_1) \neq 0$ and $\mathrm{gf}(\Gamma_2, g_2) \neq 0$. This implies $g_1 \in \mathrm{gr}(\Gamma_1)$ and $g_2 \in \mathrm{gr}(\Gamma_2)$. Conversely, suppose $g_1 \in \mathrm{gr}(\Gamma_1)$ and $g_2 \in \mathrm{gr}(\Gamma_2)$. Then $\mathrm{gf}(\Gamma_1, g_1) \neq 0$ and $\mathrm{gf}(\Gamma_2, g_2) \neq 0$ and by Lemma 4.2 we have $\mathrm{gf}(\Gamma, g_1 + g_2) \neq 0$, hence, $g_1 + g_2 \in \Gamma$. □

The following is a special case of Lemma 4.2.

COROLLARY 4.3 *Let w be a double occurrence word and set $\Gamma = \overline{\Gamma}(w)$ and $\Gamma' = \overline{\Gamma}(waa)$ where a is a letter that is not in w. Then $\mathrm{gr}(\Gamma') = \mathrm{gr}(\Gamma)$. Moreover,*

$$\mathrm{gs}(\Gamma') = \{(g, 2 \cdot \mathrm{gf}(\Gamma, g)) : g \in \mathrm{gr}(\Gamma)\}.$$

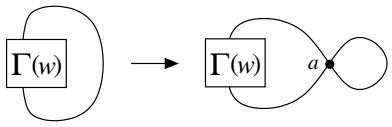


Figure 21: Replacing an edge by a loop to obtain Γ'

Proof. The operation that transforms Γ into Γ' can be thought of as replacing an edge of Γ by a loop as is depicted in Figure 21. Then putting $\Gamma_1 = \Gamma$ and replacing Γ_2 by an edge in Figures 20(e),(f),(h), and (i) we have all possible ribbon graphs of Γ'. In any case, the ribbon graphs of Γ' in comparison to Γ have an additional vertex and an additional boundary component, hence, $\mathrm{gr}(\Gamma') = \mathrm{gr}(\Gamma)$. The result on the genus spectrum of Γ' follows from arguments similar to those in the proof of Lemma 4.2. □

DEFINITION 4.8 Let w and w' be double occurrence words. We call w' a *loop nesting* of w if there exists a sequence of words $w = w_0, w_1, \ldots, w_n = w'$ such that w_i is a cyclic permutation of $w_{i-1}a_ia_i$ for some letter a_i not in w_{i-1}, for all $1 \leq i \leq n$.

The following corollary is a result of repeated application of Corollary 4.3.

COROLLARY 4.4 *Let w' be a loop nesting of w and set $\Gamma = \overline{\Gamma}(w)$ and $\Gamma' = \overline{\Gamma}(w')$. If the sizes of w and w' are m and n respectively, then $\mathrm{gs}(\Gamma') = \{(g, 2^{n-m} \cdot \mathrm{gf}(\Gamma, g)) : g \in \mathrm{gr}(\Gamma)\}$.*

4.3 Generalized genus spectrum for double occurrence words

Now we extend the definition of the genus spectrum to double occurrence words.

DEFINITION 4.9 Let w be a double occurrence word. The *generalized genus frequency for 1 or 2 boundary components* of w at genus g, denoted by $\mathrm{gf}_i(w, g)$ for $i = 1, 2$, respectively, is the number of possible ribbon graph constructions of $\overline{\Gamma}(w)$ where the closure edge of $\Gamma(w)$ is traced by one boundary component or two distinct boundary components, respectively. The *generalized genus spectrum* of w, denoted by $\mathrm{gs}(w)$, is the set of triples $(g, \mathrm{gf}_1(w, g), \mathrm{gf}_2(w, g))$ for all $g \in \mathrm{gr}(\overline{\Gamma}(w))$.

Note that a double occurrence word w and its reverse w^R have a isomorphic assembly graphs with corresponding closure edges and hence, the generalized genus spectrum is invariant with respect to reverse equivalent double occurrence words. Then because of the correspondence between reverse equivalent double occurrence words (Lemma 3.8 in [2]) and assembly graphs with endpoints, the above definition may be defined similarly on assembly graphs with endpoints so that $\mathrm{gf}_i(\Gamma, g) = \mathrm{gf}_i(w, g)$ whenever $\Gamma = \Gamma(w)$ for $i = 1, 2$.

Now we consider the generalized genus spectrums for repeat and return words.

LEMMA 4.3 *For every double occurrence word w, we have $\mathrm{gf}_1(w, 0) = 0$.*

Proof. Assume to the contrary that w is a double occurrence word satisfying $\mathrm{gf}_1(w, 0) \neq 0$. Then there is a ribbon graph of $\overline{\Gamma}(w)$ with genus 0 where the closure edge of $\Gamma(w)$ is traced by a single boundary component. Now let Γ be the graph obtained by joining two copies of $\overline{\Gamma}(w)$ through its closure edges. Then by Theorem 4.1, we have $-1 \in \mathrm{gr}(\Gamma)$. This is a contradiction as the genus is always non-negative. □

