involve

a journal of mathematics

msp

2016 vol. 9, no. 1

involve

msp.org/involve

Cover: Alex Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2016 is US $160/year for the electronic version, and $215/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

INVOLVE 9:1 (2016)
dx.doi.org/10.2140/involve.2016.9.1

Using ciliate operations to construct chromosome phylogenies

Jacob L. Herlin, Anna Nelson and Marion Scheepers

(Communicated by Joseph A. Gallian)

Whole genome sequencing has revealed several examples where genomes of different species are related by permutation. The number of certain types of rearrangements needed to transform one permuted list into another can measure the distance between such lists. Using an algorithm based on three basic DNA editing operations suggested by a model for ciliate micronuclear decryption, this study defines the distance between two permutations to be the number of ciliate operations the algorithm performs during such a transformation. Combining well-known clustering methods with this distance function enables one to construct corresponding phylogenies. These ideas are illustrated by exploring the phylogenetic relationships among the chromosomes of eight fruit fly (*Drosophila*) species, using the well-known UPGMA algorithm on the distance function provided by the ciliate operations.

Over evolutionary time, "local" DNA editing events such as nucleotide substitutions, deletions or insertions diversify the set of DNA sequences present in organisms. Results of whole genome sequencing suggest that also "global" DNA editing events diversify these DNA sequences.

Consider two species S_1 and S_2 with a common ancestor whose genome was organized over n linear chromosomes. A gene G of the ancestor was inherited as gene G_1 by species S_1 and as gene G_2 by species S_2. G_1 and G_2 are *orthologous* genes, or simply *orthologs*. Assume that the species S_1 and S_2 each also has n chromosomes, and that for each ancestral chromosome i, the orthologs of any ancestral gene on chromosome i are also in the descendant species S_1 and S_2 on the corresponding chromosome i. This assumption is known, in the context of certain

MSC2010: 05E15, 20B99, 92-08, 92D15, 92D99.
Keywords: permutations, reversals, block interchanges, fruit fly, ciliate, phylogeny.
This research was started during the Summer of 2011 when Herlin and Nelson participated in the Boise State University Mathematics REU program, funded by NSF grant DMS 1062857. We gratefully acknowledge funding by the NSF and by Boise State University.

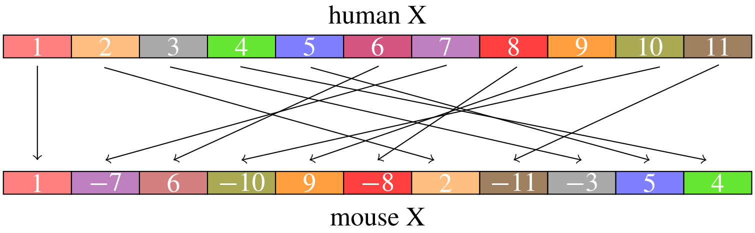


Figure 1. The permutation between 11 synteny blocks of the human and the mouse X chromosomes. A negative symbol denotes an orientation change by a 180° rotation of a synteny block. The lengths of synteny blocks are not to scale. See Figure 2 of [Pevzner and Tesler 2003].

fruit fly species, as the *Muller hypothesis*[1]. In this paper we shall assume the Muller hypothesis for our applications.

It may happen that the order in which orthologs on chromosome i appear in species S_1 is different from the order in which they appear in species S_2. In this case chromosome i in each of these two species can be partitioned into a number, say k, of *synteny blocks*[2]: a synteny block is a maximal list of adjacent orthologous genes that have the same adjacencies in the two species. In this definition of a synteny block, we permit blocks consisting of single genes. An endpoint of a synteny block is also called a *breakpoint*. Synteny blocks may have opposite orientation in two species. Thus the synteny blocks of chromosome i of species S_1 are a *signed* permutation of the corresponding synteny blocks of chromosome i of species S_2. This phenomenon is observed in several branches in the tree of life. Figure 1 illustrates the phenomenon for 11 synteny blocks of orthologous genes in the X chromosome of human and mouse.

Since the appearance of [Sturtevant and Dobzhansky 1936] and [Dobzhansky and Sturtevant 1938] on fruit fly genomes, it has been popular to use *reversals*[3] as the primary global DNA sequence editing operation to describe phylogenetic relationships among genomes. See, for example, [Bafna and Pevzner 1995; Hannenhalli and Pevzner 1999].

An insightful phylogenetic analysis that includes fine structural elements of reversals is given in [Bhutkar et al. 2008]. It addresses the question of whether reversals can occur at arbitrary locations in the genome of an organism. Certain locations,

[1] Named after H. J. Muller [1940] who observed that for the data then known for relatives of *Drosophila melanogaster*, this assumption is true even for chromosome arms.

[2] This definition of a synteny block is more restrictive than the one used in [Bhutkar et al. 2008]: The latter allows for differences in gene order up to a certain threshold, and does not allow for single gene blocks. See the section “An application to genome phylogenetics” for more information.

[3] A reversal is a rotation of a DNA segment through 180°. Reversals are also called inversions.

which would disrupt the coding region of an essential gene, would not be observed in extant organisms. Similarly, locations that negatively affect the fitness of organisms would disappear over time due to "purifying selection". Additionally, certain sequence motifs may actually promote DNA recombination that results in a genome rearrangement. For example, [Coghran and Wolfe 2002] reports a correlation between *breakpoints*[4] associated with rearrangements, and repetitive DNA. This point is also considered in [Bhutkar et al. 2008]. In the review [Hughes 2000], a similar correlation between rearrangements in bacterial genomes and repetitive DNA is discussed. These considerations suggest that genome rearrangement events that lead to the diverse genomes we observe in nature are not arbitrary, but constrained by contexts. In this paper, we explore the use of *context-directed* DNA recombination events to analyze genome rearrangements and to construct a phylogeny based on these.

In recent years, transpositions and block interchanges have also been considered as possible global DNA sequence editing operations [Bafna and Pevzner 1998; Coghran and Wolfe 2002; Mira and Meidanis 2007; Yancopoulos et al. 2005]. In a *block interchange*, two disjoint segments of a chromosome exchange locations without changing orientation. Thus, in Figure 1, synteny blocks 2 and 7 would have been a block interchange if synteny block 7 did not also undergo a reversal. A *transposition* is a special block interchange where the two segments that exchange location are adjacent. In Figure 1, synteny blocks 4 and 5 illustrate a transposition.

On page 1661 of [Bhutkar et al. 2008], in the discussion of their selection of genes to which their analysis of rearrangements in fruit fly genomes apply, the authors indicate that genes deemed to have been relocated by a transposition rather than a reversal have been explicitly removed from the analysis. Thus, the analysis of [Bhutkar et al. 2008] features reversals exclusively. On the other hand, the analysis in [Coghran and Wolfe 2002] of rearrangements in the genomes of two nematode species includes reversals, transpositions and *translocations*. A translocation occurs when segments from two different chromosomes exchange positions. In this paper, we explore only reversals and block interchanges (both constrained by contexts) in the analysis of rearrangements.

Experimental results from ciliate laboratories present us with examples of DNA editing operations that routinely occur during developmental processes in these organisms. The textbook [Ehrenfeucht et al. 2004] and the two surveys [Prescott 1994; 2000] give a good starting point for information about these "ciliate operations" and the corresponding biological background. We shall call the yet to be fully identified system in ciliates that accomplishes micronuclear decryption[5], the *ciliate decryptome*.

[4] Referring to the mouse X chromosome in Figure 1, a breakpoint is a transition point between synteny blocks that are not consecutively numbered.

[5] Some details regarding this process are given below in Section 1.

We shall illustrate how to use “ciliate operations” to deduce potential phylogenetic relationships from genome rearrangement phenomena. Previous work, including [Bafna and Pevzner 1995; Bhutkar et al. 2008; Hannenhalli and Pevzner 1999], used unconstrained reversals to deduce phylogenetic relationships. Our main ideas are to use ciliate genomic elements to model two genomes related by permutations of locations and orientations of synteny blocks, to apply the context-directed DNA operations of the ciliate decryptome to define a distance function between the relevant permuted genomes, and to then use a classical distance-based algorithm to derive phylogenies. Of the several different distance-based algorithms available, we selected the UPGMA algorithm[6].

Then we apply these ideas to chromosomes of eight species of fruit flies (*Drosophila*) to obtain a phylogeny for each of these chromosomes.

The use of ciliate operations as the basis for deriving a distance function has the attractive feature that the ciliate decryptome is programmable [Nowacki et al. 2007], and the computational steps taken by the decryptome can be monitored under laboratory conditions [Möllenbeck et al. 2008]. Thus, there are extant organisms that are poised to be employed as DNA computing devices naturally equipped to determine phylogenetic relationships among permuted genomes.

Our paper is organized as follows: In Section 1 we briefly describe ciliate nuclear duality. This duality is the basis for modeling pairs of genomes related by permutation as genetic elements of the ciliate genome. In Section 2 we briefly describe the context-directed DNA operations of the ciliate decryptome. In Section 3 we introduce and analyze the mathematical notion of a pointer list. In Section 4 we model relevant features of the ciliate decryptome’s DNA operations by mathematical operations on pointer lists. In Section 5 we describe an algorithm which we call the HNS algorithm, that uses these operations on pointer lists to compute the distance between chomosomes that are related by permutation. In Section 6 we use data downloaded from flybase.org and the HNS and UPGMA algorithms to construct phylogenies over eight species for each of the fruit fly chromosomes. In the closing section, we discuss possible future directions related to this work.

1. Ciliates and nuclear duality

A ciliate is a single cell eukaryote that hosts two types of nuclei: one type, the macronucleus, contains the transcriptionally active somatic genome, while the other type, the micronucleus, contains a transcriptionally silent germline-like genome. The micronuclear genome is, in the technical sense of the word, an encrypted version of the macronuclear genome. Special events in the ciliate life cycle predictably trigger

[6] Descriptions of UPGMA can be found in Chapter 27 of [Barton et al. 2007], available online, or in the textbook [Clote and Backofen 2000].

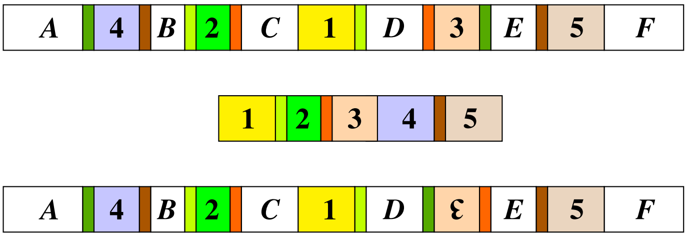


Figure 2. The top diagram depicts a possible micronuclear precursor, and the bottom diagram is another possible micronuclear precursor of the macronuclear gene in the middle diagram.

conjugation between a pair of mating-compatible cells. Conjugation results in what amounts to a Diffie–Hellman exchange[7] between two conjugants, the formation of a new micronucleus in each, and the decryption of one copy of the new micronuclear genome to establish a replacement macronuclear genome, while in each conjugant the instances of its preexisting genome are discarded. Readers interested in a thorough survey of ciliate nuclear duality could consult [Prescott 1994].

The relationship between micro- and macronuclear DNA. To describe the experimentally observed relationship between the micronuclear and macronuclear DNA molecules, consider Figure 2.

The micronuclear DNA sequences in the top and the bottom rows of Figure 2 each have three types of regions: The white blocks, labeled with letters, are called *internal eliminated sequences* (IESs). The blocks labeled with numbers are called *macronuclear destined sequences* (MDSs), while the narrow strips are called *pointers*. As the micronuclear precursors show, there are two copies of each pointer. For example, MDS 2 has a pointer on the left flank that is identical to the pointer on the right flank of MDS 1. This pointer will be called the "1-2 *pointer*". And MDS 2 has a pointer on its right flank which is identical to the pointer on the left flank of MDS 3. This pointer is called the "2-3 *pointer*". The other pointers are named similarly. Also note that MDS 1 does not have a pointer on its left flank, and MDS 5 does not have a pointer on its right flank. As MDS 3 and the pointers on its flanks show in the bottom row of Figure 2, in the micronuclear precursor, an MDS plus its flanking pointer(s), as a unit, can be in a 180-degree rotated orientation of the corresponding components in the macronuclear gene. The corresponding macronuclear sequence in the middle row of Figure 2 contains only one of each of the pointers present in its micronuclear precursor, and all the MDSs, but none

[7] A Diffie–Hellman exchange is a cryptographic protocol for secure exchange of a secret key in a hostile environment. The conjugants exchange a haploid copy of the germline genome, which is an encrypted version of the somatic genome.

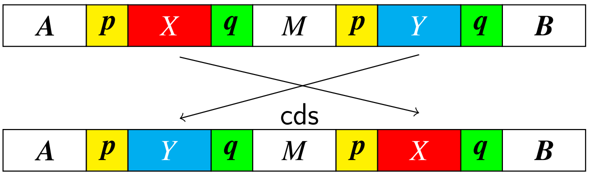


Figure 3. Context-directed block swaps: the $\boldsymbol{p} \cdots \boldsymbol{q} \cdots \boldsymbol{p} \cdots \boldsymbol{q}$ pointer context permits swapping the DNA segments X and Y.

of the IESs of the micronuclear precursor. In the macronuclear sequence, these components occur in a specific order, which we call the *canonical order*.

In shorthand, the micronuclear precursor in the top row of Figure 2 is [4, 2, 1, 3, 5], while the micronuclear precursor in the bottom row of Figure 2 is [4, 2, 1, −3, 5].

2. The ciliate DNA operations

We now turn to the actual ciliate algorithm that processes micronuclear precursors to produce their corresponding macronuclear versions. The articles [Angeleska et al. 2007; Prescott et al. 2003] propose hypotheses about biochemical processes that perform the decryption algorithm in ciliates. We do not address the biochemical foundations here.

The textbook [Ehrenfeucht et al. 2004] describes three DNA editing operations underlying this decryption process. There is experimental evidence that these three operations accomplish the decryption process. The article [Möllenbeck et al. 2008] gives experimental data about the DNA products of intermediate steps of the ciliate algorithm. We henceforth assume that the three operations that produce macronuclear molecules from their micronuclear precursors are as proposed in [Ehrenfeucht et al. 2004]: context-directed block interchanges (swaps), context-directed reversals and context-directed excisions.

Context-directed block interchanges (swaps). The top strip in Figure 3 represents a segment of DNA in a micronuclear chromosome of some ciliate. The symbols $\boldsymbol{p}$ and $\boldsymbol{q}$ denote identified pointers, while A, B, M, X, Y represent segments of DNA.

The three necessary conditions to swap segments X and Y are:

(1) X and Y both have an occurrence of each of the pointers $\boldsymbol{p}$ and $\boldsymbol{q}$ at their flanks;

(2) the pointer pair $\boldsymbol{p}$, $\boldsymbol{q}$ appears in the (alternating) context $\cdots \boldsymbol{p} \cdots \boldsymbol{q} \cdots \boldsymbol{p} \cdots \boldsymbol{q} \cdots$;

(3) neither occurrence of the pointer $\boldsymbol{p}$ nor of pointer $\boldsymbol{q}$ is flanked by a pair of successively numbered MDSs. For specificity consider Figure 4, where numbered blocks denote MDSs while lettered blocks denote IESs. The X of Figure 3 may be taken to be the segment $2B$ of Figure 4, while the Y of Figure 3 may be taken to be the segment $D3$ of Figure 4.

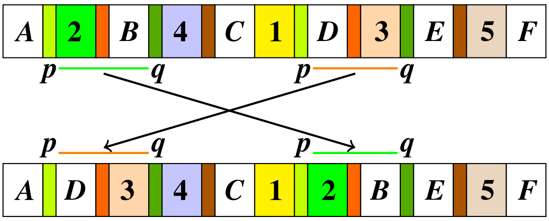


Figure 4. The top diagram depicts a possible micronuclear precursor, and the bottom diagram is the result of cds applied to the pointer pair $\boldsymbol{p} = (1, 2)$ and $\boldsymbol{q} = (3, 4)$.

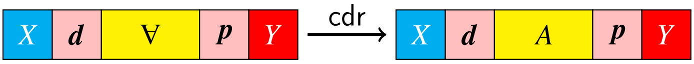


Figure 5. Context-directed reversal: the $-\boldsymbol{p}\cdots\boldsymbol{p}$ or $\boldsymbol{p}\cdots-\boldsymbol{p}$ pointer context-permits 180° rotation of flanked segment A.

Only when *all three* conditions are met is an interchange of the segments X and Y permitted. The result of this swap is depicted in the bottom strip of Figure 3. The reader may check, by comparing the bottom strips of Figures 3 and 4, that subsequent to an application of cds the contextual conditions (1) and (2) are still valid, but condition (3) is no longer met: indeed, one occurrence of each of the pointers $\boldsymbol{p}$ and $\boldsymbol{q}$ is now flanked by successively numbered MDSs.

Context-directed reversal. To describe a context-directed reversal, consider the left strip in Figure 5. It depicts a segment of DNA appearing in the micronucleus.

To rotate the yellow segment, labeled by an upside-down A, by 180°, that is, to reverse A, two necessary contextual conditions must be met:

(1) A is flanked by a pointer $\boldsymbol{p}$ and by the 180° rotation[8] of $\boldsymbol{p}$;

(2) neither occurrence of $\boldsymbol{p}$ is flanked by successively numbered MDSs. For specificity, consult Figure 6, where numbered blocks denote MDSs and lettered blocks denote IESs. The A of Figure 5 corresponds to the segment $-2C4D$ of Figure 6.

Only when *both* of these contextual requirements are met is rotation of the segment flanked by the relevant pointer context permitted. The result of this context-directed reversal is depicted by the right strip in Figure 5, and the corresponding bottom strip of Figure 6. As illustrated, subsequent to a context-directed reversal, one of the occurrences of the pointer $\boldsymbol{p}$ now has successively numbered MDSs on both flanks and no further applications of cdr are permitted to this pointer context.

[8] In text, the 180° rotation of $\boldsymbol{p}$ will be denoted $-\boldsymbol{p}$.

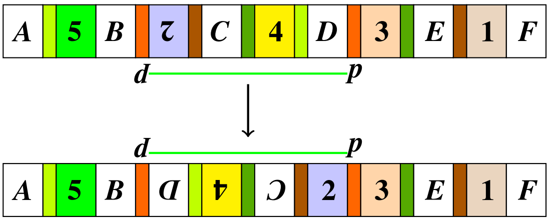


Figure 6. The top row depicts a possible micronuclear precursor. The bottom row results from cdr applied to the pointer $\boldsymbol{p} = (2, 3)$.

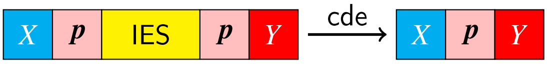


Figure 7. Context-directed excision: the IES flanked by pointer $\boldsymbol{p}$ on both sides is removed, along with one copy of $\boldsymbol{p}$.

Context-directed excision. To describe context-directed excision, consider Figure 7. In it, the pointer $\boldsymbol{p}$ flanks a DNA segment identified as an IES (the yellow segment). This context $\boldsymbol{p}\,\mathsf{IES}\,\boldsymbol{p}$ permits the excision of the IES segment plus one of the pointers, with the result of joining the DNA segments flanking the original pair of pointers, to the flanks of the remaining pointer.

Observe that context-directed block interchanges and context-directed reversals do not decrease or increase the length of the string they operate on, and they retain all the pointers. But context-directed excision, as illustrated in Figure 7, changes the pointer contexts by deleting selected pointers and IESs.

3. Pointer lists

Pointers are an essential ingredient of the three DNA editing operations. We exploit this central role of pointers by now basing our computational formalism (that mathematically models these three ciliate operations) on pointers. Towards this end, we introduce the notion of a *pointer list*[9].

Definition 1. A finite sequence $P := [x_1, \ldots, x_m]$ of integers is said to be a *pointer list* if it satisfies the following six conditions:

(1) m is an even positive integer.

(2) There is a unique i with $\mu = |x_i| = \min\{|x_j| : 1 \le j \le m\}$.

(3) There is a unique j with $\lambda = |x_j| = \max\{|x_i| : 1 \le i \le m\}$.

[9] In anticipation of wider applicability of the notion of a pointer list, we give a definition that is more general than the specific instance of it that we need.

(4) For each $i \in \{1, \dots, m\}$ with $\mu < |x_i| < \lambda$, there is a unique $j \in \{1, \dots, m\} \backslash \{i\}$ such that $|x_i| = |x_j|$.

(5) For each odd $i \in \{1, \dots, m\}$, $x_i \le x_{i+1}$ and $x_i \cdot x_{i+1} > 0$.

(6) Whenever $i \in \{1, \dots, m\}$ is odd, there is no j such that $|x_i| < |x_j| < |x_{i+1}|$ or $|x_{i+1}| < |x_j| < |x_i|$.

The following two mathematical facts are important in reasoning about ciliate operations on pointer lists.

Lemma 1. *Let $[x_1, x_2, \dots, x_{m-1}, x_m]$ be a pointer list. If i and j are distinct indices for which $|x_i| = |x_j|$, then x_i and x_j have the same sign if, and only if, i and j have distinct parity.*

Lemma 2. *If $[x_1, x_2, \dots, x_{m-1}, x_m]$ is a pointer list of length larger than* 4, *then at least one of the following three statements is false*:

(a) $(\forall i)(x_i \neq x_{i+1})$.

(b) $(\forall i)(\forall j)$(*If* $|x_i| = |x_j|$, *then* $x_i = x_j$).

(c) $(\forall i)(\forall j)(\forall k)(\forall \ell)$(*If* $i \neq k$, $j \neq \ell$, $i < j$ *and* $x_i = x_k$ *and* $x_j = x_\ell$, *then either* $i < j < \ell < k$ *or* $i < k < j < \ell$).

In the interest of readability, the somewhat lengthy, yet elementary, proofs of these facts are left to the reader.

Pointer lists to which we will apply the ciliate operations come about as follows: Let $\mathbb{Z}$ denote the set of integers. For a set S, the symbol ${}^{<\omega}S$ denotes the set of finite sequences with entries from S. For an integer z, we define

$$\check{z}(1) = \begin{cases} z & \text{if } z = |z|, \\ z - 1 & \text{otherwise,} \end{cases}$$

and in all cases, $\check{z}(2) = \check{z}(1) + 1$. Then define the function $\pi : {}^{<\omega}\mathbb{Z} \to {}^{<\omega}\mathbb{Z}$ by

$$\pi([z_1, \dots, z_k]) = [\check{z}_1(1), \check{z}_1(2), \dots, \check{z}_k(1), \check{z}_k(2)].$$

Thus, for example, $\pi([-1, 4, 3, 5, 2, -9, 7, 10, -8, 6])$ is the sequence

$$[-2, -1, 4, 5, 3, 4, 5, 6, 2, 3, -10, -9, 7, 8, 10, 11, -9, -8, 6, 7].$$

It can be verified that this sequence is indeed a pointer list. The following lemma captures this fact.

Lemma 3. *For each finite sequence $M := [s_1, s_2, \dots, s_n]$ of nonzero integers such that there is an integer m for which $\{|s_i| : 1 \le i \le n\} = \{m + 1, \dots, m + n\}$, the sequence $\pi(M)$ is a pointer list.*

The proof consists of verifying that $\pi(M)$ meets all stipulations of Definition 1.

4. The ciliate operations on pointer lists

We now introduce three special functions, cde, cdr and cds, from ${}^{<\omega}\mathbb{Z}$ to ${}^{<\omega}\mathbb{Z}$, inspired by the three ciliate operations, as follows. Let $P := [x_1, \dots, x_m]$ be a given finite sequence.

Context directed excision:

$$\mathsf{cde}(P) = \begin{cases} P & \text{if there is no } i \text{ with } x_i = x_{i+1}, \\ [x_1, \dots, x_{i-1}, x_{i+2}, \dots, x_m] & \text{for } i \text{ minimal with } x_i = x_{i+1}, \end{cases}$$

Context-directed reversal:

$$\mathsf{cdr}(P) = \begin{cases} P & \text{if there are no } i < j \\ & \text{with } x_i = -x_j, \\ [x_1, \dots, x_{i-1}, x_i, \underline{-x_j, \dots, -x_{i+1}}, x_{j+1}, \dots, x_m] & \text{for the minimal } i \text{ with} \\ & x_i = -x_j \text{ for a } j > i. \end{cases}$$

Context-directed block swaps: We set $\mathsf{cds}(P) = P$ if there are no $i < j < k < \ell$ with $x_i = x_k$ and $x_j = x_\ell$. However, if there are $i < j < k < \ell$ with $x_i = x_k$ and $x_j = x_\ell$, then choose the least such i, and for it the least corresponding j, and define $\mathsf{cds}(P)$ to be

$$[x_1, \dots, x_i, \underline{x_k, \dots, x_\ell}, x_j, \dots, x_{k-1}, \underline{x_{i+1}, \dots, x_{j-1}}, x_{\ell+1}, \dots, x_m].$$

These three operations have now been defined on arbitrary finite sequences of integers. They behave rather well on the subset $\mathsf{PL} = \{\sigma \in {}^{<\omega}\mathbb{Z} : \sigma \text{ is a pointer list}\}$ of their domain, as stated in the next two theorems. In the interest of readability the proofs have been omitted.

Theorem 4. *If P is a pointer list of length larger than 4, then at least one of the following statements is true*:

(1) $\mathsf{cde}(P) \neq P$.

(2) $\mathsf{cdr}(P) \neq P$.

(3) $\mathsf{cds}(P) \neq P$.

Theorem 5 (pointer list preservation). *Let $P = [x_1, \dots, x_m]$ be a pointer list. Then each of $\mathsf{cde}(P)$, $\mathsf{cdr}(P)$ and $\mathsf{cds}(P)$ is a pointer list.*

A finite sequence σ is a *fixed point* of a function $F : {}^{<\omega}\mathbb{Z} \to {}^{<\omega}\mathbb{Z}$ if $F(\sigma) = \sigma$.

Theorem 6. *If P is a pointer list of length larger than 4 and not a fixed point of $F \in \{\mathsf{cdr}, \mathsf{cds}\}$, then $F(P)$ is not a fixed point of* cde.

5. The HNS algorithm

Call a pointer list a *destination* if it is one of the following: $[\mu, \lambda]$, $[-\lambda, -\mu]$, or for some integer z with $|z| \notin \{\lambda, \mu\}$, the pointer list is one of $[z, \lambda, \mu, z]$ or $[z, -\mu, -\lambda, z]$.

Let P be a pointer list. Letting $\mathsf{cde}^i(P)$ denote the i-th iteration of cde on P, define $e(P)$ to be the minimal value of i such that $\mathsf{cde}^{i+1}(P) = \mathsf{cde}^i(P)$. Then define $\mathsf{E}(P) = \mathsf{cde}^{e(P)}(P)$.

In the following theorem, recall that a finite sequence σ is a *fixed point* of a function $F : {}^{<\omega}\mathbb{Z} \to {}^{<\omega}\mathbb{Z}$ if $F(\sigma) = \sigma$.

Theorem 7. *For a given pointer list P_0, define the sequence $P_0, P_1, \ldots, P_i, \ldots$ so that*

$$P_{i+1} = \begin{cases} \mathsf{E}(P_i) & \textit{if } P_i \textit{ is not a } \mathsf{cde} \textit{ fixed point,} \\ \mathsf{cds}(P_i) & \textit{if } P_i \textit{ is a } \mathsf{cde}, \textit{ but not a } \mathsf{cds} \textit{ fixed point,} \\ \mathsf{cdr}(P_i) & \textit{if } P_i \textit{ is a } \mathsf{cde} \textit{ and a } \mathsf{cds} \textit{ but not a } \mathsf{cdr} \textit{ fixed point.} \end{cases}$$

Then the sequence $P_0, P_1, \ldots, P_i, \ldots$ terminates in a destination.

Proof. By Theorem 5, each term in this sequence is a pointer list. By Theorem 4, as long as such a pointer list has more than four terms, it is not a fixed point of the ciliate operations. By Theorem 6, the sequence does not terminate with an application of cds or of cdr, but with an application of E. Each application of E reduces the length of a pointer list that is not a fixed point for E by a positive even number of terms. According to the definitions of the ciliate operations, the pointers with absolute value λ and μ are never excised, and thus are present in any fixed point of a ciliate operation. Thus, a fixed point consisting of only two terms necessarily consists of the terms with absolute values λ and μ. As such, a two-term result is still a pointer list by Theorem 5. Stipulation (5) of Definition 1 shows that this fixed point must be $[\mu, \lambda]$ or $[-\lambda, -\mu]$. Since applications of cde remove terms that are equal and adjacent, a four-term fixed point must contain, in addition to terms with absolute values μ and λ, two terms of equal absolute value. If these two terms have opposite sign, the pointer list is not a fixed point for cdr. Thus, these two terms must be of the same sign. But then, as the pointer list is a fixed point of cde, these two terms are not adjacent. Moreover, their absolute value is strictly between μ and λ. Now stipulation (5) of Definition 1 implies that this pointer list is one of the two remaining claimed destinations. □

Thus the following algorithm, which we call the HNS algorithm, halts:

(1) Input: A pointer list P, its length $|P|$ and integers r and s;

(2) Iteratively apply cde until a cde fixed point is reached. With each application, decrease $|P|$ by 2. Then proceed to (3).

(3) If P is a fixed point of cds, proceed to (4). Else, apply cds, increase s by 1, and return to (1).

(4) If P is a fixed point of cdr, terminate the algorithm and report the current values of P, r and s. Else, apply cdr, increase r by 1, and return to (1).

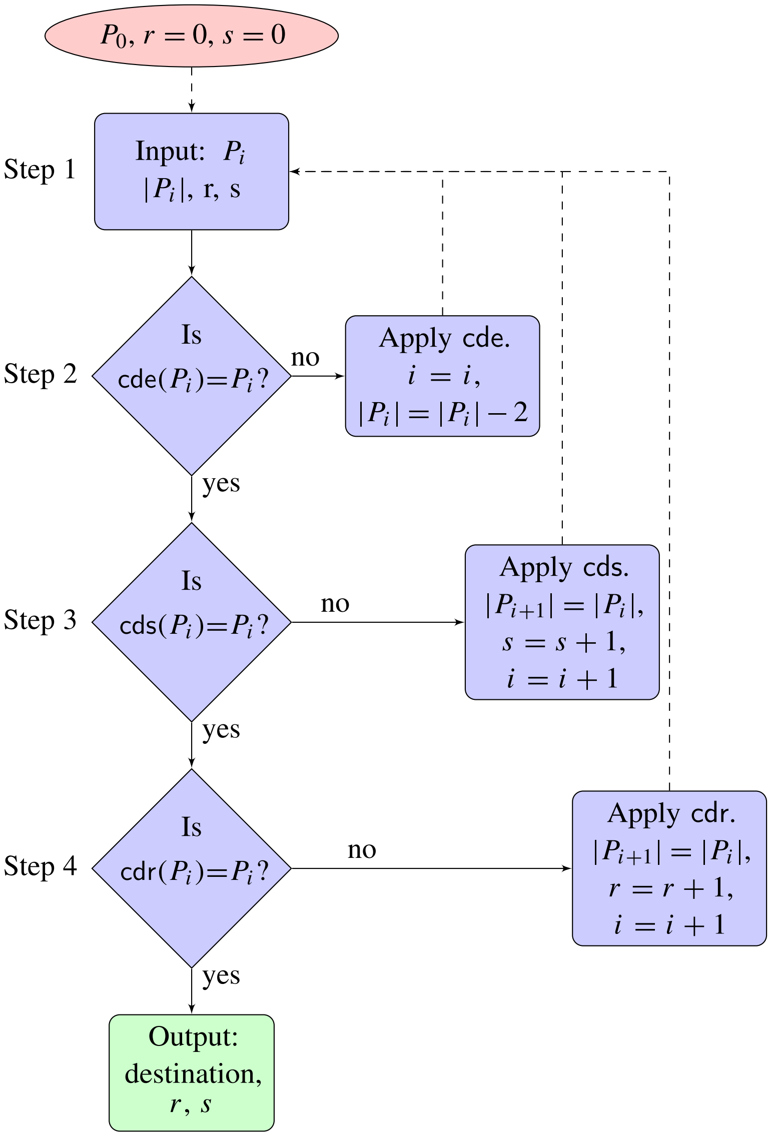


Figure 8. A flow diagram for the HNS algorithm.

Figure 8 depicts the algorithm in flow-diagram style. Let the original length of the pointer list P be denoted $|P|$.

In step (2), the algorithm examines $|P|-1$ adjacent pairs. If P is not a cde fixed point, then with the application of cde, $|P|$ decreases by 2. In this step we update the length of the resulting P with each nontrivial application of cde.

In step (3), the algorithm starts with a position $k < |P|$ and then chooses a position $\ell > k+1$ with $x_k = x_\ell$ if any. This takes at most $(|P|-1)+(|P|-2)+\cdots+2$ search steps, which is $O(|P|^2)$. If this search fails, proceed to step (4). Else, suppose a successful $k+1 < \ell < |P|$ is found. Then for $k < j < \ell$, search for an $m > \ell$ with $x_m = x_\ell$. This would require at most $(\ell - k)(|P| - \ell)$ steps. If this fails, proceed to step (4). Else, execute a cds based on the found quadruple (k, j, ℓ, m), increase s by 1, and return to step (1). Step (3) is completed in $O(|P|^2)$ search steps.

In step (4), the algorithm starts with a position $k < |P|$ and then scans positions $j > k$ until it finds an $x_j = -x_k$. The worst case scenario for this search is also $(|P|-1)+(|P|-2)+\cdots+2$, or $O(|P|^2)$. If the search succeeds, the result of cdr is obtained in at most $|P|-1$ search steps. Increase r by 1, and return to step (2). Else, if the search fails, terminate the algorithm and report the current values of P, r and s.

In one cycle of executing steps until return to step (1), the worst case scenario employs at most $O(|P|^2)$ search and execution steps. For the next round, an upper bound is$O((|P|-1)^2) = O(|P|^2)$. This continues for at most $|P|/2$ rounds. Thus a global upper bound, in terms of the length of the initial pointer list, is $O(|P|^3)$.

The efficiency of this algorithm that produces from an initial pointer list a fixed point for the operations cde, cds and cdr in $O(|P|^3)$ steps can probably be improved. Additionally, this algorithm most likely does not minimize the number of steps taken, using cde, cds and cdr, to reduce a pointer list to a fixed point.

In our phylogenetic application below, any calibration of time span in terms of the number of operations required is based on the above HNS algorithm as computational standard for the calibration.

6. An application to genome phylogenetics

As illustrated in Figure 1, for organisms S_1 and S_2 there may be synteny blocks of orthologous genes on corresponding chromosomes. Choose S_1 as reference and number the synteny blocks in their $5'$ to $3'$ order of appearance on S_1's chromosome as $1, 2, 3, \ldots, n$. In species S_2, the synteny blocks of these same genes may appear in a different order, and individual synteny blocks may also appear in orientation opposite from the orientation in S_1. Write the corresponding list of numbers in their order of appearance on S_2's chromosome, making the number negative if the synteny block orientation is opposite to that in S_1. The result is a signed permutation of the list $1, 2, 3, \ldots, n$.

Now imagine that the list of synteny blocks for S_1 are the MDSs of a ciliate macronuclear gene G, while the signed permutation that represents the corresponding list of synteny blocks for S_2 is the micronuclear precursor of G. Take the number of operations the ciliate decryptome performs to convert the micronuclear precursor to its macronuclear version G as a measure of the evolutionary distance between the two chromosomes of S_1 and S_2. We used the HNS algorithm to simulate the actions of the ciliate decryptome on the set of highly permuted genomes from various species of fruit flies.

The fruit fly genome is organized in four[10] chromosomes, enumerated 1, 2, 3 and 4. These four chromosomes are traditionally divided into six so-called Muller elements. The left and right arms of chromosome 2 are each one of these Muller elements, and it is similar for chromosome 3. Chromosome 1 is the X chromosome. The correspondence of chromosomal material to Muller elements is as follows:

chromosome	1 = X	2L	2R	3L	3R	4
Muller element	A	B	C	D	E	F

The fruit fly genome has at least 13,600 confirmed genes (and counting), but is not expected to host significantly more genes. Recall that our definition of a synteny block is more restrictive than the one used in [Bhutkar et al. 2008], where "microinversions" are permitted. See, for example, Table 1 on page 1662 of [loc. cit.] for data on these more relaxed synteny blocks relative to the genome of *D. melanogaster*. Between two species, the number of synteny blocks can still be well over a thousand, as can be gleaned from Table 1 of [loc. cit.], where the more relaxed definition of synteny block actually provides a lower bound on the number of synteny blocks as defined in our paper.

According to findings of [Bhutkar et al. 2008], 95% of orthologous genes between two species are present on the same Muller element. For the species we are using, with one exception to be noted now, evidence suggests that all orthologous genes are present on the same Muller elements. Using data obtained from flybase.org, we examined the permutation structure of these for the eight species *D. melanogaster*, *D. yakuba*, *D. erecta*, *D. sechellia*, *D. mojavensis*, *D. simulans*, *D. grimshawi* and *D. virilis*. As illustrated in Figure 3 of [Bhutkar et al. 2008], there is a translocation of genes between Muller elements B and C for *D. erecta*, one of the species in our sample. Thus we combined Muller elements B and C into one computational unit (chromosome 2) for our application. Thus, we refer to the five units A, B/C, D, E and F in the remainder of this discussion.

For each of the five units we computed, using in-house developed software written in Python, the number of applications of context-directed swaps or context-directed reversals performed by the HNS algorithm to permute the synteny block order of one species to produce the corresponding synteny block order of another species. This was done with each species considered as reference species. Since HNS gives preference to block interchanges, the number of reversals in our derived data is low.

Note that although we used the full gene lists from flybase.org, using pointer lists and ciliate operations automatically reduces to performing ciliate sorting operations on synteny blocks between pairs of species.

[10] There are exceptions: see, for example, Figure 1 of [Schaeffer et al. 2008]. None of the exceptional species is considered in our paper.

From our data about the number of context-directed swaps, s, and reversals, r, we define a corresponding distance matrix by using the formula $s + r/2$. As the reader would observe from examining our data, this in fact does define a metric[11].

Then we applied the unweighted pair group method with arithmetic mean, also known as the UPGMA algorithm[12], to these metrics. We used an in-house developed MAPLE implementation of UPGMA to compute these phylogenies.