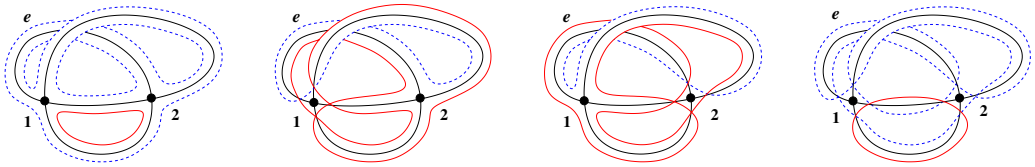
THEOREM 4.2 *If w is a return word on n letters, then*

$$\mathrm{gs}(w) = \{(0, 0, 2^n)\}.$$

If w is a repeat word on $n > 1$ letters, then

$$\mathrm{gs}(w) = \begin{cases} \{(0, 0, 2), (1, 0, 2^n - 2)\} & \text{if } n \text{ is odd,} \\ \{(1, 2, 2^n - 2)\} & \text{if } n \text{ is even.} \end{cases}$$

Proof. (Return Words): Note that $\overline{\Gamma}(aa)$ has a single vertex a and the two ribbon graphs of $\overline{\Gamma}(aa)$ each have three boundary components, hence, $\mathrm{gs}(\overline{\Gamma}(w)) = \{(0,2)\}$. Furthermore, for any return word w on n letters, note that w is a loop nesting of aa. Consequently, by Corollary 4.3, we have $\mathrm{gs}(\overline{\Gamma}(w)) = \{(0,2^n)\}$. Since $\mathrm{gf}_1(w,0) = 0$ by Lemma 4.3, the result on the genus spectrum of return words follows.



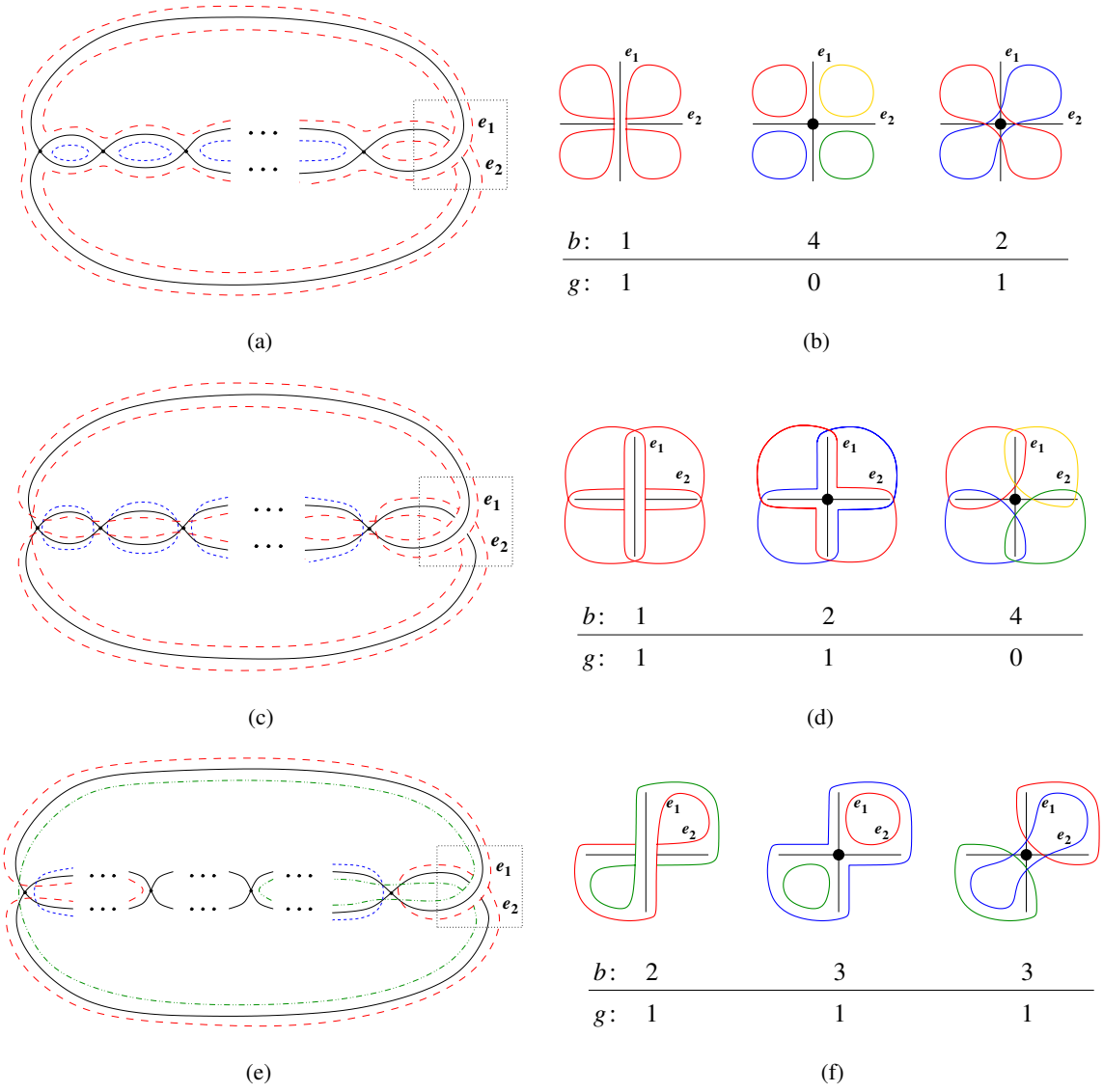
(a) No connections changed (b) Connection changed at vertex 1 (c) Connection changed at vertex 2 (d) Connection changed at both vertices