The Appendix contains the data, derived distance matrices and corresponding phylogenetic trees for the five units in Figures 9, 10, 11, 12 and 13. An entry in the format "$r : s$" in row i and column j of a table is interpreted as follows: r denotes the number of context-directed reversals (cdr operations), while s denotes the number of context-directed block interchanges (cds operations) executed by the HNS algorithm to convert the permutation of the species in row i to that of the species in column j. Thus the species in column j is the *reference species*. The total for whole genomes is given in Figure 14.

We used the timeline given in Figures 1 and 3 of [Hahn et al. 2007] to calibrate the timeline in our phylogenetic trees[13]. This calibration is a rough timeline: Our work describes evolutionary relationships among instances of a specific chromosome present in these eight species. The evolutionary timeline for a chromosome need not agree with the evolutionary timeline for speciation. According to Figures 1 and 3 of [Hahn et al. 2007], the time span from the earliest common ancestor of our species is roughly 60 million years.

Discussion

Comparison of our results in the Appendix and the results of [Bhutkar et al. 2008, Table 2] show a significant difference in the number of sorting operations, with ours typically higher. One reason for these differences lies in our definition of synteny blocks: We allow blocks consisting of a single gene, and we do not allow blocks containing different gene orders. Thus, we have a larger number of synteny blocks to be sorted, and our computations took into account all orthologous genes. This point is illustrated by comparing the number of synteny blocks for Muller element E

[11] There are strong grounds for equating the value of two reversals with that of a single swap. As computations show, the result (given in the Appendix) is a matrix that is symmetric over its diagonal. It is also evident that the number of sorting operations to sort permutation α to obtain permutation β, plus the number of sorting operations to sort permutation β to permutation γ, is no smaller than the number of sorting operations to directly sort permutation α to permutation γ. Thus, the triangle inequality holds.

[12] This is Algorithm 4.1 in [Clote and Backofen 2000]. A good exposition is also given in Chapter 27 of [Barton et al. 2007], available online at www.evolution-textbook.org.

[13] We could have used alternative timelines, such as the timelines given in the figure at the DroSpeGe website http://insects.eugenes.org/DroSpeGe/. Whichever published timeline one chooses will determine the corresponding calibration applied to our data.

computational unit	cso
A	266.75
B/C	207.75
D	247.25
E	364.25
F	8.5

Table 1. Ciliate sorting operations since most recent common ancestor of all considered species.

for *D. yakuba*, *D. sechellia* and *D. simulans* (computed relative to *D. melanogaster*) reported in [loc. cit., Table 5] with the actual number of sorting operations reported for these species (with *D. melanogaster* as reference) in our Figure 12. Moreover, whereas in [loc. cit.] the authors used unconstrained reversals as sorting operation, we used context-directed reversals. Additionally, in [loc. cit.] genes that suggest that a transposition is responsible for the rearrangement were excluded from the analysis. We included all orthologous genes since the sorting operation of context-directed swaps (block interchanges) accounts also for transpositions.

Comparison of the phylogenies in the Appendix with the phylogeny in [loc. cit., Figure 8] or with the phylogeny of sequenced species at flybase.org[14] indicate that our placement of *D. sechellia* is in all cases quite different. The placement of *D. mojavensis, D. virilis* and *D. grimshawi* relative to each other and to the other species agrees with both of these phylogenies for all but Muller elements A and E.

By using the UPGMA algorithm to construct phylogenies from distance matrices, we assumed a uniform rate of evolution for the Muller elements. Comparing these uniform rates among the different chromosomes indicates that no two individual chromosomes undergo permutations at the same rates. Our sorting data suggests the upper bounds in Table 1 on the number of ciliate sorting operations (cso) since the most recent common ancestor of all the species considered.

These numbers were computed by taking the largest ciliate sorting distance achieved between a pair of the considered species, and dividing[15] by 2 to obtain an estimate of the number of ciliate sorting operations to each species' corresponding genomic element since their most recent common ancestor.

The Muller F element has undergone remarkably few permutations in comparison with the other Muller elements. Muller element E appears to be the most susceptible to permutation, while Muller element F appears the most "resistant" to permutation. This, however, may be a biased view of susceptibility to permutation since these computational units do not harbor the same number of genes or synteny blocks. As

[14] http://flybase.org/static_pages/species/sequenced_species.html

[15] Using our hypothesis of uniform rate of evolution.

indicated in [Hochman 1971], chromosome 4 (Muller element F) is generally a very small chromosome: it may contain fewer than 100 genes (see, for example, the results regarding Muller element F for various species in [Schaeffer et al. 2008]). The other Muller elements each contains well over 1000 genes. Thus one would expect the number of rearrangements needed to sort one species' chromosome 4 gene content to that of another species to be relatively low in comparison with the other, larger, chromosomes.

Tables 5 and 6 of [Bhutkar et al. 2008] report rearrangement rates that are computed from the number of synteny blocks relative to *D. melanogaster*, the nucleotide length of the Muller element, and the estimated divergence time for the species in question. These rates assume that arbitrary reversals cause the rearrangements and thus ignore genes deemed to have been moved by other sorting mechanisms, and use a definition of synteny block that ignores certain rearrangements. In the case of our context-directed sorting operations, a more appropriate measure of "susceptibility to permutation" should probably take into account additional parameters regarding nucleotide patterns in the Muller elements. Progress in this regard would address the third[16] and fourth[17] questions raised in [Schaeffer et al. 2008, pp. 1603–1604], phrased for arbitrary reversals, and may also indicate whether context-directed reversals and block interchanges are more suitable sorting operations for phylogenetic analyses based on permutations of genomic material. Such rearrangement rates may be used as "susceptibility coefficients", measuring the susceptibility of a genomic element to rearrangement.

According to [Bhutkar et al. 2008, Figure 3], the F element of *D. willistoni* (which is not among the species we considered) has been absorbed in the E-element of *D. willistoni*. It would be interesting to "distill" the *D. willistoni* F-element from the *D. willistoni* E-element, and compare its level of permutation relative to the F-element of the eight species in our study. Establishing susceptibility coefficients may enable us to obtain from the current permutation state of the distilled *D. willistoni* F-element, and established evolutionary time distances for the fruit fly phylogeny, an estimate of when absorption of the F-element into the E-element took place.

Similarly, by separating the treatment of the B and C elements, calculating the corresponding susceptibility coefficients of these elements, and distilling the B-element components and the C-element components for *D. ananassae*, one may be able to estimate when these transpositions occurred. Figure 3 of [loc. cit.] also indicates that part of *D. pseudoobscura*'s Muller A element was transposed to its Muller E element. Susceptibility coefficients may be useful in estimating when

[16] "...how do new inversions originate?" This can be expanded to include the question of how new block interchanges originate.

[17] "...what is the molecular basis for gene arrangement polymorphism?"

this transposition occurred. An investigation of the structural properties of the chromosomes involved in these interchromosomal translocations may also reveal if any DNA motifs promote these translocations.

The differences in phylogenies for different chromosomal domains in the considered species suggest the possibility of inferring from Mendelian inheritance hypotheses and diploidy of the fruit fly genomes, interbreeding among ancestor species that would produce the observed chromosomal configurations.

We relied on the UPGMA algorithm for constructing our phylogenies. Other clustering techniques such as neighbor joining, or several other algorithms, as, for example, in [Clote and Backofen 2000], may reveal finer details than the technique applied here.

While using ciliate operations to compute the permutation-based distances between pairs of species, we found permutations which are not reducible to each other by ciliate operations. In contrast to the case for unrestricted block interchanges and unrestricted reversals, not all permutations are invertible by context-directed block interchanges and reversals. When our algorithm terminates with a destination of length 4 instead of 2, this indicates that the two permutations involved in the distance measure require an additional transposition to complete the transformation. Though we have not done so in our current paper, the fact of noninvertibility by ciliate decryptome operations could be taken as an additional parameter in measuring evolutionary distance. Instead, in this paper we counted this additional transposition needed at the end as a single step towards the distance. An argument can be made that the necessity of this additional transposition should be accounted for more significantly in computing evolutionary distance. It also raises the question of determining an easily applicable characterization of permutations that are invertible by constrained block interchanges or reversals. The problem of mathematically characterizing permutations that are invertible by context-directed operations has been solved in subsequent work [Adamyk et al. $\geq$ 2016].

Finally, although the HNS algorithm finds in polynomial time the data needed to construct a distance matrix, we do not propose that this algorithm finds optimal data in the following sense: when one permutation can be transformed to another by means of context-directed reversals and block interchanges, what is the least number of these operations needed for such a transformation? The answer for context-directed block interchanges has been obtained in [Adamyk et al. $\geq$ 2016]. The minimal number of operations may depend on strategic sorting decisions made while sorting a permutation. One may inquire whether certain permutations require less strategic decision making in order to obtain a successful sorting. The permutations requiring the least number of strategic decisions for context-directed block interchanges have been characterized in [Anderson et al. $\geq$ 2016], but a complete answer is currently not known.

Appendix: The distance matrices underlying the application of UPGMA to the five chromosomes of eight fruit fly species

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir		32:431	38:463	31:438	33:403	40:426	29:414	35:514
D. gri	26:434		36:446	35:430	36:381	45:404	40:391	45:504
D. sim	36:464	34:447		35:460	8:268	19:311	21:282	26:505
D. moj	29:439	37:429	41:457		41:407	34:434	36:422	37:515
D. mel	37:401	40:379	6:269	45:405		1:171	35:93	19:482
D. ere	36:428	43:405	29:306	42:430	3:170		25:182	28:499
D. yak	43:407	40:391	31:277	50:415	11:105	25:182		29:481
D. sec	43:510	39:507	20:508	39:514	7:488	22:502	17:487	

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir	0.0	447.0	482.0	453.5	419.5	446.0	428.5	531.5
D. gri	447.0	0.0	464.0	447.5	399.0	426.5	411.0	526.5
D. sim	482.0	464.0	0.0	477.5	272.0	320.5	292.5	518.0
D. moj	453.5	447.5	477.5	0.0	427.5	451.0	440.0	533.5
D. mel	419.5	399.0	272.0	427.5	0.0	171.5	110.5	491.5
D. ere	446.0	426.5	320.5	451.0	171.5	0.0	194.5	513.0
D. yak	428.5	411.0	292.5	440.0	110.5	194.5	0.0	495.5
D. sec	531.5	526.5	518.0	533.5	491.5	513.0	495.5	0.0

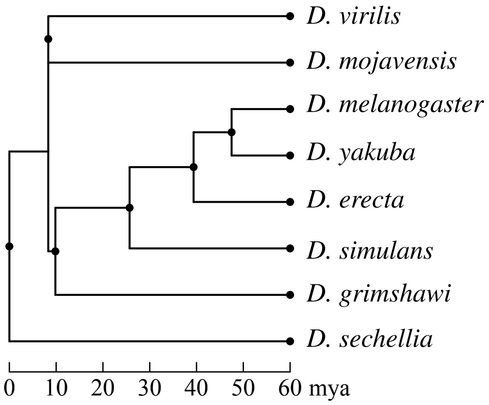


Figure 9. Data, distance matrix and resulting phylogeny for the Muller A-element.

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir		62:255	44:290	31:161	41:262	51:324	63:332	38:375
D. gri	56:258		43:318	38:187	52:280	62:342	50:370	45:393
D. sim	48:288	49:315		53:205	2: 95	24:188	16:223	9:256
D. moj	67:143	32:190	55:204		49:173	51:254	48:285	47:319
D. mel	59:253	58:277	8: 92	49:173		19:159	101:145	3:229
D. ere	45:327	44:351	14:193	57:251	11:163		9:249	32:275
D. yak	49:339	42:374	14:224	44:287	7:192	15:246		34:286
D. sec	52:368	49:391	7:257	41:322	3:229	38:272	46:280	

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir		286	312	176.5	282.5	349.5	363.5	394
D. gri	286		339.5	206	306	373	395	415.5
D. sim	312	339.5		231.5	96	200	331	260.5
D. moj	176.5	206	231.5		197.5	279.5	309	342.5
D. mel	282.5	306	96	197.5		168.5	195.5	230.5
D. ere	349.5	373	200	279.5	168.5		253.5	291
D. yak	363.5	395	331	309	195.5	253.5		303
D. sec	394	415.5	260.5	342.5	230.5	291	303	

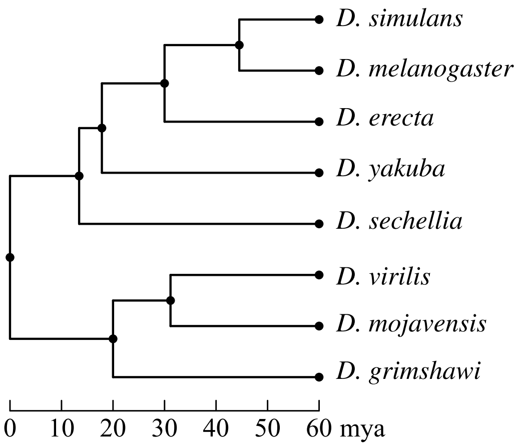


Figure 10. Data, distance matrix and resulting phylogeny for the Muller B/C-element.

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir		21:124	60:193	27:113	69:175	58:160	53:231	60:450
D. gri	27:121		51:210	29:154	52:187	56:174	56:244	59:460
D. sim	68:189	59:206		65:219	2: 69	5: 56	10:129	2:390
D. moj	23:115	23:157	59:222		53:214	59:192	55:257	51:469
D. mel	69:175	62:182	2: 69	81:200		8: 35	10:109	0:388
D. ere	66:156	58:173	7: 55	67:188	10: 34		12: 79	90:337
D. yak	59:228	64:240	26:121	71:249	14:107	18: 76		12:416
D. sec	54:453	55:462	2:390	49:470	0:388	4:380	12:416	

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir	0	134.5	223	126.5	209.5	189	257.5	480
D. gri	134.5	0	235.5	168.5	213	202	272	489.5
D. sim	223	235.5	0	251.5	70	58.5	134	391
D. moj	126.5	168.5	251.5	0	240.5	221.5	284.5	494.5
D. mel	209.5	213	70	240.5	0	39	114	388
D. ere	189	202	58.5	221.5	39	0	85	382
D. yak	257.5	272	134	284.5	114	85	0	422
D. sec	480	489.5	391	494.5	388	382	422	0

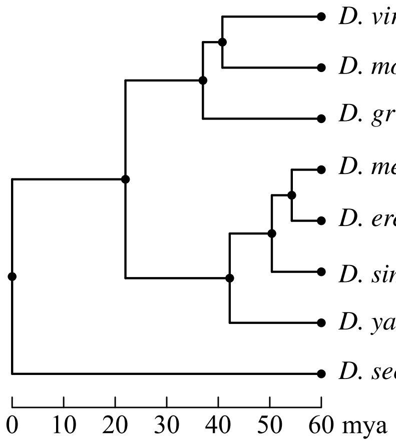


Figure 11. Data, distance matrix and resulting phylogeny for the Muller D-element.

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir		47:634	40:451	27:340	40:432	46:551	42:436	41:598
D. gri	47:634		47:616	25:549	46:602	55:664	54:603	47:705
D. sim	52:445	57:611		45:213	8: 71	142:241	13: 75	14:347
D. moj	89:309	53:535	39:216		39:185	43:401	31:194	45:446
D. mel	44:430	54:598	8: 71	39:185		196:196	7: 38	21:334
D. ere	50:549	55:664	10:307	39:403	6:291		12:291	38:428
D. yak	54:430	62:599	19: 72	43:188	15: 34	8:293		23:334
D. sec	51:593	53:702	14:347	39:449	5:342	38:428	9:341	

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir		657.5	471	353.5	452	574	457	618.5
D. gri	657.5		639.5	561.5	625	691.5	630	728.5
D. sim	471	639.5		235.5	75	312	81.5	354
D. moj	353.5	561.5	235.5		204.5	422.5	209.5	468.5
D. mel	452	625	75	204.5		294	41.5	344.5
D. ere	574	691.5	312	422.5	294		297	447
D. yak	457	630	81.5	209.5	41.5	297		345.5
D. sec	618.5	728.5	354	468.5	344.5	447	345.5	

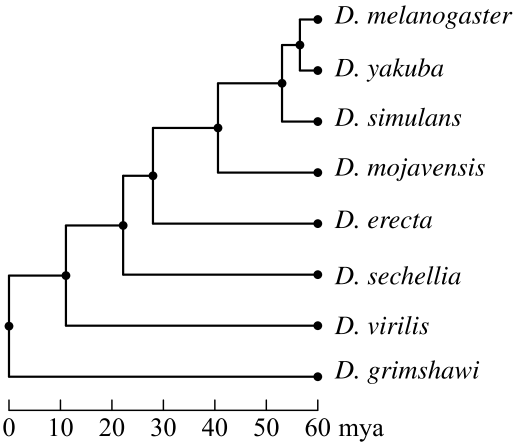


Figure 12. Data, distance matrix and resulting phylogeny for the Muller E-element.

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir		3:5	12:8	2:1	11:6	11:6	11:6	8:13
D. gri	3:5		10:12	3:4	11:9	11:9	11:9	10:12
D. sim	8:10	8:13		10:9	4:5	4:5	4:5	4:13
D. moj	2:1	3:4	6:11		7:7	7:7	7:7	9:12
D. mel	9:7	7:11	6:4	7:7		0:0	0:0	0:12
D. ere	9:7	11:9	6:4	9:6	0:0		0:0	0:12
D. yak	9:7	7:11	6:4	7:7	0:0	0:0		0:12
D. sec	8:13	8:13	2:14	9:12	0:12	0:12	0:12	

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir		6.5	14	2	11.5	11.5	11.5	17
D. gri	6.5		17	5.5	14.5	14.5	14.5	17
D. sim	14	17		14	7	7	7	15
D. moj	2	5.5	14		10.5	10.5	10.5	16.5
D. mel	11.5	14.5	7	10.5		0	0	12
D. ere	11.5	14.5	7	10.5	0		0	12
D. yak	11.5	14.5	7	10.5	0	0		12
D. sec	17	17	15	16.5	12	12	12	

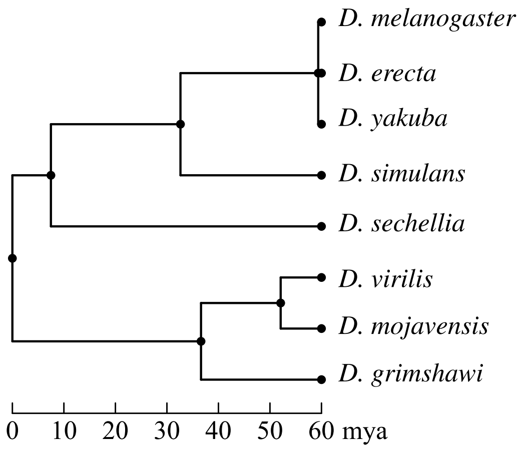


Figure 13. Data, distance matrix and resulting phylogeny for the Muller F-element.