Figure 22: Ribbon graphs of $\overline{\Gamma}(1212)$

(Repeat Words): The proof is by induction on the size n of the repeat word w. When $n = 2$, $w = 1212$ and we consider all possible ribbon graphs of $\overline{\Gamma}(w)$ in Figure 22. In each ribbon graph there are 2 boundary components, hence, each has genus 1. In Figures 22(a) and 22(d) the closure edge of $\Gamma(w)$, labeled e, is traced by a single boundary component and in 22(b) and 22(c) e is traced by distinct boundary components. Thus, we have $\mathrm{gs}(1212) = \{(1,2,2)\}$ as a base case for induction.

Now suppose the result holds for repeat words of size up to $n-1$. Let w and w' be the return words of size $n-1$ and n, respectively. To prove that the theorem holds for n, we consider the ribbon graphs of $\overline{\Gamma}(w)$ and by adding a vertex to $\overline{\Gamma}(w)$ we obtain the ribbon graphs of $\overline{\Gamma}(w')$. We consider cases on the parity of $n-1$.

(i) Let $n-1$ be even. In Figures 23(a), 23(c), and 23(e), we consider ribbon graphs of $\overline{\Gamma}(w)$ where the connections are changed at none of the vertices, the connections are changed at all of the vertices, and the connections are changed at some but not all of the vertices, respectively. The dotted square in each of these figures is the location where we plan to add a vertex in order to obtain a ribbon graph of $\overline{\Gamma}(w')$. In Figures 23(b), 23(d), and 23(f) we depict for each of the respective cases in Figures 23(a), 23(c), 23(e), the global connections of the boundaries that trace the edges e_1 and e_2 before adding the vertex (left), after adding the vertex (middle), and after changing the connection at that vertex (right). For each global connection, we see the number of boundary components b tracing the edges e_1 and e_2 and the genus g of the ribbon graph; before adding the vertex g is given by the induction hypothesis and then is calculated by how b changes as we add a vertex or change the connection at that vertex.



b:	1	4	2
g:	1	0	1

b:	1	2	4
g:	1	1	0

b:	2	3	3
g:	1	1	1

Figure 23: Case (i): $n-1$ is even

Note that the closure edge of $\Gamma(w')$ is e_1 and in each global connection in Figure 23 after adding a vertex, e_1 is traced by distinct boundary components. Then $\mathrm{gf}_1(w, g) = 0$ for all $g \in \mathrm{gr}(\overline{\Gamma}(w))$. Also, there are exactly two cases where $g = 0$, namely, in Figures 23(b) (middle) and 23(d) (right). In every other case the genus of the ribbon graph for $\overline{\Gamma}(w')$ is 1. Thus, we have $\mathrm{gs}(w') = \{(0, 0, 2), (1, 0, 2^n - 2)\}$, as desired.

(ii) Figure 24 is similar to Figure 23, except now we are considering $\overline{\Gamma}(w)$ where w is a repeat word of odd size $n - 1$.

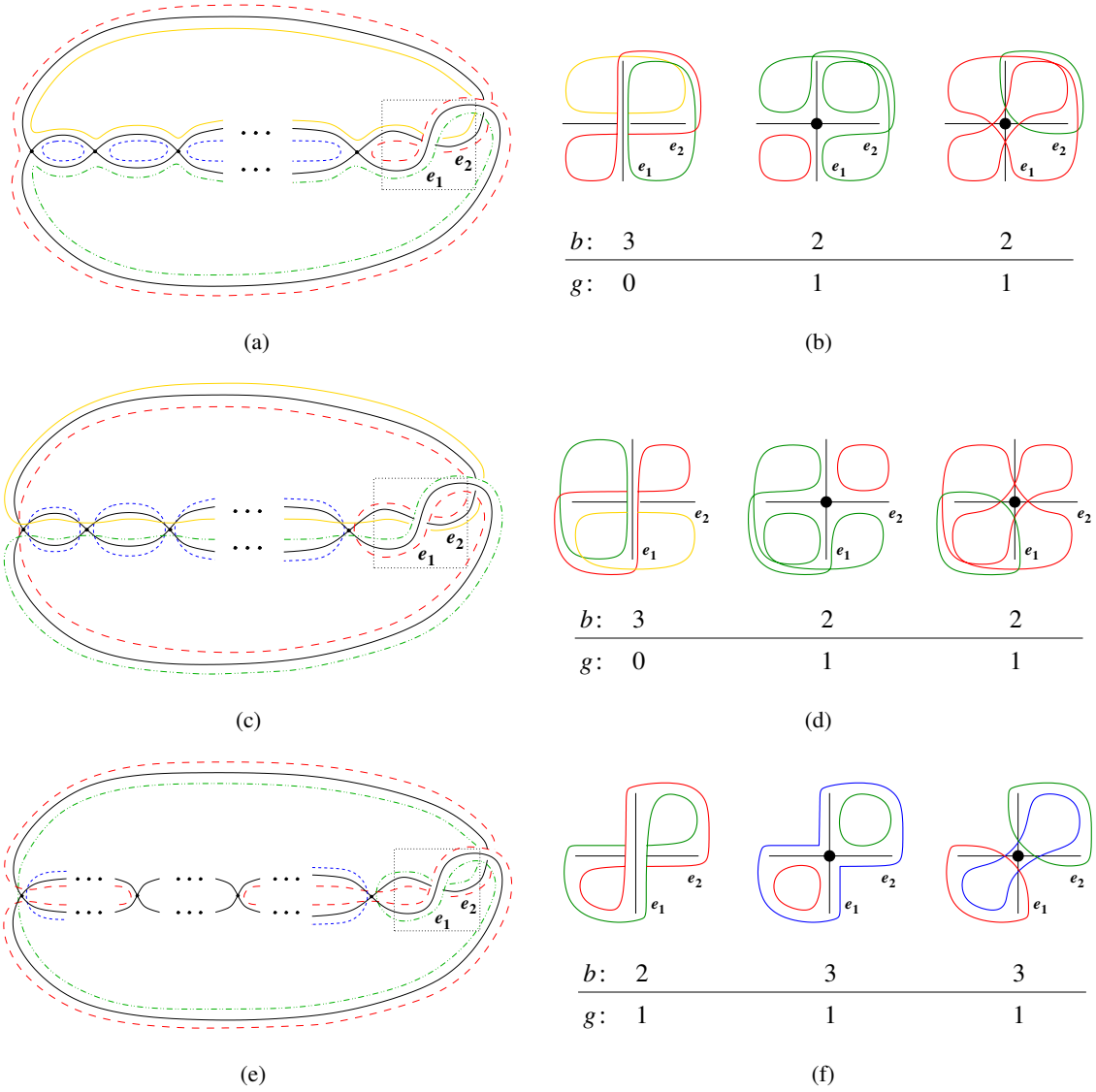


Figure 24: Case (ii): $n - 1$ is odd

Note that the closure edge of $\Gamma(w')$ with label e_2 is traced by a single boundary component in exactly two cases of the ribbon graph for $\overline{\Gamma}(w)$, namely, those cases depicted in Figures 24(b) (middle) and 24(d) (right). Also, the genus for the ribbon graph of $\overline{\Gamma}(w')$ in all cases is 1. Thus, $\mathrm{gs}(w') = \{(1, 2, 2^n - 2)\}$, as desired.

□

As a direct result, we obtain the genus ranges for the assembly graphs associated with repeat words and return words.

COROLLARY 4.5 *If w is a repeat word on n letters, then $\mathrm{gr}(w) = \{0\}$. If w is a return word on $n > 1$ letters, then $\mathrm{gr}(w) = \{1\}$ if n is even or $\mathrm{gr}(w) = \{0, 1\}$ if n is odd.*

Now we prove a generalized version of Theorem 4.1.

THEOREM 4.3 *Let w_1 and w_2 be double occurrence words and let $w = w_1 * w_2$ be their concatenation. Then*

$$\mathrm{gf}_1(w, g) = \sum_{g_1+g_2-1=g} \mathrm{gf}_1(w_1, g_1) \cdot \mathrm{gf}_1(w_2, g_2) + \sum_{g_1+g_2=g} \mathrm{gf}_1(w_1, g_1) \cdot \mathrm{gf}_2(w_2, g_2) + \mathrm{gf}_2(w_1, g_1) \cdot \mathrm{gf}_1(w_2, g_2)$$

and

$$\mathrm{gf}_2(w, g) = \sum_{g_1+g_2=g} \mathrm{gf}_2(w_1, g_1) \cdot \mathrm{gf}_2(w_2, g_2)$$

Proof. We argue similarly to the proof of Theorem 4.1 but with added focus on the generalized genus spectrum.

Set $\Gamma_1 = \overline{\Gamma}(w_1)$ and $\Gamma_2 = \overline{\Gamma}(w_2)$ with closure edges e_1 and e_2, respectively. Then $\Gamma = \overline{\Gamma}(w)$ is precisely the same as the graph obtained by connecting Γ_1 and Γ_2 through edges e_1 and e_2.

Let F_1 and F_2 be ribbon graphs of Γ_1 and Γ_2, respectively, and let F be the ribbon graph of Γ obtained from connecting Γ_1 and Γ_2 through edges e_1 and e_2 without changing the connection at any of the vertices in Γ_1 or Γ_2. Then the boundary components of F are determined by F_1 and F_2 as we have depicted in Figure 25. Let $g_i = g(F_i)$ for $i = 1, 2$ and consider the following cases on the number of boundary components tracing e_1 and e_2 to obtain values for $\mathrm{gf}_1(w, g_1 + g_2)$ and $\mathrm{gf}_2(w, g_1 + g_2)$.