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir		165:1449	194:1405	118:1053	194:1278	206:1467	198:1419	182:1950
D. gri	165:1449		187:1602	130:1324	197:1459	229:1593	211:1617	206:2074
D. sim	194:1405	187:1602		208:1106	24:508	194:801	64:722	55:1511
D. moj	118:1053	130:1324	208:1106		189:986	194:1288	177:1165	189:1761
D. mel	194:1278	197:1459	24:508	189:986		224:561	153:385	43:1445
D. ere	206:1467	229:1593	194:801	194:1288	224:561		58:801	188:1551
D. yak	198:1419	211:1617	64:722	177:1165	153:385	58:801		98:1529
D. sec	182:1950	206:2074	55:1511	189:1761	43:1445	188:1551	98:1529	

	D. vir	*D. gri*	*D. sim*	*D. moj*	*D. mel*	*D. ere*	*D. yak*	*D. sec*
D. vir	0	1231.5	1502	1112	1375	1570	1518	2041
D. gri	1231.5	0	1695.5	1389	1557.5	1707.5	1722.5	2177
D. sim	1502	1695.5	0	1210	520	898	746	1538.5
D. moj	1112	1389	1210	0	1080.5	1385	1253.5	1655.5
D. mel	1375	1557.5	520	1080.5	0	673	461.5	1463.5
D. ere	1570	1707.5	898	1385	673	0	830	1645
D. yak	1518	1722.5	746	1253.5	461.5	830	0	1578
D. sec	2041	2177	1538.5	1655.5	1463.5	1645	1578	0

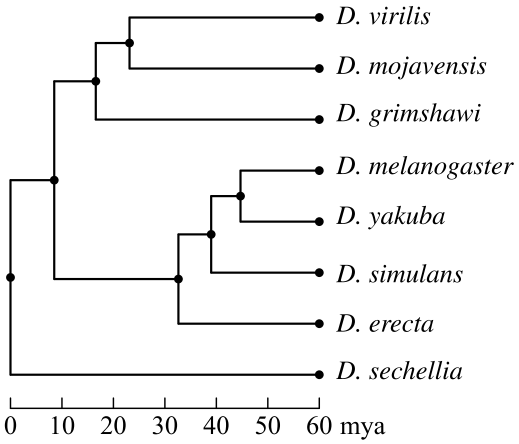


Figure 14. Data, distance matrix and resulting phylogeny for the whole genome.

Acknowledgement

We gratefully acknowledge that the advice of a very careful referee helped to greatly improve the readability of this paper.

References

[Adamyk et al. $\geq$ 2016] K. Adamyk, E. Holmes, G. Mayfield, D. J. Moritz, and M. Scheepers, "Games, genomes and graphs", preprint.

[Anderson et al. $\geq$ 2016] C. Anderson, M. Scheepers, M. Warner, and H. Wauck, "On permutations optimized for sorting by ciliate operations", preprint.

[Angeleska et al. 2007] A. Angeleska, N. Jonoska, M. Saito, and L. F. Landweber, "RNA-guided DNA assembly", *J. Theoret. Biol.* **248**:4 (2007), 706–720. MR 2899092

[Bafna and Pevzner 1995] V. Bafna and P. Pevzner, "Sorting by reversals: genome rearrangements in plant organelles and evolutionary history of the X chromosome", *Mol. Biol. Evol.* **12**:2 (1995), 239–246.

[Bafna and Pevzner 1998] V. Bafna and P. A. Pevzner, "Sorting by transpositions", *SIAM J. Discrete Math.* **11**:2 (1998), 224–240. MR 99e:05002 Zbl 0973.92014

[Barton et al. 2007] N. H. Barton, D. E. G. Briggs, J. A. Eisen, D. B. Goldstein, and N. H. Patel, *Evolution*, Cold Spring Harbor Laboratory Press, 2007.

[Bhutkar et al. 2008] A. Bhutkar, S. W. Schaeffer, S. M. Russo, M. Xu, T. F. Smith, and W. M. Gelbart, "Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes", *Genetics* **179**:3 (2008), 1657–1680.

[Clote and Backofen 2000] P. Clote and R. Backofen, *Computational molecular biology: an introduction*, John Wiley & Sons, Ltd., Chichester, 2000. MR 2002h:92021 Zbl 0955.92013

[Coghran and Wolfe 2002] A. Coghran and K. H. Wolfe, "Fourfold faster rate of genome rearrangement in nematodes than in *Drosophila*", *Genome Res.* **12**:6 (2002), 857–867.

[Dobzhansky and Sturtevant 1938] T. Dobzhansky and A. H. Sturtevant, "Inversions in the chromosomes of *Drosophila pseudoobscura*", *Genetics* **23**:1 (1938), 28–64.

[Ehrenfeucht et al. 2004] A. Ehrenfeucht, T. Harju, I. Petre, D. M. Prescott, and G. Rozenberg, *Computation in living cells: gene assembly in ciliates*, Springer, Berlin, 2004. Zbl 1069.68048

[Hahn et al. 2007] M. Hahn, M. Han, and S.-G. Han, "Gene family evolution across 12 *Drosophila* genomes", *PLOS Genetics* **3**:11 (2007), 2135–2146.

[Hannenhalli and Pevzner 1999] S. Hannenhalli and P. A. Pevzner, "Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals", *J. ACM* **46**:1 (1999), 1–27. MR 2000j:92013 Zbl 1064.92510

[Hochman 1971] B. Hochman, "Analysis of chromosome 4 in *Drosophila melanogaster* II: ethyl methanesulfonate induced lethals", *Genetics* **67**:2 (1971), 235–252.

[Hughes 2000] D. Hughes, "Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes", *Genome Biology* **1**:6 (2000), 1–8.

[Mira and Meidanis 2007] C. Mira and J. Meidanis, "Sorting by block-interchanges and signed reversals", pp. 670–676 in *Proceedings of the Fourth International Conference on Information Technology: New Generations* (Las Vegas, NV, 2007), edited by S. Latifi, IEEE Computer Society, Los Alamitos, CA, 2007.

[Möllenbeck et al. 2008] M. Möllenbeck, Y. Zhou, A. R. O. Cavalcanti, F. Jönsson, B. P. Higgins, W.-J. Chang, S. Juranek, T. G. Doak, G. Rozenberg, H. J. Lipps, and L. F. Landweber, "The pathway to detangle a scrambled gene", *PLOS One* **3**:6 (2008), e2330.

[Muller 1940] H. J. Muller, "Bearings of the *Drosophila* work on systematics", pp. 185–268 in *The New Systematics*, edited by J. Huxley, Clarendon Press, Oxford, 1940.

[Nowacki et al. 2007] M. Nowacki, V. Vijayan, Y. Zhou, K. Schotanus, T. Doak, and L. Landweber, "RNA-mediated epigenetic programming of a genome-rearrangement pathway", *Nature* **451**:7175 (2007), 153–158.

[Pevzner and Tesler 2003] P. A. Pevzner and G. Tesler, "Genome rearrangements in mammalian evolution: lessons from human and mouse genomes", *Genome Res.* **13**:1 (2003), 37–45.

[Prescott 1994] D. M. Prescott, "The DNA of ciliated protozoa", *Microbiol. Rev.* **58**:2 (1994), 233–267.

[Prescott 2000] D. M. Prescott, "Genome gymnastics: unique modes of DNA evolution and processing in ciliates", *Nature Reviews Genetics* **1** (2000), 191–198.

[Prescott et al. 2003] D. M. Prescott, A. Ehrenfeucht, and G. Rozenberg, "Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates", *J. Theoret. Biol.* **222**:3 (2003), 323–330. MR 2067536

[Schaeffer et al. 2008] S. W. Schaeffer, A. Bhutkar, B. F. McAllister, M. Matsuda, L. M. Matzkin, P. M. O'Grady, C. Rohde, V. L. S. Valente, M. Aguadé, W. W. Anderson, K. Edwards, A. C. L. Garcia, J. Goodman, J. Hartigan, E. Kataoka, R. T. Lapoint, E. R. Lozovsky, C. A. Machado, M. A. F. Noor, M. Papaceit, L. K. Reed, S. Richards, T. T. Rieger, S. M. Russo, H. Sato, C. Segarra, C. R. Smith, T. F. Smith, V. Strelets, Y. N. Tobari, Y. Tomimura, M. Wasserman, T. Watts, R. Wilson, K. Yoshida, T. A. Markow, W. M. Gelbart, and T. C. Kaufman, "Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps", *Genetics* **179**:3 (2008), 1601–1655.

[Sturtevant and Dobzhansky 1936] A. H. Sturtevant and T. Dobzhansky, "Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use in the study of the history of the species", *PNAS* **22**:7 (1936), 448–450.

[Yancopoulos et al. 2005] S. Yancopoulos, O. Attie, and R. Friedberg, "Efficient sorting of genomic permutations by translocation, inversion and block interchange", *Bioinformatics* **21**:16 (2005), 3340–3346.

Received: 2013-01-23 Revised: 2014-12-11 Accepted: 2014-12-21

jlherlin@indiana.edu
Department of Mathematics, Indiana University, Rawles Hall,
831 East 3rd Street, Bloomington, IN 47405, United States

anelson@math.utah.edu
Department of Mathematics, University of Utah, 155 S 1400 E,
Room 233, Salt Lake City, UT 84112-0090, United States

mscheepe@boisestate.edu
Department of Mathematics, Boise State University,
Boise, ID 83725, United States

INVOLVE 9:1 (2016)
dx.doi.org/10.2140/involve.2016.9.27

On the distribution of the greatest common divisor of Gaussian integers

Tai-Danae Bradley, Yin Choi Cheng and Yan Fei Luo

(Communicated by Kenneth S. Berenhaut)

For a pair of random Gaussian integers chosen uniformly and independently from the set of Gaussian integers of norm x or less as x goes to infinity, we find asymptotics for the average norm of their greatest common divisor, with explicit error terms. We also present results for higher moments along with computational data which support the results for the second, third, fourth, and fifth moments. The analogous question for integers is studied by Diaconis and Erdős.

1. Introduction

In this paper, we study questions related to the size of the greatest common divisor of pairs of randomly chosen Gaussian integers. In particular, in Theorem 1, we first calculate the probability that a pair of random Gaussian integers, chosen uniformly and independently from the set of all Gaussian integers with norm x or less, has greatest common divisor $\pm\kappa$ or $\pm i\kappa$ for a fixed Gaussian integer κ. The main term for this probability in the case where $\kappa=1$ was first given by Collins and Johnson [1989, Theorem 8]. We refine their results by providing the expression for the more general case in addition to giving an explicit error term for all cases. In Theorem 2 we derive the expected norm of the greatest common divisor between a pair of Gaussian integers with norm x or less. Finally, in Theorem 3 and Conjecture 4, we present an expression for higher moments of the norm of the greatest common divisor between a pair of Gaussian integers with norm x or less. We expect our results to generalize to principal ideal domains without too much difficulty. More generally, our results should hold for the ring of integers in an algebraic number field, though our techniques will need to be modified to deal with class number greater than one and infinite unit group. We expect the ideas in [Micheli and Ferraguti 2015] could help address this question and would be an interesting direction to explore further. Of further interest are function field analogues. Some interesting results in this direction may be found in [Micheli and Schnyder 2015].

MSC2010: 11N37, 11A05, 11K65, 60E05.
Keywords: Gaussian integer, gcd, moment, Dedekind zeta function.

Similar questions have also been studied for the case of rational integers. Originally, Mertens [1874] proved that the probability that a pair of rational integers chosen uniformly and independently at random from $\{1, 2, \dots, x\}$ are relatively prime is asymptotic to $1/\zeta(2)$, as x tends to infinity, where ζ is the Riemann zeta function. Christopher [1956, Theorem 1] generalized Mertens' result by finding the probability that two integers have greatest common divisor k for a fixed k larger than 1. An asymptotic expression for the moments of the greatest common divisor was first derived by Cesàro [1885], and Diaconis and Erdős [2004, Theorem 2] later extended his work by explicitly calculating the error term. In particular, the expected value for the greatest common divisor between a pair of random integers chosen independently and uniformly from the set $\{1, 2, \dots, x\}$ is

$$\frac{1}{\zeta(2)} \log x + O(1), \tag{1}$$

while the n-th moment is given by

$$\frac{x^{n-1}}{n+1}\left(\frac{2\zeta(n)}{\zeta(n+1)} - 1\right) + O(x^{n-2}\log x) \qquad \text{for } n \geq 2. \tag{2}$$

The goal of the present paper is to show that (1) has an analogous counterpart in the ring of Gaussian integers as stated in Theorem 2 at the end of this section. Further, we show that (2) also has an analogous form as presented in Theorem 3 and Conjecture 4. Before proceeding, we first give the following preliminary definitions and remark.

Definition. The norm of a Gaussian integer $\alpha = a + bi$ for rational integers a and b is defined by $N(\alpha) = a^2 + b^2$.

Most of our results will be in terms of the norms of Gaussian integers and not the integers themselves.

Remark. Given two Gaussian integers η and μ, a greatest common divisor, denoted (η, μ), is defined to be a Gaussian integer κ such that κ is a divisor of both η and μ, and if there is any other common factor between η and μ, then it must also be a factor of κ. From this definition, it becomes clear that (η, μ) is unique only up to its associates. In other words, $(\eta, \mu) = \kappa$, $-\kappa$, $i\kappa$, and $-i\kappa$. Our calculations, however, will be performed via ideals for reasons that will soon become apparent. For a Gaussian integer η, we say $\mathfrak{n}$ is the ideal such that

$$\mathfrak{n} = (\eta) = (-\eta) = (i\eta) = (-i\eta),$$

and the norm of $\mathfrak{n}$ is defined by $N(\mathfrak{n}) = N(\eta)$. Accordingly, the definition of the greatest common divisor for a pair of ideals is this:

Definition (greatest common divisor of two ideals). For a ring R, let $\mathfrak{n}, \mathfrak{m} \subset R$ be ideals. The greatest common divisor $(\mathfrak{n}, \mathfrak{m})$ is defined to be the ideal $\mathfrak{K} \subset R$ which satisfies the following:

(1) $\mathfrak{n} \subset \mathfrak{K}$ and $\mathfrak{m} \subset \mathfrak{K}$.

(2) If there exists some ideal $\mathfrak{a} \subset R$ such that $\mathfrak{n} \subset \mathfrak{a}$ and $\mathfrak{m} \subset \mathfrak{a}$, then $\mathfrak{K} \subset \mathfrak{a}$.

In other words, $(\mathfrak{n}, \mathfrak{m})$ is the smallest ideal that contains all the elements of both $\mathfrak{n}$ and $\mathfrak{m}$. When applied to the ring of Gaussian integers, a Dedekind domain, it is clear that $(\mathfrak{n}, \mathfrak{m})$ is unique.

Definition (the Dedekind zeta function of $\mathbb{Q}(i)$). For the number field $\mathbb{Q}(i)$, the complex-valued Dedekind zeta function is defined for $\mathrm{Re}(s) > 1$ by

$$\zeta_{\mathbb{Q}(i)}(s) = \sum_{\mathfrak{a} \subset \mathbb{Z}[i]} \frac{1}{N(\mathfrak{a})^s} = \frac{1}{4} \sum_{\substack{(a,b)\in\mathbb{Z} \\ (a,b)\neq(0,0)}} \frac{1}{(a^2+b^2)^s},$$

where the first summation is over the nonzero ideals $\mathfrak{a}$ of the ring of Gaussian integers $\mathbb{Z}[i]$.

In order to find the expression for the expected norm of a greatest common divisor between a pair of Gaussian integers of norm x or less, we will first derive the necessary probability distribution function of Theorem 1:

Theorem 1. *Let $\mathfrak{n}$ and $\mathfrak{m}$ be nonzero ideals chosen independently and uniformly at random from the set of ideals in $\mathbb{Z}[i]$ with norm x or less. The probability that $(\mathfrak{n}, \mathfrak{m}) = \mathfrak{K}$ is*

$$\frac{1}{\zeta_{\mathbb{Q}(i)}(2) N(\mathfrak{K})^2} + O\left(\frac{1}{x^{2/3} N(\mathfrak{K})^{4/3}}\right).$$

This probability will allow us to calculate the expected norm of the greatest common divisor between a pair of ideals:

Theorem 2. *Let $\mathfrak{n}$ and $\mathfrak{m}$ be nonzero ideals chosen independently and uniformly at random from the set of ideals in $\mathbb{Z}[i]$ with norm x or less. The expected norm of the greatest common divisor of $\mathfrak{n}$ and $\mathfrak{m}$ is*

$$\frac{\pi}{4\zeta_{\mathbb{Q}(i)}(2)} \log x + O(1).$$

We will then prove the following result regarding the n-th moment for $n > 2$:

Theorem 3. *Let $\mathfrak{n}$ and $\mathfrak{m}$ be nonzero ideals chosen independently and uniformly at random from the set of ideals in $\mathbb{Z}[i]$ with norm x or less. For $n > 2$, there exists a constant $c_n \in \mathbb{R}$ such that*

$$E_x\{N(\mathfrak{n}, \mathfrak{m})^n\} \sim c_n x^{n-1},$$

where $E_x\{N(\mathfrak{n}, \mathfrak{m})^n\}$ denotes the n-th moment of the norm of the greatest common divisor of $\mathfrak{n}$ and $\mathfrak{m}$.

Lastly, we will present numerical data which provide strong evidence for the following conjecture regarding the constant of Theorem 3 for all $n \geq 2$:

Conjecture 4. *For $n \geq 2$,*

$$E_x\{N(\mathfrak{n},\mathfrak{m})^n\} \sim \frac{4}{\pi(n+1)}\left(\frac{2\zeta_{\mathbb{Q}(i)}(n)}{\zeta_{\mathbb{Q}(i)}(n+1)} - 1\right)x^{n-1}.$$

The proof of Theorem 1 will be given in Section 2 and that of Theorem 2 will be given in Section 3. Finally, in Section 4, we prove Theorem 3 and present Conjecture 4 along with computational data which support the conjecture for the second, third, fourth, and fifth moments.

2. Probability distribution function

Before deriving the expression for the probability of Theorem 1, we first define the following two functions:

Definition (the Möbius function). For an ideal $\mathfrak{n}$, the Möbius Function $\mu(\mathfrak{n})$ is defined by

$$\mu(\mathfrak{n}) = \begin{cases} 1 & \text{if } \mathfrak{n} = (1), \\ (-1)^t & \text{if } \mathfrak{n} = \mathfrak{p}_1\mathfrak{p}_2\cdots\mathfrak{p}_t \text{ for distinct prime ideals } \mathfrak{p}_i, \\ 0 & \text{otherwise.} \end{cases}$$

We will use the following identity

$$\sum_{\mathfrak{d}|\mathfrak{n}} \mu(\mathfrak{d}) = \begin{cases} 1 & \text{if } \mathfrak{n} = (1), \\ 0 & \text{if } \mathfrak{n} \neq (1), \end{cases} \tag{3}$$

as well as the generating function

$$\sum_{\mathfrak{n}\subset\mathbb{Z}[i]} \frac{\mu(\mathfrak{n})}{N(\mathfrak{n})^s} = \frac{1}{\zeta_{\mathbb{Q}(i)}(s)} \quad \text{for } \operatorname{Re}(s) > 1.$$

Definition (the sum-of-two-squares function). For $n \in \mathbb{Z}$, let the sum-of-two-squares function $r(n,2)$ represent the number of ways that n can be expressed as a sum of two squares. Thus,

$$r(n,2) = \tfrac{1}{4}\#\{\mathfrak{a} \subset \mathbb{Z}[i] : N(\mathfrak{a}) = n\}.$$

We will need the result of Sierpiński [1906] (for a statement in English, see [Schinzel 1972, (1)])

$$\sum_{n=1}^{x} r(n,2) = \pi x + O(x^{1/3}). \tag{4}$$

The error term $O(x^{1/3})$ has been improved by Huxley [2003] to $O(x^{131/416+\epsilon})$, but the former is sufficient for our purposes. We shall also use

$$\sum_{n=1}^{x} \frac{r(n,2)}{n} = \pi(S + \log x) + O(x^{-1/2}) \tag{5}$$

[Sierpiński 1907], where S denotes Sierpiński's constant $S \approx 2.58/\pi$. This also has the alternate expressions

$$S = \frac{1}{\pi} \lim_{z\to\infty}\left(4\zeta(z)\beta(z) - \frac{\pi}{z-1}\right) = \gamma + \frac{\beta'(1)}{\beta(1)}$$

[Finch 2003, p. 123], where $\beta(z)$ is the Dirichlet beta function and γ is the Euler–Mascheroni constant.

With these functions at hand, we may now proceed to calculate the desired probability. To do so, we will need two preliminary results. The first is the total number of pairs of ideals generated by Gaussian integers with norm at most x. The second result is the number of those pairs which have greatest common divisor $\mathfrak{K}$. The expressions for each of these are derived in the following two lemmas.

Lemma 5. *The total number of pairs of nonzero ideals* $\mathfrak{n}$ *and* $\mathfrak{m}$ *in* $\mathbb{Z}[i]$ *with norm* x *or less is*

$$\frac{\pi^2 x^2}{16} + O(x^{4/3}).$$

Proof. Let $\mathfrak{n}$ and $\mathfrak{m}$ be nonzero ideals. Then

$$\#\{\mathfrak{n}, \mathfrak{m} \subset \mathbb{Z}[i]^2 : N(\mathfrak{n}), N(\mathfrak{m}) \le x\} = \sum_{\substack{\mathfrak{n}\subset\mathbb{Z}[i] \\ N(\mathfrak{n})\le x}} \sum_{\substack{\mathfrak{m}\subset\mathbb{Z}[i] \\ N(\mathfrak{m})\le x}} 1,$$

and we may rewrite this as

$$\frac{1}{16} \sum_{N(\mathfrak{n})=1}^{\lfloor x\rfloor} r(N(\mathfrak{n}),2) \sum_{N(\mathfrak{m})=1}^{\lfloor x\rfloor} r(N(\mathfrak{m}),2),$$

which by (4) equals

$$\tfrac{1}{16}(\pi x + O(\lfloor x\rfloor^{1/3}))^2.$$

Further, since $O(\lfloor x\rfloor) = O(x)$, we may expand $(\pi x + O(x^{1/3}))^2$ and obtain $\pi^2x^2 + 2\pi x\, O(x^{1/3}) + O(x^{2/3})$, which reduces to $\pi^2 x^2 + O(x^{4/3})$. Thus, the total number of $\mathfrak{n}$ and $\mathfrak{m}$ with norm at most x is

$$\frac{\pi^2 x^2}{16} + O(x^{4/3}). \qquad \square$$

Lemma 6. *The total number of pairs of nonzero ideals* $\mathfrak{n}$ *and* $\mathfrak{m}$ *in* $\mathbb{Z}[i]$ *with norm* x *or less having greatest common divisor* $\mathfrak{K}$ *is*

$$\frac{\pi^2x^2}{16\zeta_{\mathbb{Q}(i)}(2)k^2}+O\left(\frac{x^{4/3}}{k^{4/3}}\right),$$

where $k=N(\mathfrak{K})$.

Proof. Let $\mathfrak{n}$ and $\mathfrak{m}$ be nonzero ideals. The number of pairs of $\mathfrak{n}$ and $\mathfrak{m}$ with norm x or less which are relatively prime is

$$\begin{aligned}\#\big\{\mathfrak{n},\mathfrak{m}\subset\mathbb{Z}[i]^2 : N(\mathfrak{n}),N(\mathfrak{m})\le x \text{ and } (\mathfrak{n},\mathfrak{m})=(1)\big\} &= \sum_{\substack{\mathfrak{n}\subset\mathbb{Z}[i]\\ N(\mathfrak{n})\le x}}\ \sum_{\substack{\mathfrak{m}\subset\mathbb{Z}[i]\\ N(\mathfrak{m})\le x\\ (\mathfrak{n},\mathfrak{m})=(1)}} 1\\ &= \sum_{\substack{\mathfrak{n}\subset\mathbb{Z}[i]\\ N(\mathfrak{n})\le x}}\ \sum_{\substack{\mathfrak{m}\subset\mathbb{Z}[i]\\ N(\mathfrak{m})\le x}}\ \sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ \mathfrak{d}|(\mathfrak{n},\mathfrak{m})}} \mu(\mathfrak{d}),\end{aligned}$$

where in the last line we used identity (3). Reindexing with $\mathfrak{n}=\mathfrak{d}\mathfrak{n}'$ and $\mathfrak{m}=\mathfrak{d}\mathfrak{m}'$, where the norms of $\mathfrak{n}'$ and $\mathfrak{m}'$ range from 1 to $x/N(\mathfrak{d})$, we may rewrite this as

$$\begin{aligned}&\sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})\le x}} \mu(\mathfrak{d}) \sum_{\substack{\mathfrak{n}'\subset\mathbb{Z}[i]\\ N(\mathfrak{n}')\le x/N(\mathfrak{d})}}\ \sum_{\substack{\mathfrak{m}'\subset\mathbb{Z}[i]\\ N(\mathfrak{m}')\le x/N(\mathfrak{d})}} 1\\ &\qquad=\frac{1}{16}\sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})\le x}} \mu(\mathfrak{d}) \sum_{N\mathfrak{n}'=1}^{\lfloor x/N\mathfrak{d}\rfloor} r(N(\mathfrak{n}'),2) \sum_{N\mathfrak{m}'=1}^{\lfloor x/N\mathfrak{d}\rfloor} r(N(\mathfrak{m}'),2).\end{aligned}$$

As in Lemma 5, this reduces to

$$\frac{1}{16}\sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})\le x}} \mu(\mathfrak{d})\left(\frac{\pi^2x^2}{N(\mathfrak{d})^2}+O\left(\frac{x}{N(\mathfrak{d})}\right)^{4/3}\right).$$

We then distribute the summation to obtain

$$\frac{\pi^2x^2}{16}\sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})\le x}}\frac{\mu(\mathfrak{d})}{N(\mathfrak{d})^2}+O\left(\sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})\le x}}\left(\frac{x}{N(\mathfrak{d})}\right)^{4/3}\right). \tag{6}$$

To evaluate the main term, we call on the generating function $\sum_{\mathfrak{n}\subset\mathbb{Z}[i]}\mu(\mathfrak{n})/N(\mathfrak{n})^s = 1/\zeta_{\mathbb{Q}(i)}(s)$ for $\mathrm{Re}(s)>1$ to see that

$$\sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})\le x}}\frac{\mu(\mathfrak{d})}{N(\mathfrak{d})^2}=\frac{1}{\zeta_{\mathbb{Q}(i)}(2)}-\sum_{n=x+1}^{\infty}\ \sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})=n}}\frac{\mu(\mathfrak{d})}{N(\mathfrak{d})^2},$$

which implies

$$\left|\frac{1}{\zeta_{\mathbb{Q}(i)}(2)} - \sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})\leq x}} \frac{\mu(\mathfrak{d})}{N(\mathfrak{d})^2}\right| \leq \sum_{n=x+1}^{\infty} \frac{1}{n^2} \sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})=n}} 1 = \frac{1}{4}\sum_{n=x+1}^{\infty} \frac{r(n,2)}{n^2}.$$

Now we note that $r(n,2) \leq 4\sigma_0(n) = o(n^{\epsilon})$ for all $\epsilon > 0$, where σ_0 represents the number of divisors of n. Thus

$$\frac{1}{4}\sum_{n=x+1}^{\infty}\frac{r(n,2)}{n^2} \leq \sum_{n=x+1}^{\infty}\frac{o(n^{\epsilon})}{n^2} = o(x^{\epsilon-1}),$$

and so

$$\left|\frac{1}{\zeta_{\mathbb{Q}(i)}(2)} - \sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})\leq x}} \frac{\mu(\mathfrak{d})}{N(\mathfrak{d})^2}\right| \leq o(x^{\epsilon-1}) \quad \text{or} \quad \sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})\leq x}} \frac{\mu(\mathfrak{d})}{N(\mathfrak{d})^2} = \frac{1}{\zeta_{\mathbb{Q}(i)}(2)} + o(x^{\epsilon-1}).$$

For the error term of (6), we have

$$\sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})\leq x}} \left(\frac{1}{N(\mathfrak{d})}\right)^{4/3} = \sum_{n=1}^{x}\frac{1}{n^{4/3}} \sum_{\substack{\mathfrak{d}\subset\mathbb{Z}[i]\\ N(\mathfrak{d})=n}} 1 = \frac{1}{4}\sum_{n=1}^{x}\frac{r(n,2)}{n^{4/3}}$$

and again use the bound $r(n,2) \leq o(n^{\epsilon})$ to see that

$$\frac{1}{4}\sum_{n=1}^{x}\frac{r(n,2)}{n^{4/3}} \leq \sum_{n=1}^{x} o(n^{\epsilon-4/3}),$$

which equals $o(x^{\epsilon-1/3}) + o(1)$. From this it is clear that

$$O\left(x^{4/3}\sum_{n=1}^{x}\frac{r(n,2)}{n^{4/3}}\right) = O(o(x^{4/3})) = O(x^{4/3}).$$

Thus (6) becomes

$$\frac{\pi^2 x^2}{16\zeta_{\mathbb{Q}(i)}(2)} + o(x^{\epsilon-1}) + O(x^{4/3}),$$

which allows us to conclude

$$\#\big\{\mathfrak{n},\mathfrak{m}\subset\mathbb{Z}[i]^2 : N(\mathfrak{n}), N(\mathfrak{m}) \leq x \text{ and } (\mathfrak{n},\mathfrak{m}) = (1)\big\} = \frac{\pi^2 x^2}{16\zeta_{\mathbb{Q}(i)}(2)} + O(x^{4/3}).$$

Having counted the number of relatively prime $\mathfrak{n}$ and $\mathfrak{m}$ within a given norm, we can now reindex to obtain the number of them which have $(\mathfrak{n},\mathfrak{m}) = \mathfrak{K}$. Letting $\mathfrak{n} = \mathfrak{n}'\mathfrak{K}$ and $\mathfrak{m} = \mathfrak{m}'\mathfrak{K}$, we see that $\mathfrak{n}'$ and $\mathfrak{m}'$ are relatively prime if and only if $\mathfrak{n}$ and $\mathfrak{m}$ have $\mathfrak{K}$ as their greatest common divisor. Hence, the number of relatively prime pairs $\mathfrak{n}'$ and $\mathfrak{m}'$ with norm y or less must be equivalent to the number of pairs $\mathfrak{n}$ and $\mathfrak{m}$

with norm yk or less (where $k = N(\mathfrak{K})$) having greatest common divisor $\mathfrak{K}$. Thus,

$$\begin{aligned}\#\{\mathfrak{n}, \mathfrak{m} \subset \mathbb{Z}[i]^2 : N(\mathfrak{n}), N(\mathfrak{m}) \le x \text{ and } (\mathfrak{n}, \mathfrak{m}) = \mathfrak{K}\}\\ &= \#\{\mathfrak{n}', \mathfrak{m}' \subset \mathbb{Z}[i]^2 : N(\mathfrak{n}'), N(\mathfrak{m}') \le x/k \text{ and } (\mathfrak{n}', \mathfrak{m}') = (1)\}\\ &= \frac{\pi^2 x^2}{16\zeta_{\mathbb{Q}(i)}(2)k^2} + O\Big(\frac{x^{4/3}}{k^{4/3}}\Big), \qquad \square\end{aligned}$$

Lastly, the probability that $\mathfrak{n}$ and $\mathfrak{m}$, having norm at most x, will have greatest common divisor $\mathfrak{K}$ is defined to be the number of pairs of ideals of norm x or less which have greatest common divisor $\mathfrak{K}$ divided by the total number of pairs of ideals of norm x or less. Thus, by Lemmas 5 and 6,

$$\begin{aligned}P_x\{\mathfrak{n}, \mathfrak{m} \subset \mathbb{Z}[i]^2 : N(\mathfrak{n}), N(\mathfrak{m}) \le x \text{ and } (\mathfrak{n}, \mathfrak{m}) = \mathfrak{K}\}\\ &= \Big(\frac{\pi^2 x^2}{16} + O(x^{4/3})\Big)^{-1}\Big(\frac{\pi^2 x^2}{16\zeta_{\mathbb{Q}(i)}(2)k^2} + O\Big(\frac{x^{4/3}}{k^{4/3}}\Big)\Big). \qquad (7)\end{aligned}$$

We can rewrite $(\pi^2 x^2/16 + O(x^{4/3}))^{-1}$ as $16\pi^{-2}x^{-2}(1 + O(x^{-2/3}))^{-1}$, which is equal to $16\pi^{-2}x^{-2}(1 + O(x^{-2/3}))$ since $(1 + f(x))^{-1} = 1 + O(f(x))$ for $f(x)$ tending towards 0 as x approaches infinity.

Line (7) then becomes

$$\begin{aligned}\pi^{-2}x^{-2}(1 + O(x^{-2/3}))\Big(\frac{\pi^2 x^2}{\zeta_{\mathbb{Q}(i)}(2)k^2} + O\Big(\frac{x^{4/3}}{k^{4/3}}\Big)\Big)\\ &= (1 + O(x^{-2/3}))\Big(\frac{1}{\zeta_{\mathbb{Q}(i)}(2)k^2} + O\Big(\frac{1}{x^{2/3}k^{4/3}}\Big)\Big)\\ &= \frac{1}{\zeta_{\mathbb{Q}(i)}(2)k^2} + O\Big(\frac{1}{x^{2/3}k^{4/3}}\Big) + O\Big(\frac{1}{x^{2/3}k^2}\Big) + O\Big(\frac{1}{x^{4/3}k^{4/3}}\Big),\end{aligned}$$

or finally

$$\frac{1}{\zeta_{\mathbb{Q}(i)}(2)k^2} + O\Big(\frac{1}{x^{2/3}k^{4/3}}\Big),$$

completing the proof of Theorem 1. The following corollary is a direct consequence of Theorem 1 for the special case when $\mathfrak{K} = (1)$.

Corollary 7. *The probability that a pair of Gaussian integers with norm x or less are relatively prime is*

$$\frac{1}{\zeta_{\mathbb{Q}(i)}(2)} + O\Big(\frac{1}{x^{2/3}}\Big).$$

In effect, Corollary 7 tells us that for x large, the probability that two Gaussian integers are relatively prime is asymptotic to $(\zeta_{\mathbb{Q}(i)}(2))^{-1}$ as x tends towards infinity. This is in agreement with the work of Collins and Johnson who state the probability as $(\zeta_{\mathbb{Q}(i)}(2))^{-1} = (\zeta(2)L(2, \chi))^{-1} \approx 0.6637$, where $L(2, \chi)$ is a Dirichlet L-series and χ the primitive Dirichlet character modulo 4.

3. Expected value

Having derived the probability distribution function found in Theorem 1, we are ready to find an expression for the expected value of our random variable, $N(\mathfrak{n}, \mathfrak{m}) = k$, where the norm of $\mathfrak{n}$ and $\mathfrak{m}$ ranges from 1 to x. To do this, we must express our probability in terms of k as well. The modification is simple, however. Since the number of ideals with norm k is equivalent to $r(k, 2)/4$, the probability that the greatest common divisor of $\mathfrak{n}$ and $\mathfrak{m}$ has norm k must be

$$P_x\{N(\mathfrak{n}, \mathfrak{m}) = k\} = \frac{r(k, 2)}{4\zeta_{\mathbb{Q}(i)}(2)k^2} + O\left(\frac{r(k, 2)}{x^{2/3}k^{4/3}}\right).$$

Then, by definition of expected value

$$\begin{aligned} E_x\{N(\mathfrak{n}, \mathfrak{m})\} &= \sum_{k=1}^{x} k P_x\{N(\mathfrak{n}, \mathfrak{m}) = k\} \\ &= \sum_{k=1}^{x} k\left(\frac{r(k, 2)}{4\zeta_{\mathbb{Q}(i)}(2)k^2} + O\left(\frac{r(k, 2)}{x^{2/3}k^{4/3}}\right)\right) \\ &= \frac{1}{4\zeta_{\mathbb{Q}(i)}(2)} \sum_{k=1}^{x} \frac{r(k, 2)}{k} + O\left(\frac{1}{x^{2/3}} \sum_{k=1}^{x} \frac{r(k, 2)}{k^{1/3}}\right). \end{aligned} \tag{8}$$

Using Stieltjes integration by parts to evaluate the error term, we obtain

$$\begin{aligned} \sum_{k=1}^{x} \frac{r(k, 2)}{k^{1/3}} &= x^{-1/3} \sum_{k=1}^{x} r(k, 2) - 4 - \int_1^x (\pi k + O(k^{1/3}))\left(-\tfrac{1}{3}k^{-4/3}\right) dk \\ &= \frac{3\pi}{2} x^{2/3} + O(\log x), \end{aligned}$$

which implies

$$O\left(x^{-2/3} \sum_{k=1}^{x} \frac{r(k, 2)}{k^{1/3}}\right) = O(1 + x^{-2/3}\log x) = O(1).$$

The main term of (8) can be rewritten using Sierpiński's identity from (5). Thus the expected value is equal to

$$\frac{1}{4\zeta_{\mathbb{Q}(i)}(2)}\left(\pi(S + \log x) + O\left(\frac{1}{x^{1/2}}\right)\right) + O(1)$$

or

$$\frac{\pi}{4\zeta_{\mathbb{Q}(i)}(2)} \log x + O(1).$$

This completes the proof of Theorem 2.