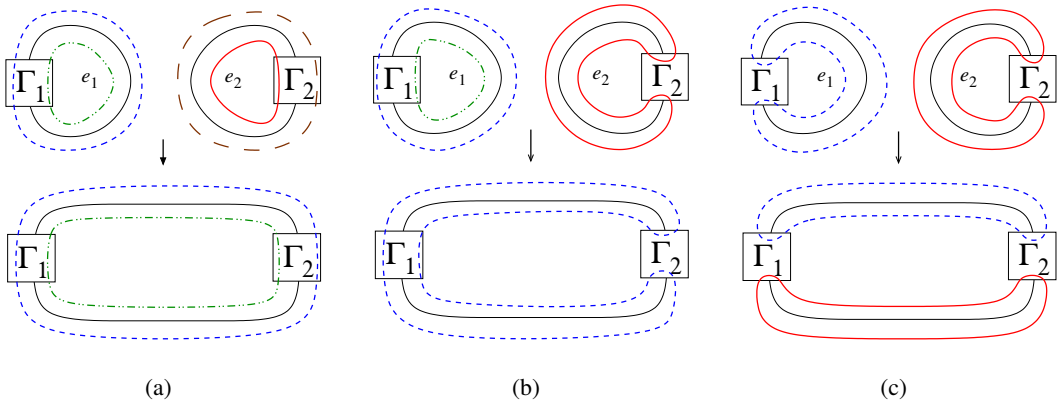


Figure 25: Boundary components before and after connecting graphs Γ_1 and Γ_2

(i) Suppose e_1 and e_2 are both traced by distinct boundary components in F_1 and F_2, respectively. Then Figure 25(a) shows that the closure edge of $\Gamma(w)$ in F is also traced by distinct boundary components. Furthermore, as we saw in Theorem 4.1 (part (i)), $g(F) = g_1 + g_2$. Since there are $\mathrm{gf}_2(w_1, g_1)$ ribbon graphs F_1 and $\mathrm{gf}_2(w_2, g_2)$ ribbon graphs F_2 satisfying the above, the possibilities come together to form $\mathrm{gf}_2(w_1, g_1) \cdot \mathrm{gf}_2(w_2, g_2)$ ribbon graphs F that contribute to $\mathrm{gf}_2(w, g_1 + g_2)$.

(ii) Suppose e_1 is traced by distinct boundary components in F_1 and e_2 is traced by a single boundary component in F_2. Then Figure 25(b) shows that the closure edge of $\Gamma(w)$ is also traced by a single boundary component. Also, from the proof of Theorem 4.1 (part (i)), we have that $g(F) = g_1 + g_2$. Since there are $\mathrm{gf}_2(w_1, g_1)$ ribbon graphs F_1 and $\mathrm{gf}_1(w_2, g_2)$ ribbon graphs F_2 satisfying the above, the possibilities come together to form $\mathrm{gf}_2(w_1, g_1) \cdot \mathrm{gf}_1(w_2, g_2)$ ribbon graphs F that contribute to $\mathrm{gf}_1(w, g_1 + g_2)$.

(iii) Suppose e_1 is traced by a single boundary component in F_1 and e_2 is traced by distinct boundary components in F_2. Then, similar to the last case with e_1 and e_2 transposed, the possibilities for F_1 and F_2 come together the form $\mathrm{gf}_1(w_1, g_1) \cdot \mathrm{gf}_2(w_2, g_2)$ ribbon graphs F that contribute to $\mathrm{gf}_1(w, g_1 + g_2)$.

(iv) Suppose e_1 and e_2 are traced by a single boundary component in F_1 and F_2, respectively. Then Figure 25(c) shows that the closure edge of $\Gamma(w)$ in F is also traced by a single boundary component. Also, as we saw in the proof of Theorem 4.1 (part (ii)), we have $g(F) = g_1 + g_2 - 1$. Since there are $\mathrm{gf}_1(w_1, g_1)$ ribbon graphs F_1 and $\mathrm{gf}_1(w_2, g_2)$ ribbon graphs F_2 satisfying the above, the possibilities come together to form $\mathrm{gf}_1(w_1, g_1) \cdot \mathrm{gf}_1(w_2, g_2)$ ribbon graphs F that contribute to $\mathrm{gf}_1(w, g_1 + g_2 - 1)$.

Now since all ribbon graphs F of Γ come from some case (i)-(iv) above, where g_1 and g_2 range over $\mathrm{gr}(\Gamma_1)$ and $\mathrm{gr}(\Gamma_2)$, respectively, the result follows. $\square$

Chapter 5

Comparison between Nesting Index and Genus Range

In this chapter we present some results which draw from both Chapter 3 on the nesting index property and Chapter 4 on the genus range property. In particular, we construct double occurrence words that realize certain values for nesting index and genus range. The first result shows that there exists a word with nesting index 1 and genus range $[0, n]$ for arbitrary $n > 0$.

LEMMA 5.1 *Let* $w = 123123$ *and let* w^n *be the word obtained by concatenating* $n \geq 1$ *copies of* w. *Then* $\mathrm{NI}(w^n) = 1$ *and* $\mathrm{gr}(\overline{\Gamma}(w^n)) = [0, n]$.