4. Higher moments

Finally, we show that there exists some constant $c_n \in \mathbb{R}$ such that the main term of the n-th moment of $N(\mathfrak{n}, \mathfrak{m})$ must be of the form $c_n x^{n-1}$ for $n > 2$. Let $N(\mathfrak{n}), N(\mathfrak{m}) \le x$ with $(\mathfrak{n}, \mathfrak{m}) = \mathfrak{K}$ and restrict $N(\mathfrak{K})$ to the interval $\bigl(x/(j+1), x/j\bigr]$. We may then write $\mathfrak{n} = \mathfrak{n}'\mathfrak{K}$ and $\mathfrak{m} = \mathfrak{m}'\mathfrak{K}$, where $(\mathfrak{n}', \mathfrak{m}') = (1)$. The restriction on the norm of $\mathfrak{K}$ allows us to see that $N(\mathfrak{n}'), N(\mathfrak{m}') < x(j+1)/x$, which implies $N(\mathfrak{n}'), N(\mathfrak{m}') \le j$. Now define

$$f(j) = \#\bigl\{(\mathfrak{n}', \mathfrak{m}') \subset \mathbb{Z}[i]^2 : N(\mathfrak{n}'), N(\mathfrak{m}') \le j \text{ and } (\mathfrak{n}', \mathfrak{m}') = (1)\bigr\}$$

for $j \in \mathbb{N}$. By Lemma 6,

$$\begin{aligned} f(j) &= \frac{\pi^2 j^2}{16\zeta_{\mathbb{Q}(i)}(2)} + O(j^{4/3}) \\ &= O(j^2). \end{aligned}$$

Our reindexing above shows that this expression for $f(j)$ also gives us the number of pairs of ideals with norm x or less having greatest common divisor $\mathfrak{K}$, where $x/(j+1) < N(\mathfrak{K}) \le x/j$. Thus the n-th moment of $N(\mathfrak{n}, \mathfrak{m})$ is given by

$$E_x\{N(\mathfrak{n}, \mathfrak{m})^n\} = \frac{1}{\pi^2 x^2/16 + O(x^{4/3})} \Biggl(\sum_{j=1}^{x} f(j) \sum_{\substack{\mathfrak{K} \subset \mathbb{Z}[i] \\ x/(j+1) < N(\mathfrak{K}) \le x/j}} N(\mathfrak{K})^n \Biggr). \tag{9}$$

We next turn our attention to the inner sum of (9). First note that

$$\sum_{\substack{\mathfrak{K} \subset \mathbb{Z}[i] \\ x/(j+1) < N(\mathfrak{K}) \le x/j}} N(\mathfrak{K})^n = \frac{1}{4} \sum_{k=\lceil x/(j+1) \rceil}^{\lfloor x/j \rfloor} k^n r(k, 2),$$

where $k = N(\mathfrak{K})$. Then Stieltjes integration by parts yields

$$\begin{aligned} \frac{1}{4} \sum_{k=\lceil x/(j+1) \rceil}^{\lfloor x/j \rfloor} k^n r(k,2) &= \left\lfloor \frac{x}{j} \right\rfloor^n \sum_{k=1}^{\lfloor x/j \rfloor} r(k,2) - \left\lceil \frac{x}{j+1} \right\rceil^n \sum_{k=1}^{\lceil x/(j+1) \rceil} r(k,2) \\ &\qquad - \int_{\lceil x/(j+1) \rceil}^{\lfloor x/j \rfloor} n t^{n-1} (\pi t + O(t^{1/3}))\, dt \\ &= \frac{\pi}{4(n+1)} x^{n+1} \left(\frac{1}{j^{n+1}} - \frac{1}{(j+1)^{n+1}} \right) + O\left(\frac{x}{j} \right)^{n+1/3}. \end{aligned}$$

The numerator of $E_x\{N(\mathfrak{n}, \mathfrak{m})^n\}$ is now equal to

$$\frac{\pi}{4(n+1)} x^{n+1} \sum_{j=1}^{x} O(j^2) \left(\frac{(j+1)^{n+1} - j^{n+1}}{j^{n+1}(j+1)^{n+1}} \right) + x^{n+1/3} \sum_{j=1}^{x} O(j^2) O\left(\frac{1}{j^{n+1/3}} \right). \tag{10}$$

The sum on the left is

$$\sum_{j=1}^{x} O(j^2) O\left(\frac{1}{j(j+1)^{n+1}}\right) = \sum_{j=1}^{x} O\left(\frac{1}{j^{-1}(j+1)^{n+1}}\right),$$

which is bounded above by $\sum_{j=1}^{x} O(1/j^n)$. For x tending toward infinity and $n \geq 2$, this converges to some constant $c_n' \in \mathbb{R}$. A similar argument shows that the second sum of (10) is likewise convergent for $n > 2$. We thus conclude that the main term of $E_x\{N(\mathfrak{n}, \mathfrak{m})^n\}$ is of the form $c_n' x^{n+1}$.

Finally, we divide this by the total number of pairs of ideals $\mathfrak{n}$, $\mathfrak{m}$ with norm at most x to obtain the main term of the n-th moment of $N(\mathfrak{n}, \mathfrak{m})$ for $n > 2$

$$\frac{c_n' x^{n+1}}{\pi^2 x^2/16 + O(x^{4/3})} = c_n x^{n-1} \frac{1}{1 + O(x^{-2/3})},$$

where $c_n = 16 c_n'/\pi^2$. Since

$$(1 + O(x^{-2/3}))^{-1} = 1 + O(x^{-2/3}),$$

it follows that for x tending to infinity

$$E_x\{N(\mathfrak{n}, \mathfrak{m})^n\} \sim c_n x^{n-1}.$$

With this, we bring the proof of Theorem 3 to an end and close by restating our conjecture regarding the constant of $E_x\{N(\mathfrak{n}, \mathfrak{m})^n\}$ for all $n \geq 2$. We also include numerical evidence below which provides support for the conjecture in the cases when $n = 2, 3, 4$ and 5.

Conjecture 4. *For $n \geq 2$,*

$$E_x\{N(\mathfrak{n}, \mathfrak{m})^n\} \sim \frac{4}{\pi(n+1)} \left(\frac{2\zeta_{\mathbb{Q}(i)}(n)}{\zeta_{\mathbb{Q}(i)}(n+1)} - 1 \right) x^{n-1}.$$

Using Matlab, we first compiled a list of all pairs of Gaussian integers in the first quadrant with norm x or less and used the Euclidean algorithm to find all possible greatest common divisors. We determined the n-th moment by raising the norm of each greatest common divisor to the n-th power, summed the terms together, and then divided the result by the total number of pairs of Gaussian integers in the first quadrant with norm x or less. We have graphed the results in Figures 1–4 below for the cases when $n = 2, 3, 4$ and 5 with $x = 50,000$. In Table 1, we have listed the main term of the best fit curve corresponding to each graph as compared against the conjectured main term for each value of n.

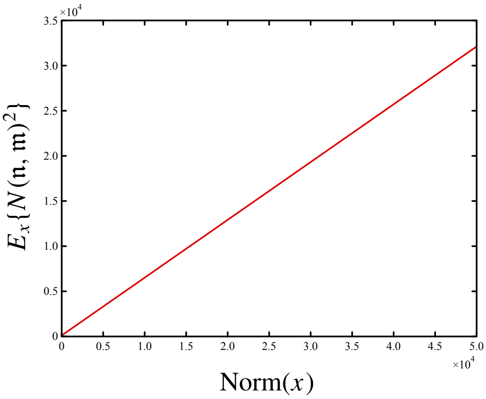


Figure 1. The graph of $E_x\{N(\mathfrak{n}, \mathfrak{m})^2\}$ for $1 \leq x \leq 50,000$. The best fit curve is $0.63952x + 0.5753$.

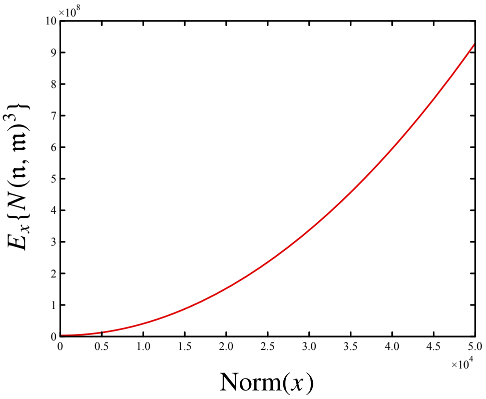


Figure 2. The graph of $E_x\{N(\mathfrak{n}, \mathfrak{m})^3\}$ for $1 \leq x \leq 50,000$. The best fit curve is $0.37018x^2 + 0.69337x - 584.8498$.

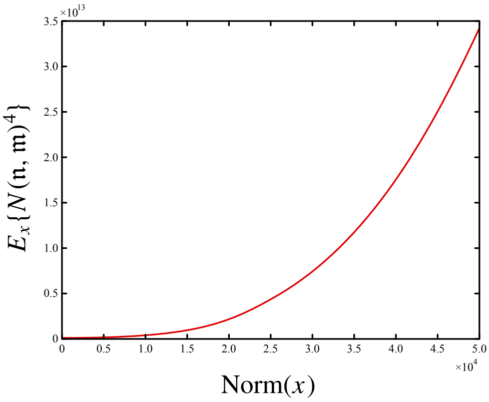


Figure 3. The graph of $E_x\{N(\mathfrak{n}, \mathfrak{m})^4\}$ for $1 \leq x \leq 50,000$. The best fit curve is $0.27238x^3 + 0.80149x^2 - 3723.1433x + 12324561.4508$.

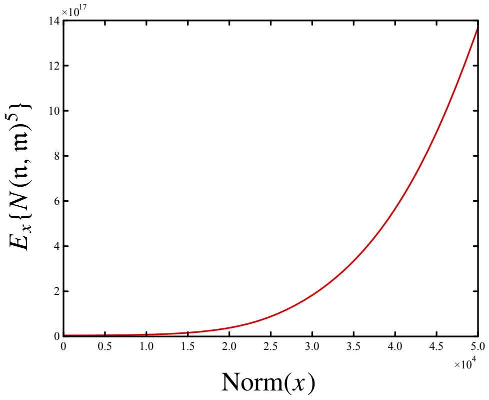


Figure 4. The graph of $E_x\{N(\mathfrak{n}, \mathfrak{m})^5\}$ for $1 \leq x \leq 50,000$. The best fit curve is $0.21914x^4 + 0.92436x^3 - 9773.8223x^2 + 92150266.2382x - 190551355734.3794$.

moment (n)	numerically derived term	conjectured term
2	$0.63952x$	$0.67364x$
3	$0.37018x^2$	$0.37444x^2$
4	$0.27238x^3$	$0.27309x^3$
5	$0.21914x^4$	$0.21928x^4$

Table 1. The main term of the n-th moment of the norm of the greatest common divisor of pairs of Gaussian integers with norm at most x.

Acknowledgements

The authors were supported by the Rich Summer Internship grant of the Dr. Barnett and Jean Hollander Rich Scholarship Fund and would like to thank both the selection committee and donors. For additional funding, the first author was supported by the NIH Maximizing Access to Research Careers (MARC) U-STAR grant [5T34 GM007639], the second author by the National Science Foundation [DMS-1201446] and the third author by the City College Fellowships Program. Gratitude also goes to Joseph Dacanay whose help with Matlab laid the groundwork for our experimental data. We thank Giacomo Micheli for making us aware of his work and for his comments on our paper. The authors are also grateful to the referee for helpful comments regarding the statement and proof of Theorem 3. We especially thank our advisor, Dr. Brooke Feigon, without whose patient instruction and insightful comments this work would not have been possible.

References

[Cesàro 1885] E. Cesàro, "Étude moyenne du plus grand commun diviseur de deux nombres", *Ann. Mat. Pura Appl.* **13**:2 (1885), 233–268. Zbl 17.0144.05

[Christopher 1956] J. Christopher, "The asymptotic density of some k-dimensional sets", *Amer. Math. Monthly* **63** (1956), 399–401. MR 20 #3832 Zbl 0070.04101

[Collins and Johnson 1989] G. E. Collins and J. R. Johnson, "The probability of relative primality of Gaussian integers", pp. 252–258 in *Symbolic and algebraic computation* (Rome, 1988), edited by P. Gianni, Lecture Notes in Comput. Sci. **358**, Springer, Berlin, 1989. MR 90m:11165

[Diaconis and Erdős 2004] P. Diaconis and P. Erdős, "On the distribution of the greatest common divisor", pp. 56–61 in *A festschrift for Herman Rubin*, edited by A. DasGupta, IMS Lecture Notes Monogr. Ser. **45**, Inst. Math. Statist., Beachwood, OH, 2004. MR 2005m:60011 Zbl 1268.11139

[Finch 2003] S. R. Finch, *Mathematical constants*, Encyclopedia of Mathematics and its Applications **94**, Cambridge University Press, 2003. MR 2004i:00001 Zbl 1054.00001

[Huxley 2003] M. N. Huxley, "Exponential sums and lattice points, III", *Proc. London Math. Soc.* (3) **87**:3 (2003), 591–609. MR 2004m:11127 Zbl 1065.11079

[Mertens 1874] F. Mertens, "Ueber einige asymptotische Gesetze der Zahlentheorie", *J. Reine Angew. Math.* **77** (1874), 289–338. MR 1579608

[Micheli and Ferraguti 2015] G. Micheli and A. Ferraguti, "On Mertens–Cesáro theorem for number fields", preprint, 2015. To appear in *Bull. Aust. Math. Soc.* arXiv 1409.6527

[Micheli and Schnyder 2015] G. Micheli and R. Schnyder, "On the density of coprime m-tuples over holomorphy rings", *Int. J. Number Theory* (online publication September 2015).

[Schinzel 1972] A. Schinzel, "Wacław Sierpiński's papers on the theory of numbers", *Acta Arith.* **21** (1972), 7–13. (errata insert). MR 46 #9b Zbl 0243.01028

[Sierpiński 1906] W. Sierpiński, "O pewnem zagadnieniu z rachunku funkcyj asymptotycznych", *Prace Matematyczno-Fizyczne* **17** (1906), 77–118.

[Sierpiński 1907] W. Sierpiński, "O sumowaniu szeregu $\sum_{n>a}^{n\le b} \tau(n) f(n)$, gdzie $\tau(n)$ oznacza liczbę rozkładów liczby n na sumę kwadratów dwóch liczb całkowitych", *Prace Matematyczno-Fizyczne* **18** (1907), 1–59. JFM 38.0319.02

Received: 2013-03-27 Revised: 2015-01-09 Accepted: 2015-01-28

tai.danae@gmail.com *Department of Mathematics, The Graduate Center, CUNY, 365 5th Avenue, New York, NY 10016, United States*

cycsano@hotmail.com *Department of Mathematics, The Graduate Center, CUNY, 365 5th Avenue, New York, NY 10016, United States*

fay.or.flymorning@gmail.com *GACE Consulting Engineers PC, 105 Madison Avenue, 6th Floor, New York, NY 10016, United States*

INVOLVE 9:1 (2016)
dx.doi.org/10.2140/involve.2016.9.41

Proving the pressing game conjecture on linear graphs

Eliot Bixby, Toby Flint and István Miklós

(Communicated by Joshua Cooper)

The pressing game on black-and-white graphs is the following: given a graph $G(V, E)$ with its vertices colored with black and white, any black vertex v can be pressed, which has the following effect: (1) all neighbors of v change color; i.e., white neighbors become black and vice versa; (2) all pairs of neighbors of v change adjacency; i.e., adjacent pairs become nonadjacent and nonadjacent ones become adjacent; and (3) v becomes a separated white vertex. The aim of the game is to transform G into an all-white, empty graph. It is a known result that the all-white empty graph is reachable in the pressing game if each component of G contains at least one black vertex, and for a fixed graph, any successful transformation has the same number of pressed vertices.

The pressing game conjecture states that any successful pressing sequence can be transformed into any other successful pressing sequence with small alterations. Here we prove the conjecture for linear graphs, also known as paths. The connection to genome rearrangement and sorting signed permutations with reversals is also discussed.

1. Introduction

Sorting signed permutations by reversals (or inversions as biologists call it) is the first genome rearrangement model introduced in the scientific literature. The hypothesis that reversals change the order and orientation of genes — called genetic factors at the time — arose in [Sturtevant 1921] and was implicitly verified upon the discovery of chromosomes [Sturtevant and Novitski 1941]. At the same time, geneticists realized that “the mathematical properties of series of letters subjected to the operation of successive inversions do not appear to be worked out” [Sturtevant and Tan 1937]. In constructing phylogenies, maximum parsimony — supposing the

MSC2010: primary 05A05; secondary 05CXX.

Keywords: bioinformatics, sorting by reversals, pressing game, irreducible Markov chain.

This paper presents the results of the undergraduate research of E. Bixby and T. Flint in the 2012 Fall semester at the Budapest Semesters in Mathematics.

least evolutionary change as the most likely explanation — is a desirable characteristic. As such, the construction of minimum length sorting by reversals is both a biologically and mathematically interesting problem. This computational problem was rediscovered at the end of the 20th century, and its solution is known as the Hannenhalli–Pevzner theorem [1995; 1999].

The Hannenhalli–Pevzner theorem gives a polynomial running time algorithm that finds one such minimum length sorting sequence, that is, a series of reversals that transforms one signed permutation into another. However, there might be multiple solutions, and the number of solutions typically grows exponentially with the length of the permutation. Therefore, a(n almost) uniform sampler is required which gives a set of solutions from which statistical properties of the solutions can be calculated. The Markov chain Monte Carlo method (MCMC) is a typical approach to such sampling. MCMC starts with an arbitrary solution, and applies random perturbations on it, thus exploring the solution space. In the case of most parsimonious reversal sorting sequences, two distinct methods of perturbation have been considered:

(1) The first approach encodes the most parsimonious reversal sorting sequences with the intermediate permutations which appear as the result of the perturbations: $\pi_{\text{start}} = \pi_1$ is transformed into π_2, which is transformed into $\pi_3, \ldots,$ which is transformed into $\pi_n = \pi_{\text{end}}$. Then it cuts out a random interval from this sequence, $\pi_i, \pi_{i+i}, \ldots, \pi_j$ and gives a new, random sorting sequence between the permutations at the beginning and end of the window, namely, between π_i and π_j.

(2) The second approach encodes the scenarios with the series of mutations applied, and perturbs them in a sophisticated way, described in detail later in this paper.

As random perturbations are applied, the Markov chain randomly explores the solution space and will be at a random state after some number of steps. This random state is described by its distribution over the state space. A Markov chain is said to *converge* to a distribution ϕ if the distribution of its random state after some number of steps converges to ϕ as the number of steps tends to infinity.

A Markov chain for sampling purposes should fulfill two conditions: (a) it must converge to the uniform distribution, and as such must be irreducible, namely, from any solution the chain must be able to get to any another solution, and (b) the convergence must be fast.

Unfortunately, the first approach has been shown to be slowly mixing [Miklós et al. 2010]. This means that the necessary number of steps in the Markov chain to sufficiently approximate the uniform distribution grows exponentially with the length of the permutation. Therefore this approach is not applicable in practice.

Unfortunately, it is not known whether or not the second approach is irreducible, let alone whether or not it is rapidly mixing. In this paper, we take a step towards proving that this method is, in fact, irreducible.

This paper is organized in the following way. In Section 2, we define the problem of sorting by reversals, and the combinatorial tools necessary: the graph of desire and reality and the overlap graph. Then we introduce the pressing game on black-and-white graphs, and show that they correspond to the shortest reversal scenarios in a subset of permutations that typically appear in biology. We finish the section by stating the pressing game conjecture, a proof of which would imply the second method is irreducible. In Section 3, we prove the conjecture for linear graphs, also known as paths. The paper is finished with a discussion and conclusions.

2. Preliminaries

Definition. A *signed permutation* is a permutation of numbers from 1 to n, where each number has a $+$ or $-$ sign.

While the number of length n permutations is $n!$, the number of length n signed permutations is $2^n \times n!$.

Definition. A *reversal* takes any contiguous piece of a signed permutation and reverses both the order of the numbers and the sign of each number. It is also allowed that a reversal takes only a single number from the signed permutations; in that case, it changes the sign of this number.

For example, the following reversal flips the $-3\ {+6}\ {-5}\ {+4}\ {+7}$ segment:

$$+8\ {-1}\ {-3}\ {+6}\ {-5}\ {+4}\ {+7}\ {-9}\ {+2} \quad \Rightarrow \quad +8\ {-1}\ {-7}\ {-4}\ {+5}\ {-6}\ {+3}\ {-9}\ {+2}.$$

The sorting by reversals problem asks for the minimum number of reversals necessary to transform a signed permutation into the identity permutation, i.e., the signed permutation $+1\ +2\ \cdots\ +n$. This number is called the *reversal distance*, and the reversal distance of a signed permutation π is denoted by $d_{\mathrm{REV}}(\pi)$. To solve this problem, we have to introduce two discrete mathematical objects, the graph of desire and reality and the overlap graph. The graph of desire and reality is a drawn graph, meaning both edges and vertex locations affect the properties of the graph. The overlap graph is a graph in terms of standard graph theory.

The *graph of desire and reality* for a signed permutation can be constructed in the following way. Each signed number is replaced with two unsigned numbers; $+i$ becomes $2i-1, 2i$, and similarly, $-i$ becomes $2i, 2i-1$. The resulting length $2n$ permutation is framed between 0 and $2n+1$. Each number including 0 and $2n+1$ will represent one vertex in the graph of desire and reality. They are drawn in the same order along a line as they appear in the permutation.

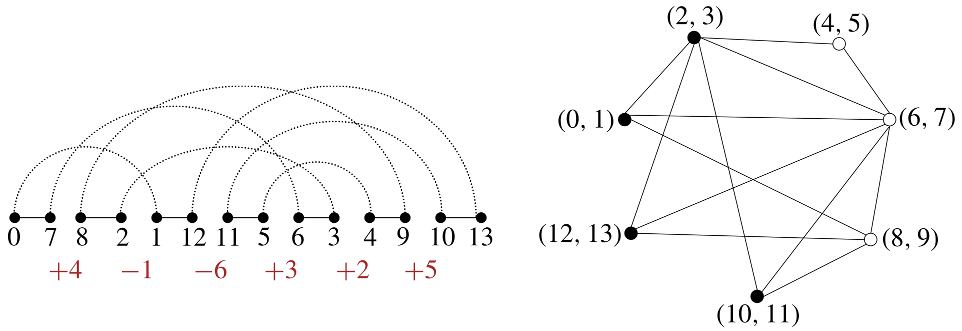


Figure 1. The graph of desire and reality and the overlap graph of the signed permutation $+4\ -1\ -6\ +3\ +2\ +5$.

We index the positions of the vertices starting with 1, and each pair of vertices in positions $2i-1$ and $2i$ are connected with an edge drawn as a straight line. We call these edges the *reality edges*. Each pair of vertices for numbers $2i$ and $2i+1$, $i = 0, 1, \ldots, n$ are connected with an edge drawn as an arc above the line of the vertices, and they are named the *desire edges*. The explanation for these names is that the reality edges describe what we see in the current permutation, and the desire edges describe the desired adjacencies in the final graph (the identity permutation): we would like 1 to be next to 0, 3 to be next to 2, etc.

Each desire edge is incident to two reality edges. We will call these edges the *legs* of the desire edge. A desire edge is called *oriented* if it spans an odd number of vertices. The rationale of this naming is that its legs point in the same direction; see, for example, the desire edge connecting 0 and 1 in Figure 1. A desire edge is called *unoriented* if it spans an even number of vertices and in this case, its legs indeed point in different directions; see, for example, the desire edge connecting 4 and 5 or the desire edge connecting 8 and 9 in Figure 1.

The *overlap graph* is constructed from the graph of desire and reality in the following way. The vertices of the overlap graph are the desire edges in the graph of desire and reality. The vertices are colored either black or white. A vertex in the overlap graph is black if it corresponds to an oriented desire edge. A vertex is white if it corresponds to an unoriented desire edge. Two vertices are adjacent if the intervals spanned by the corresponding desire edges overlap but neither contains the other. In Figure 1, we give an example for the graph of desire and reality and overlap graph.

The overlap graph might be disconnected. A component is called *oriented* if it contains at least one black vertex. If the component contains only white vertices, it is called *unoriented*. A component is *nontrivial* if it contains more than one vertex.

Any reversal changes the topology of the graph of desire and reality on two reality edges. Any desire edge is incident to two reality edges, and we say that the reversal *acts on* this desire edge if it changes the topology on the two incident reality edges.

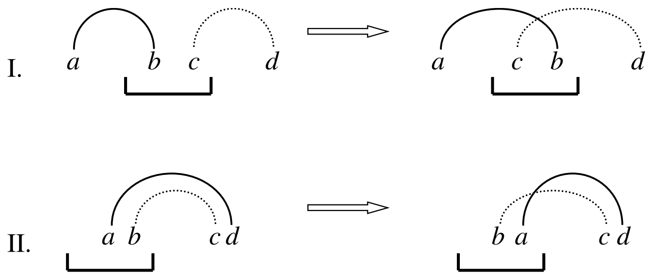


Figure 2. This picture shows how a reversal can change the overlap of two desire edges. The reversed fragment is indicated with a thick black line.

Any reversal in the underlying permutation also has the effect of reversing some segment of vertices in the graph of desire and reality. How do reversals acting on oriented desire edges change the graph of desire and reality and thus the overlap graph? We present a lemma below explaining this.

Lemma 1. *Fix a reversal, and let v be an oriented desire edge on which the reversal acts. Then the reversal*

(1) *changes whether any desire edge crossing v is oriented,*

(2) *changes whether any pair of desire edges crossing v overlaps, and*

(3) *causes the desire edge itself to become an unoriented edge without any overlapping edges (that is, neighbors in the overlap graph).*

Proof. (1) The reversal flips one of the legs of each overlapping desire edge. Therefore it changes the parity of the number of vertices below the desire edge and thus whether or not it is oriented.

(2) If two edges both overlap with v but not with each other because the intersection of their interval is empty, then the two edges must come from the two ends of v; see also Figure 2, case I. A reversal acting on v will change the order of one of the endpoints of their interval, so they will indeed overlap. If two edges overlap with v, but not with each other, since the interval of one of them contains the interval of the other, then they come from one end of v. It is easy to see that after the reversal they will overlap by definition; see Figure 2, case II. It is also easy to see that any overlapping pairs of edges which also overlap with each other are the two cases illustrated on the right-hand side of Figure 2, so after the reversal, they will not overlap.

(3) Finally, the oriented edge on which the reversal acts becomes an unoriented edge forming a small cycle with a reality edge, and thus it cannot overlap with any other desire edge. □

This lemma also shows the connection between sorting by reversals and the pressing game on black-and-white graphs: pressing a black vertex in an overlap graph is equivalent to reversing the corresponding desire edge. Below we define the pressing game on black-and-white graphs:

Definition. Given a graph $G(V, E)$ with its vertices colored with black and white, any black vertex v can be pressed, which has the following effect: (a) all neighbors of v change color, meaning that white neighbors become black and *vice versa*; (b) all pairs of neighbors of v change adjacency, meaning that adjacent pairs become nonadjacent and nonadjacent ones become adjacent; (c) finally, v becomes a separated white vertex. The aim of the game is to transform G into an all-white, empty graph.

If each component of G contains at least one black vertex, then the pressing game always has at least one solution, as it turns out, by the Hannenhalli–Pevzner theorem:

Theorem 2 [Hannenhalli and Pevzner 1999]. *Let π be a permutation whose overlap graph does not contain any nontrivial unoriented component. Then the reversal distance $d_{\text{REV}}(\pi)$, namely, the minimum number of reversals necessary to sort the permutation is*

$$d_{\text{REV}}(\pi) = n + 1 - c(\pi),$$

where n is the length of the permutation π and $c(\pi)$ is the number of cycles in the graph of desire and reality.

If the permutation π' contains a nontrivial unoriented component, then

$$d_{\text{REV}}(\pi') > n + 1 - c(\pi').$$

It is easy to see that any reversal can increase the number of cycles in the graph of desire and reality at most by 1, and the identity permutation contains $n+1$ cycles; hence the Hannenhalli–Pevzner theorem also says that if a permutation does not contain any nontrivial unoriented components, then any optimal reversal sorting sequence increases the number of cycles to $n+1$ without creating any nontrivial unoriented components. It is also true that these reversals can be chosen to act on oriented desire edges. Below we state this theorem.

Theorem 3. *Let π be a permutation which is not the identity permutation and whose overlap graph does not contain any nontrivial unoriented component. Then a reversal exists that acts on an oriented desire edge, increases $c(\pi)$ by* 1, *and does not create any nontrivial unoriented components.*

Furthermore, if G is an arbitrary black-and-white graph such that each component contains at least one black vertex, then at least one black vertex can be pressed without making a nontrivial unoriented component.

The proof can be found in [Bergeron 2001], and we skip it here. The proof considers only the overlap graph, and in fact, it indeed works for every black-and-white graph. A clear consequence is the following theorem.

Theorem 4. *Let G be a black-and-white graph such that each component contains at least one black vertex. Then G can be transformed into the all-white empty graph in the pressing game.*

Proof. It is sufficient to iteratively use Theorem 3. Indeed, according to Theorem 3, we can find a black vertex v such that pressing it does not create a nontrivial all-white component; on the other hand, v becomes a separated white vertex, and it will remain a separated white vertex afterward. Hence, the number of vertices in nontrivial components decreases at least by one, and in a finite number of steps, G is transformed into the all-white, empty graph. □

Consider the set of vertices as an alphabet; any sequence over this alphabet is called a *pressing sequence*. It is a *valid* pressing sequence when each vertex is black when it is pressed, and it is *successful* if it is valid and leads to the all-white, empty graph. The length of the pressing sequence is the number of vertices pressed in it. The following theorem is also true.

Theorem 5. *Let G be a black-and-white graph such that each component contains at least one black vertex. Then every successful pressing sequence of G has the same length.*

The proof can be found in [Hartman and Verbin 2006]. We are ready to state the pressing sequence conjecture.

Conjecture 6. *Let G be a black-and-white graph such that each component contains at least one black vertex. Construct a metagraph M whose vertices are the successful pressing sequences on G. Connect two vertices if the length of the longest common subsequence of the pressing sequences they represent is at least the common length of the pressing sequences minus* 4. *The conjecture is that M is connected.*

The conjecture means that with small alterations, we can transform any pressing sequence into any other pressing sequence, regardless of the underlying graph. By "small alteration" we mean that we remove at most four (not necessarily consecutive) vertices from a pressing sequence, and add at most four vertices, not necessarily to the same places where vertices were removed, and not necessarily to consecutive places.

It is important to note that there exist sorting sequences that are not pressing sequences. Specifically, these sequences contain two reversals which act on the same location in the permutation. These sequences also correspond to cycle-increasing reversals in the graph of desire and reality. However, the infinite site model [Ma et al. 2008] corresponds to permutations whose sorting sequences are exactly the pressing sequences, and restricting ourselves to this subset of permutations is a biologically reasonable assumption.

In this paper, we prove the pressing game conjecture for linear graphs. In addition, we can prove the metagraph will be already connected if we require that neighboring vertices have a longest common subsequence at least the common length of their pressing sequences minus 2.

3. Proof of the conjecture on linear graphs

The proof of our main theorem is recursive, and for this, we need the following notations. Let G be a black-and-white graph, and v a black vertex in it. Then Gv denotes the graph we get by pressing vertex v. Similarly, if P is a valid pressing sequence of G (namely, each vertex is black when we want to press it, but P does not necessarily yield the all-white, empty graph), then GP denotes the graph we get after pressing all vertices in P in the indicated order. Finally, let P^k denote the suffix of P starting in position $k+1$.

The convenience of linear graphs is their simple structure and furthermore, their self-reducibility:

Observation. Let G be a linear black-and-white graph and v a black vertex in it. Then Gv consists of a linear graph and the separated white vertex v.

Since any separated white vertex does not have to be pressed again, it is sufficient to consider $Gv \setminus \{v\}$, which is a linear graph. We are ready to state and prove our main theorem.

Theorem 7. *Let G be an arbitrary, finite, linear black-and-white graph, and let M be the following graph. The vertices of M are the successful pressing sequences on G, and two vertices are adjacent if the length of the longest common subsequence of the pressing sequences they represent is at least the common length of the pressing sequences minus* 2. *Then M is connected.*

Proof. It is sufficient to show that for any successful pressing sequences X and $Y = v_1 v_2 \cdots v_k$, there is a series $X_1, X_2, \dots, X_m$ such that for any $i = 1, 2, \dots, m-1$, the length of the longest common subsequence of X_i and X_{i+1} is at least the common length of the sequences minus 2, and X_m starts with v_1. Indeed, then both X_m and Y start with v_1, and both X_m^1 and Y^1 are successful pressing sequences on $Gv_1 \setminus \{v_1\}$. We can use induction to transform X_m into a pressing sequence which starts with v_2; then we consider its suffix, which is a successful pressing sequence on $Gv_1 v_2 \setminus \{v_1, v_2\}$, etc.

Furthermore, it is sufficient to show that v_1 can be moved to some earlier position in some series of small alterations of the sequence, provided the intermediaries are also valid pressing sequences.

We first show that if v_1 is not in X, there exists some valid X' containing v_1, and X' differs from X by exactly one vertex. This is true for any arbitrary vertex in G and we state it in a separate lemma since we are going to use it again later.

Lemma 8. *Assume that X is a successful pressing sequence on G and that vertex v is not a separated vertex in G. Then either v is in X or there exists some valid X' containing v, and X' differs from X by exactly one vertex.*

Proof. Let $X = u_1u_2 \cdots u_k$. Assume that v is not in X. Vertex v has at least one neighbor in G and none in GX; therefore there exists at least one vertex in X which, when pressed, is adjacent to v. Consider the last such vertex, which is in position i, and call it u_i; by definition, none of the vertex pressings in X^i affect the adjacencies or color of v, so after pressing u_i, v must be a white disconnected vertex. It follows that in $Gu_1 \cdots u_{i-1}$, the vertices v and u_i have exactly the same neighbors, and as such $u_1 \cdots u_{i-1}vu_{i+1} \cdots u_k$ is a valid pressing sequence. □

We now assume that v_1 is part of the current pressing sequence, which we denote by $P_1w_1v_1P_2$, where both P_1 and P_2 might be empty.

Case 1. If w_1 and v_1 are not neighbors in GP_1, then $P_1v_1w_1P_2$ is also a valid pressing sequence, and one of the longest common subsequences of $P_1w_1v_1P_2$ and $P_1v_1w_1P_2$ is $P_1w_1P_2$, one vertex less than the original pressing sequences. In this way, we can move v_1 to a smaller index position in the pressing sequence, and this is what we want to prove.

Case 2. If w_1 and v_1 are neighbors in GP_1, then v_1 is white in GP_1, and then pressing w_1 makes it black again. However, v_1 is black in G, since it is the first vertex in the valid pressing sequence Y. As such there must exist at least one vertex in P_1 which was adjacent to a black v_1 when pressed. Let w_2 be the last such vertex in P_1, and let us denote $P_1 = P_{1a}w_2P_{1b}$.

We claim that none of the vertices in P_{1b} are neighbors of w_2 in GP_{1a}. Indeed if there were such a neighbor, call it w_3, after pressing w_2, w_3 would be adjacent to v_1. Note that w_3 cannot have already been adjacent to v_1 by linearity of GP_{1a}. As such, pressing w_3 would change the color of v_1, meaning either v_1 was black prior to pressing w_1 — a contradiction — or there were further vertices in P_{1b} which were adjacent to a black v_1 when pressed, another contradiction.

Since P_{1b} does not contain a vertex which is a neighbor of w_2 in GP_{1a}, we move w_2 next to w_1. The new pressing sequence $P_{1a}P_{1b}w_2w_1v_1P_2$ is still a valid and successful pressing sequence and the longest common subsequence of P and $P_{1a}P_{1b}w_2w_1v_1P_2$ is $P_{1a}P_{1b}w_1v_1P_2$, one vertex less than the common length of the sequences.

For sake of simplicity, we denote $P_{1a}P_{1b}$ by P_1' and now we can assume the pressing sequence is of the form $P_1'w_2w_1v_1P_2$, with P_1' and P_2 both potentially

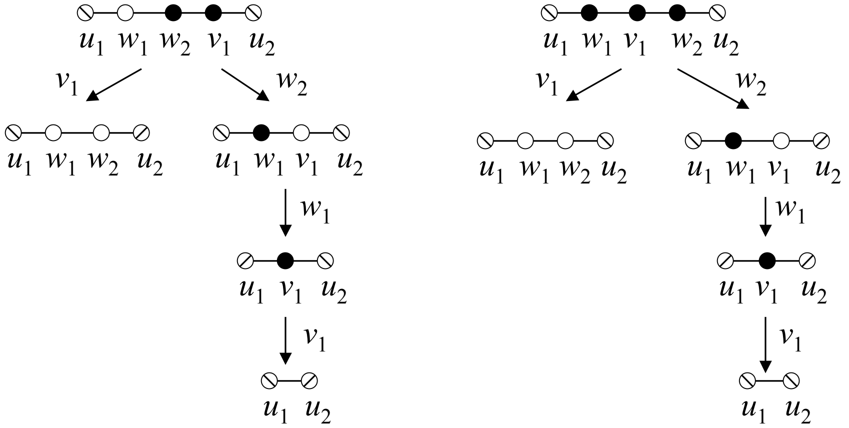
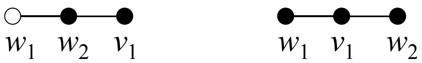


Figure 3. In the indicated two configurations, the neighbors of the $\{w_1, w_2, v_1\}$ triplet, u_1 and u_2, change color in the same way by pressing only v_1 and pressing $w_2w_1v_1$. The color change on u_1 and u_2 is indicated with the flipping of their crossing line.

empty. Since after pressing w_2, the vertices w_1 and v_1 become neighbors with w_1 being black and v_1 being white, the topology and colors of w_2, w_1 and v_1 in GP_1' is one of the following:



Case 2a. Assume that P_2 is not empty. The $\{w_1, w_2, v_1\}$ triplet has at least one neighbor (and at most two) in GP_1'; call them u_1 and u_2. Furthermore, either (1) one of u_1 and u_2 is pressed in P_2, or (2) we can replace some vertex in P_2 with u_1 or u_2 such that the resulting sequence is still valid, and successful on $GP_1'w_2w_1v_1$, due to Lemma 8. As such, we can assume that at least one neighbor of the $\{w_1, w_2, v_1\}$ triplet is pressed in P_2.

Without loss of generality, say u_1 is pressed before u_2 in P_2 and let $P_2 = P_{2a}u_1P_{2b}$. Note that we can press v_1 instead of $w_2w_1v_1$, and the resulting sequence $GP_1'v_1P_{2a}$ will be valid, as none of the vertices in P_{2a} are neighbors of w_2, w_1, or v_1. Next note from Figure 3 that the colors of u_1 and u_2 are identically altered in the pressing of either v_1 or $w_2w_1v_1$, and so we can press u_1. Figure 4 shows that the color of u_2 and a possible second neighbor of u_1 denoted by u_3 will be the same in $GP_1'w_2w_1v_1P_{2a}u_1$ and $GP_1'v_1P_{2a}u_1w_1w_2$. Therefore $P_1'v_1P_{2a}u_1w_1w_2P_{2b}$ will also be a successful pressing sequence on G, since no more vertices are affected by the given alteration of the pressing sequence. One of the longest common subsequences of $P_1'w_2w_1v_1P_{2a}u_1P_{2b}$ and $P_1'v_1P_{2a}u_1w_1w_2P_{2b}$ is $P_1'v_1P_{2a}u_1P_{2b}$, two vertices less than the entire pressing sequences. As intended, we have shown that v_1 is in a smaller index position of the pressing sequence.