Proof. Since w^n is a repeated concatenation of the repeat word 123123, by Lemma 3.2, we have $\mathrm{NI}(w^n) = 1$. Now we want to show that $\mathrm{gr}(\overline{\Gamma}(w^n)) = [0, n]$. Note that the assembly graph $\overline{\Gamma}(w^n)$ is the same as the graph obtained by connecting n copies of $\overline{\Gamma}(w)$ through the closure edges of $\Gamma(w)$. By Theorem 4.2 the closure edge of $\Gamma(w)$ is traced by distinct boundary components in all of its ribbon graphs. Also, by Corollary 4.5, $\mathrm{gr}(\overline{\Gamma}(w)) = [0, 1]$. Thus, by Theorem 4.1, $\mathrm{gr}(\overline{\Gamma}(w^n)) = [0, n]$. □

The previous construction was of a double occurrence word with a small nesting index and a large genus range. In contrast, the next result shows that there exists a word of nesting index 2 and singleton genus range $\{n\}$ for arbitrary $n > 0$.

LEMMA 5.2 *Let* $w = 123231$ *and let* w^n *be the word obtained by concatenating* $n \geq 1$ *copies of* w. *Then* $\mathrm{NI}(w^n) = 2$ *and* $\mathrm{gr}(\overline{\Gamma}(w^n)) = \{n\}$.

Proof. Since w can not be obtained by concatenating repeat words and return words, then neither can w^n, hence, by Lemma 3.2, we have that $\mathrm{NI}(w^n) \neq 1$. However, applying reduction operation 1 to w^n two times reduces w^n to ϵ and thus, we have $\mathrm{NI}(w^n) = 2$. Now we claim that $\mathrm{gr}(\overline{\Gamma}(w^n)) = \{n\}$. First, note that 2323 is a repeat word on an even number of letters, and hence, by Corollary 4.5, $\mathrm{gr}(\overline{\Gamma}(2323)) = \{1\}$. Since w is a loop nesting of 2323, we have $\mathrm{gr}(\overline{\Gamma}(w)) = \{1\}$. Also, since the closure edge of $\Gamma(w)$ is a loop, the edge

is traced by distinct boundary components in all ribbon graphs of $\overline{\Gamma}(w)$. Now since $\overline{\Gamma}(w^n)$ is the same as the graph obtained by connecting n copies of $\overline{\Gamma}(w)$ through the closure edges of $\Gamma(w)$, by Theorem 4.1, we have $\mathrm{gr}(\overline{\Gamma}(w^n)) = \{n\}$. □

It is a simple exercise to check that there is no word w with $\mathrm{NI}(w) = 1$ and $\mathrm{gr}(w) = \{n\}$ for $n > 1$. Now we use the previous two results to show that there exists a word with arbitrary genus range and nesting index not greater than 2.

THEOREM 5.1 *Let $m \leq n$ be non-negative integers that are not both zero and let $w_1 = 123231$, $w_2 = 123123$ and let $w = w_1^m w_2^{n-m}$ be the word obtained by concatenating m copies of w_1 together with $n - m$ copies of w_2. Then $\mathrm{NI}(w) = 2$ and $\mathrm{gr}(w) = [m, n]$.*

Proof. If $m = 0$, then the conditions for Lemma 5.1 are satisfied and thus, $\mathrm{NI}(w) = 1$ and $\mathrm{gr}(w) = [0, n]$. If $n = m$, then the conditions for Lemma 5.2 are satisfied and thus, $\mathrm{NI}(w) = 2$ and $\mathrm{gr}(w) = \{m\} = [m, n]$. Otherwise, we obtain w by concatenating w_1^m and w_2^{n-m} where w_1^m, by Lemma 5.1, satisfies $\mathrm{gr}(\Gamma(w_1^m)) = \{m\}$ and w_2^{n-m}, by Lemma 5.2 satisfies $\mathrm{gr}(\Gamma(w_2^{n-m})) = [0, n-m]$. Note that, by Theorem 4.3, the closure edges of $\Gamma(w_1^m)$ and $\Gamma(w_2^{m-n})$ are traced by distinct boundary components in all of their respective ribbon graphs. Now since $\overline{\Gamma}(w)$ is the same as the graph obtained by connecting $\overline{\Gamma}(w_1^m)$ and $\overline{\Gamma}(w_2^{m-n})$ through their closure edges, then by Theorem 4.1, $\mathrm{gr}(\overline{\Gamma}(w)) = [m, n]$. Also, since w can not be obtained as a concatenation of repeat and return words, $\mathrm{NI}(w) \neq 1$. However, applying reduction operation 1 to w two times reduces w to ϵ and so $\mathrm{NI}(w) = 2$. □

Note that the only genus range not recognized by the construction above is the singleton $\{0\}$. However, this can be satisfied by any repeat word w. Indeed, by Lemma 3.2, $\mathrm{NI}(w) = 1$ and, by Corollary 4.5, $\mathrm{gr}(\overline{\Gamma}(w)) = \{0\}$. Interestingly, we can also create words with arbitrary nesting index and genus range $\{0\}$.

LEMMA 5.3 *Set $w_1 = 123321$ and for $n > 1$, recursively define w_n to be the double occurrence word obtained from w_{n-1} by inserting 12213443 between every loop, that is, every subword of the form aa for some letter a in Σ, and relabeling so that the result is still a double occurrence word. Figure 26 shows the sequence of assembly graphs $\Gamma(w_0)$, $\Gamma(w_1)$, and $\Gamma(w_2)$. Then we have $\mathrm{gr}(w_n) = \{0\}$ and $\mathrm{NI}(w_n) = n$.*

Proof. Note that w_1 is a repeat word, hence, by Corollary 4.5, has genus range $\{0\}$, and each word w_n is obtained from w_{n-1} by loop nesting. Then by Corollary 4.3, we have $\mathrm{gr}(w_n) = \{0\}$.

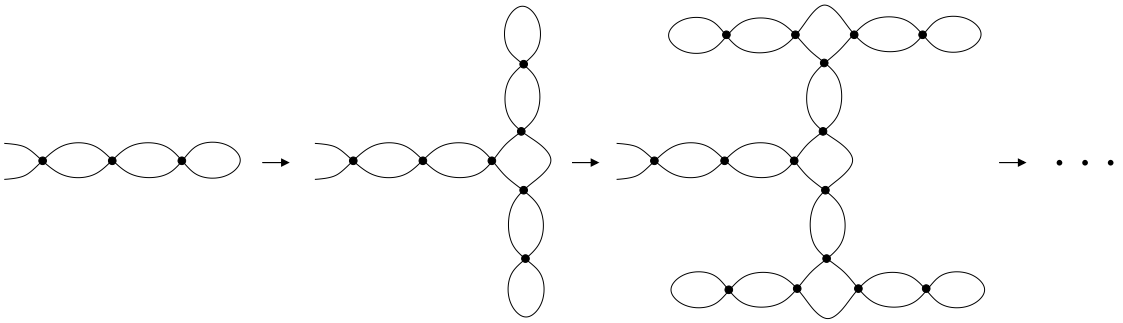


Figure 26: Sequence of assembly graphs $\Gamma(w_1), \Gamma(w_2), \Gamma(w_3), \ldots$ for w_n as defined in Lemma 5.3

Consider the double occurrence word w_n and note that removing a letter, that is, applying reduction operation 2 to w_n, provides no advantage. In other words, the shortest reduction of w_n consists of applying only reduction operation 1. Also note that by applying reduction operation 1 to w_n, we obtain w_{n-1}. It follows that $\mathrm{NI}(w_n) = \mathrm{NI}(w_{n-1}) + 1$. Since $\mathrm{NI}(w_1) = 1$, the result follows by induction on n. □

THEOREM 5.2 *There exists a word w with arbitrary nesting index ≥ 2 and arbitrary genus range.*

Proof. Let $m \leq n$ be non-negative integers. We show that there exists a word w with genus range $[m, n]$ and arbitrary nesting index at least 2. If $m = n = 0$, then by Lemma 5.3 there exists a word with arbitrary nesting index and genus range $\{0\}$. If $m \geq 0$ and $n \neq 0$, then by Lemma 5.1, there exists a word w_1 with $\mathrm{NI}(w_1) \leq 2$ and $\mathrm{gr}(w_1) = [m, n]$. From this word w_1 if we let w_2 be the double occurrence word obtained by concatenation of w_1 with 123321, then w_2 is a loop nesting of w_1, hence, $\mathrm{gr}(w_2) = \mathrm{gr}(w_1)$, and $\mathrm{NI}(w_2) = \mathrm{NI}(w_1)$. Further, for $n > 2$, if we recursively define w_n to be the word obtained from w_{n-1} by inserting 12213443 between every loop in w_{n-1}, then by arguing similarly to the proof of Lemma 5.3, we have $\mathrm{NI}(w_n) = \mathrm{NI}(w_{n-1}) + 1$. Also, since each w_n is a loop-nesting of w_{n-1}, hence, a loop-nesting of w_1, we have $\mathrm{gr}(w_n) = \mathrm{gr}(w_1) = [m, n]$. □

Chapter 6
Conclusion

In Section 3.2 we remarked that the nesting index could provide insight into the number of steps in the rearrangement processes of the micronuclear genome. While this may be true, we currently have no biological explanation for the reduction operation 2. Recall that the letters of the double occurrence word correspond to vertices in the assembly graph and, from a biological viewpoint, these vertices represent places where the DNA aligns, or "connection sites", in the recombination of the micronuclear ciliate genome. Then removing a letter from a double occurrence word may correspond to removing a "connection site", which is something that the genome obviously should not normally do. We could then improve on the biological application of the nesting index if we were to not only remove the letter ("connection site"), but then also replace that letter later in the reduction of that double occurrence word. Implementing such a reduction process by computer program may be computationally demanding without the development of sophisticated algorithms (if any exist) and we have not yet begun to explore such possibilities. In Section 3.4, however, we have given a characterization of double occurrence words that are 1-reducible and perhaps with scrambled genomes that correspond to double occurrence words that are 1-reducible, the current nesting index may more accurately predict the number of steps in the rearrangement processes of the corresponding genome.

The data in Table 1 presents some interesting trends in the nesting index of double occurrence words. In particular, we are curious about the following conjecture and open question.

CONJECTURE 1 For $n \geq 1$, the shortest word w with $\mathrm{NI}(w) = n$ has length $|w| = 2(n + \lfloor\sqrt{n-1}\rfloor)$.

QUESTION 1 *Can we characterize all double occurrence words w such that $w' - a = w$ implies $\mathrm{NI}(w') \leq \mathrm{NI}(w)$? In other words, double occurrence words w where in no way can we add a letter to w to increase its nesting index?*

From Table 1, we know that the word(s) of size 1 and nesting index 1, size 5 and nesting index 4, or size 11 and nesting index 9 have this property. These are 11, 1234254153, and cyclic permutations of

$$1, 2, 3, 4, 5, 6, 7, 8, 9, 3, 10, 6, 2, 11, 9, 5, 1, 10, 8, 4, 11, 7\ ,$$

respectively. Another example which we can easily check has this property is any word of the form $1122\cdots nn$ for arbitrary $n \geq 1$.

In Chapter 4 we were often interested in whether a particular edge e in an assembly graph Γ was traced by a single boundary component or two distinct boundary components in a particular ribbon graph of Γ and, moreover, whether this was consistent over all or only some ribbon graphs of Γ. A positive answer to the following questions would be useful in applying Theorem 4.1 to assembly graphs Γ_1 and Γ_2 that do not satisfy the conditions of part (i).

QUESTION 2 *Can we characterize assembly graphs Γ such that if the edge e in Γ is traced by a single boundary component in some ribbon graph of Γ, then there exists a ribbon graph F of Γ where $g(F) = \min(\mathrm{gr}(\Gamma))$ and e is traced by a single boundary component in F? Can we characterize assembly graphs Γ which for any edge e in Γ, there exists a ribbon graph F of Γ where $g(F) = \max(\mathrm{gr}(\Gamma))$ and e is traced by distinct boundary components in F?*

Further interest for the genus spectrum lies in determining what possible values can actually be realized as the genus spectrum of an assembly graph. For example, we believe there are words in which half of all ribbon graphs realize some genus, and the other half realize another genus.

CONJECTURE 2 If $[m, m+1]$ is realized as the genus range of some assembly graph Γ on n vertices, then there exists an assembly graph $\Gamma_{\frac{1}{2},m}$, such that $\mathrm{gs}(\Gamma_{\frac{1}{2},m}) = \{(m, 2^{n-1}), (m+1, 2^{n-1})\}$.

Although we showed in Chapter 5 that there are words with arbitrary nesting index ≥ 2 and arbitrary genus range, there is still some interest in how these properties relate to another property called the assembly number of an assembly graph. The assembly number of an assembly graph Γ is the minimum number of paths in Γ where each vertex is visited exactly once in exactly one path and a "90° turn" is made at each vertex in each path [6].

References

[1] A. Angeleska, N. Jonoska, M. Saito, L.F. Landweber, RNA-guided DNA assembly, *Journal of Theoretical Biology* 248:4 (2007) 706–720.

[2] A. Angeleska, N. Jonoska, M. Saito, DNA recombination through assembly graphs, *Discrete and Applied Math*, 157 (2009) 3020–3037.

[3] R. Arredondo, Reductions on Double Occurrence Words, Proceedings of the Fourty-fourth Southeastern International Conference on Combinatorics, Graph Theory, and Computing. *Congressus Numerantium* 218 (2013) 45–56.

[4] K. Bhandari, H.A. Dye, L.H. Kauffman, Lower bounds on virtual crossing number and minimal surface genus, in: *The Mathematics of Knots v.1*, Contributions in Mathematical and Computational Sciences, B. Markhus, V. Denis (Eds), Springer (2011) 31–43.

[5] D. Buck, E. Dolzhenko, N. Jonoska, M. Saito, K. Valencia, Genus Ranges of 4-Regular Rigid Vertex Graphs. Submitted 21 Nov 2012. `arXiv:1211.4939 [math.GT]`

[6] J. Burns, E. Dolzhenko, N. Jonoska, T. Muche, M. Saito, Four-regular Graphs with Rigid Vertices Associated to DNA Recombination, *Discrete Applied Mathematics* 161 (2013) 1378–1394.

[7] G. Cairns, D.M. Elton, The planarity problem for signed Gauss words, *Journal of Knot Theory and Its Ramifications* 2 (1993) 359–367.

[8] J.S. Carter, Classifying immersed curves, *Proc. Amer. Math. Soc.* 111:1 (1991) 281–287.

[9] W. Chang, P. Bryson, H. Liang, M. Shin, L. Landweber, The evolutionary origin of a complex scrambled gene, *Proceedings of the National Academy of Science*, 102 (2005) 15149–15154.

[10] C. Godsil, G. Royle, *Algebraic Graph Theory*, Graduate Texts in Mathematics, Volume 207, Springer-Verlag, New York, 2001.

[11] J.L. Gross, T.W. Tucker, *Topological Graph Theory*, Wiley, New York, 1987.

[12] D. Hoffman, D. Prescott, Evolution of internal eliminated segments and scrambling in the micronuclear gene encoding DNA polymerase α in two Oxytricha species, *Nucleic Acids Research* 25 (1997) 1883–1889.

[13] N. Jonoska, private communication, 2013.

[14] L. H. Kauffman, Invariant of Graphs in Three-Space, *Trans. Amer. Math. Soc.* 311:2 (1989) 697–710.

[15] L. Landweber, T. Kuo, E. Curtis, Evolution and assembly of an extremely scrambled gene, *Proceedings of the National Academy of Science* 97 (2000) 3298–3303.

[16] D. Prescott, Genome Gymnastics: Unique Models of DNA Evolution and Processing in Ciliates, *Nature Reviews Genetics* 1:3 (2000) 191–198.