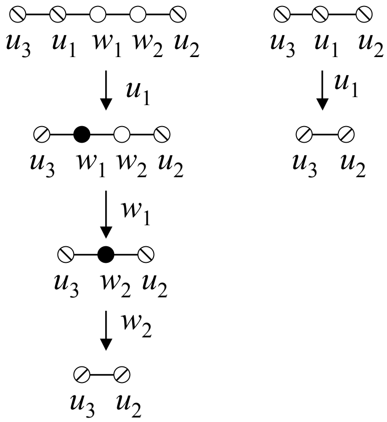


Figure 4. The color of u_2 and u_3 changes in the same way on the two indicated configurations.

Case 2b. Finally, assume that P_2 is empty. Then $GP_1'w_2w_1v_1$ is the all-white empty graph, and thus, $GP_1'w_2w_1$ contains the separated black v_1 and all separated white vertices, or contains a black v_1 connected to another black vertex and all separated and white vertices.

What follows is that GP_1' contains at most four nonisolated vertices, three of which are w_2, w_1, and v_1. Call the fourth u. If u exists, it must be black and adjacent to v_1 when v_1 is pressed. There are only four such cases, given the possible topologies for w_2, w_1, and v_1. If w_1 and w_2 are adjacent, then u is either black and adjacent to v_1 in GP_1' or it is adjacent to w_2 and is white. If w_2 and w_1 are not adjacent, then u can be adjacent to either w_2 or w_1, and must be white in both cases.

Note that all of these topologies can be described as follows; all neighbors of v_1 are black, v_1 is black, and all other vertices are white. This motivates the following lemma:

Lemma 9. *If GP is such that all neighbors of v_1 are black, v_1 is black, and all other vertices are white, and furthermore, there is a successful pressing sequence on G that starts with v_1, then there exists at least one vertex u in P such that when u is pressed u is not adjacent to v_1.*

Proof. Suppose instead that every vertex in P is adjacent to v_1 when pressed. P cannot be empty since then GP would be G and pressing v_1 in G would create an all-white nontrivial graph, contradicting that there exists a successful pressing sequence starting with pressing v_1. Furthermore, if all vertices in P are neighbors of v_1 when pressed, then P must contain an even number of vertices since v_1 is black both in G and GP.

Let $P = P_1'u_2u_1$. In order for u_1 and u_2 to be adjacent to v_1 when pressed, and for GP to fit the given criteria, GP_1' must also have v_1 and all neighbors black, and all other vertices white. By repeated application, we see that G must also fit these criteria. By assumption then, there are no black vertices not adjacent to v_1, and as such, pressing v_1 results in an all-white nontrivial graph. However, this is a contradiction, as there exists a successful pressing sequence for G in which v_1 is pressed first. □

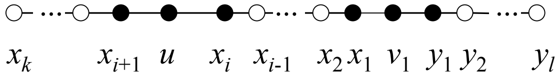
From the above lemma, we have that there exists some vertex in P_1' not adjacent to v_1 when pressed, and there are vertices which are adjacent to v_1 when pressed. For technical reasons, we have to separate them in the pressing sequence, which is doable due to the following lemma.

Lemma 10. *Let Pxu be a valid pressing sequence on G such that x is a neighbor of some v in GP and u is not a neighbor of v in GPx. Then Pux is a valid pressing sequence on G and $GPxu = GPux$.*

Proof. It is sufficient to show that x and u are not neighbors in GP. If x and u were neighbors, then the two neighbors of x would be u and v, causing u and v to become neighbors in GPx, a contradiction. □

Due to Lemma 10 it is possible to "bubble up" vertices that are not neighbors of v_1 in the pressing sequence so that the pressing sequence becomes $P_uP_nv_1$, where P_u contains the vertices that are not neighbors of v_1 when pressed and P_u contains the vertices that are neighbors of v_1 when pressed. Each bubbling-up step is allowed since the length of the longest common subsequence of two consecutive sorting sequences is their common length minus 1. We know that neither P_u nor P_n is empty due to Lemma 9 and due to the fact that w_1 and w_2 are in P_n.

Let u be the last vertex in P_u and let $P_u = P_u'u$. Without loss of generality, we can assume that u is on the left-hand side of v_1 in GP_u' and then GP_u' is



The vertices on the left-hand side of v_1 are denoted by $x_1, x_2, \ldots, x_k$ and we distinguish u amongst them. The vertices on the right-hand side of v_1 are denoted by $y_1, y_2, \ldots, y_l$.

Obviously, no x is a neighbor of any y when pressed, so we can bubble up the y vertices in P_n such that first the y vertices are pressed and then the x vertices. After a finite number of allowed alterations, $P_n = y_1y_2\cdots y_lx_1x_2\cdots x_k$.

Similarly to the previous cases, we can move down vertex u in the pressing sequence before x_i. We know that v_1 is black in $GP_u'u$ since it is black in G and neither of its neighbors is pressed in $P_u'u$. We are going to press some of the vertices amongst the x and y vertices provided that v_1 will be black after that

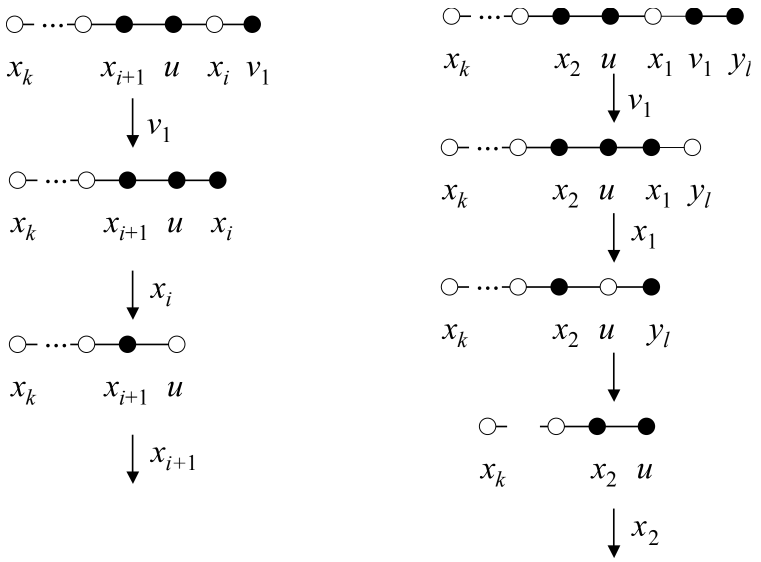
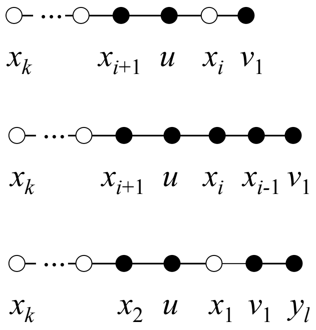

x_k x_{i+1} u x_i v_1
v_1
x_k x_{i+1} u x_i
x_i
x_k x_{i+1} u
x_{i+1}
x_k x_2 u x_1 v_1 y_l
v_1
x_k x_2 u x_1 y_l
x_1
x_k x_2 u y_l
x_k x_2 u
x_2

Figure 5. Alternative pressing sequences for two cases.

series of pressing. We consider the graph $GP'_u y_1 \cdots y_l x_1 \cdots x_{i-1}$ if v_1 is black in it (the runs of x vertices might be empty if $i = 1$), and otherwise the graph $GP'_u y_1 \cdots y_l x_1 \cdots x_{i-2}$ (also the runs of x vertices might be empty if $i = 2$) or $GP'_u y_1 \cdots y_{l-1}$ if $i = 1$ and the number of y vertices is odd (if $i = 1$ and the number of y vertices is even, then v_1 will be black in $GP'_u y_1 \cdots y_l$). We have one of the following graphs


x_k x_{i+1} u x_i v_1
x_k x_{i+1} u x_i $x_{i\text{-}1}$ v_1
x_k x_2 u x_1 v_1 y_l

on which $ux_i \cdots x_k v_1$, $ux_{i-1} \cdots x_k v_1$, $y_l ux_1 \cdots x_k v_1$ is the current successful pressing sequence, respectively.

A successful pressing sequence replacing $ux_i \cdots x_k v_1$ is $v_1 x_i \cdots x_k u$, as can be seen on the left-hand side of Figure 5. The length of the longest common subsequence of the two pressing sequences is 2 less than their common length, as required. The pressing sequence $y_l ux_1 \cdots x_k v_1$ can be replaced by $ux_1 y_l x_2 \cdots x_k v_1$ since y_l is a neighbor of neither u nor x_1. Then this pressing sequence can be

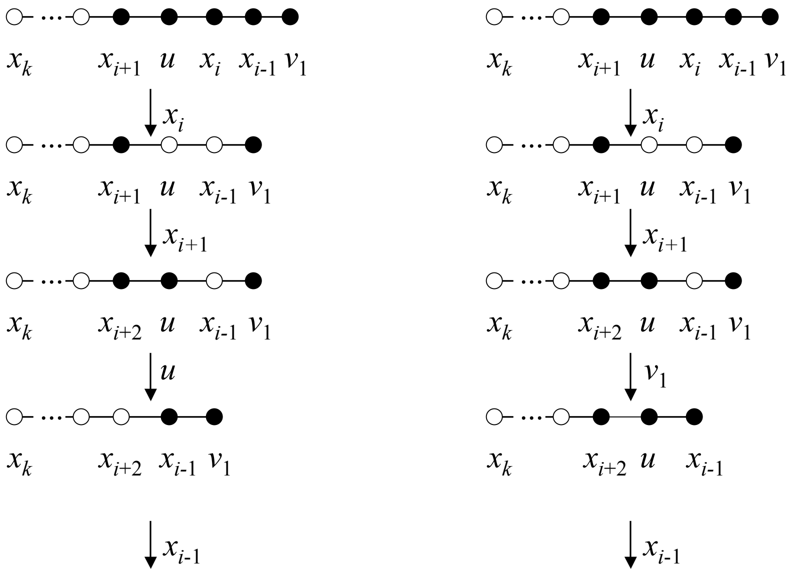

x_k x_{i+1} u x_i x_{i-1} v_1
x_i
x_k x_{i+1} u x_{i-1} v_1
x_{i+1}
x_k x_{i+2} u x_{i-1} v_1
u
x_k x_{i+2} x_{i-1} v_1
x_{i-1}
x_k x_{i+1} u x_i x_{i-1} v_1
x_i
x_k x_{i+1} u x_{i-1} v_1
x_{i+1}
x_k x_{i+2} u x_{i-1} v_1
v_1
x_k x_{i+2} u x_{i-1}
x_{i-1}

Figure 6. Changing the pressing sequence $ux_{i-1} \cdots x_k v_1$ in two steps such that v_1 is in a smaller index position.

replaced by $v_1 x_1 y_l x_2 \cdots x_k u$, as can be seen on the right-hand side of Figure 5. The length of the longest common subsequence of $ux_1 y_l x_2 \cdots x_k v_1$ and $v_1 x_1 y_l x_2 \cdots x_k u$ is again 2 less than their common length.

Finally, the pressing sequence $ux_{i-1} \cdots x_k v_1$ can be replaced in two steps; first it is changed to $x_i x_{i+1} u x_{i-1} x_{i+2} \cdots x_k v_1$, then to $x_i x_{i+1} v_1 x_{i-1} x_{i+2} \cdots x_k u$, as can be checked in Figure 6. In both steps, the length of the longest common subsequences of two consecutive pressing sequences is 2 less than their common length as required.

We proved that in any case, v_1 can be moved into a smaller index position with a finite series of allowed perturbations. Iterating this, we can move v_1 to the first position. Then we can do the same thing with v_2 on the graph $Gv_1 \setminus \{v_1\}$, and eventually transform X into Y with allowed perturbations. □

4. Discussion and conclusions

In this paper, we proved the pressing game conjecture for linear graphs. Although the linear graphs are very simple, this proof technique provides a direction for proving the general case. Indeed, it is generally true that if a vertex v is not in a successful pressing sequence P, then a successful pressing sequence P' exists which contains v and the length of the longest common subsequence of P and P' is only 1 less than their common length. Case 1 in the proof of Theorem 7 holds for arbitrary graphs, and in a working manuscript, we were able to prove that the conjecture is true for Case 2a using linear algebraic techniques similar to that used in [Hartman and Verbin 2006]. The only missing part is Case 2b, which seems to be very complicated for general graphs; for example, Lemma 10 cannot be generalized for arbitrary graphs.

A stronger theorem holds for the linear case than is conjectured for the general case. One possible direction above proving the general conjecture is to study the emerging Markov chain on the solution space of the pressing game on linear graphs. We proved that a Markov chain that randomly removes two vertices from the current pressing sequence, adds two random vertices to it, and accepts it if the result is a successful pressing sequence is irreducible. It is easy to set the jumping probabilities of the Markov chain such that it converges to the uniform distribution of the solutions. The remaining question is the speed at which this Markov chain converges.

Acknowledgements

István Miklós was supported by OTKA grant PD84297. Alexey Medvedev is thanked for fruitful discussions.

References

[Bergeron 2001] A. Bergeron, "A very elementary presentation of the Hannenhalli–Pevzner theory", pp. 106–117 in *Combinatorial pattern matching* (Jerusalem, 2001), edited by A. Amir and G. M. Landau, Lecture Notes in Comput. Sci. **2089**, Springer, Berlin, 2001. MR 1904571 Zbl 0990.68050

[Hannenhalli and Pevzner 1995] S. Hannenhalli and P. A. Pevzner, "Transforming men into mice (polynomial algorithm for genomic distance problem)", pp. 581–592 in *36th Annual Symposium on Foundations of Computer Science* (Milwaukee, WI, 1995), IEEE Comput. Soc. Press, Los Alamitos, CA, 1995. MR 1619106 Zbl 0938.68939

[Hannenhalli and Pevzner 1999] S. Hannenhalli and P. A. Pevzner, "Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals", *J. ACM* **46**:1 (1999), 1–27. MR 2000j:92013 Zbl 1064.92510

[Hartman and Verbin 2006] T. Hartman and E. Verbin, "Matrix tightness: a linear-algebraic framework for sorting by transpositions", pp. 279–290 in *String processing and information retrieval* (Glasgow, 2006), edited by F. Crestani et al., Lecture Notes in Comput. Sci. **4209**, Springer, Berlin, 2006. MR 2337809

[Ma et al. 2008] J. Ma, A. Ratan, B. J. Raney, B. B. Suh, W. Miller, and D. Haussler, "The infinite sites model of genome evolution", *Proc. Nat. Acad. Sci. USA* **105**:38 (2008), 14254–14261.

[Miklós et al. 2010] I. Miklós, B. Mélykúti, and K. Swenson, "The metropolized partial importance sampling MCMC mixes slowly on minimum reversal rearrangement paths", *IEEE/ACM Trans. Comput. Biol. Bioinform.* **4**:7 (2010), 763–767.

[Sturtevant 1921] A. H. Sturtevant, "A case of rearrangement of genes in Drosophila", *Proc. Nat. Acad. Sci. USA* **7**:8 (1921), 235–237.

[Sturtevant and Novitski 1941] A. H. Sturtevant and E. Novitski, "The homologies of chromosome elements in the genus Drosophila", *Genet.* **26** (1941), 517–541.

[Sturtevant and Tan 1937] A. H. Sturtevant and C. C. Tan, "The comparative genetics of Drosophila pseudoobscura and D. melanogaster", *J. Genet.* **34** (1937), 415–432.

Received: 2013-04-08 Revised: 2015-01-19 Accepted: 2015-01-21

eli.bixby@gmail.com
Budapest Semesters in Mathematics, H-1071 Budapest,
Bethlen Gábor tér 2, Hungary

tobycollege@gmail.com
Budapest Semesters in Mathematics, H-1071 Budapest,
Bethlen Gábor tér 2, Hungary

miklosi@renyi.hu
Rényi Institute, H-1053 Budapest,
Reáltanoda utca 13-15, Hungary

Budapest Semesters in Mathematics, H-1071 Budapest,
Bethlen Gábor tér 2, Hungary

mathematical sciences publishers msp

INVOLVE 9:1 (2016)
dx.doi.org/10.2140/involve.2016.9.57

Polygonal bicycle paths and the Darboux transformation

Ian Alevy and Emmanuel Tsukerman

(Communicated by Kenneth S. Berenhaut)

A bicycle (n, k)-gon is an equilateral n-gon whose k diagonals are of equal length. In this paper we introduce periodic bicycle (n, k)-paths, which are a natural variation in which the polygon is replaced with a periodic polygonal path, and study their rigidity and integrals of motion.

1. Background

Our motivation comes from three seemingly unrelated problems. The first is the problem of *floating bodies of equilibrium in two dimensions*. From 1935 to 1941, mathematicians at the University of Lviv, among them Stefan Banach and Mark Kac, collected mathematical problems in a book, which became known as "the Scottish book", since they often met in the Scottish Coffee House. Stanislaw Ulam posed problem 19 of this book: "Is a sphere the only solid of uniform density which will float in water in any position?" The answer in the two-dimensional case, as it turns out, depends on the density of the solid.

The second problem, known as the *tire track problem*, originated in the story, "The adventure of the priory school" by Arthur Conan Doyle, where Sherlock Holmes and Dr. Watson discuss in view of the two tire tracks of a bicycle which way the bicycle went. The problem is: "Is it possible that tire tracks other than circles or straight lines are created by bicyclists going in both directions?" As shown in Figure 1, the answer to this subtle question is affirmative.

The third problem is that of describing the trajectories of *electrons in a parabolic magnetic field*. All three problems turn out to be equivalent [Wegner 2007].

Often in mathematics it is fruitful to discretize a problem. As such, S. Tabachnikov [2006] proposed a "discrete bicycle curve" (also known as a "bicycle polygon"), which is a polygon satisfying discrete analogs of the properties of a bicycle track. The main requirement turns out to be that, in the language of discrete differential geometry, the polygon is "self-Darboux". That is, the discrete differential

MSC2010: 37J35, 52C25.
Keywords: bicycle polygons, tire track problem, floating bodies in equilibrium.

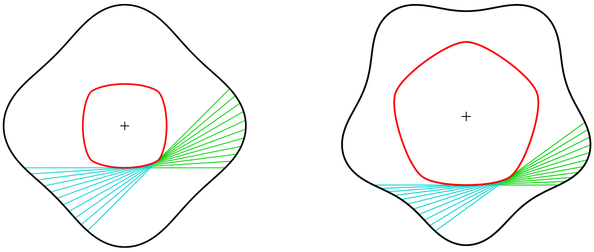


Figure 1. Ambiguous bicycle tracks. The rear-wheel track is the inner curve and the front-wheel track is the outer curve. One cannot tell which way the bicycle went because a bicycle could have followed either one of two trajectories [Wegner 2007].

geometric notion of a discrete Darboux transformation [Bobenko and Suris 2008; Tsuruga 2010], which relates one polygon to another, relates a discrete bicycle curve to itself.

The topic of bicycle curves and polygons belongs to a number of active areas of research. On the one hand, it is part of rigidity theory. As an illustration, R. Connelly and B. Csikós [2009] consider the problem of classifying first-order flexible regular bicycle polygons. Other work on the rigidity theory of bicycle curves and polygons can be found in [Csikós 2007; Cyr 2012; Tabachnikov 2006].

The topic is also part of the subject of discrete integrable systems. This point of view is taken in [Tabachnikov and Tsukerman 2013], where the authors find integrals of motion (i.e., quantities which are conserved) of bicycle curves and polygons under the Darboux transformation and recutting of polygons [Adler 1993; 1995].

In this paper, in analogy with bicycle polygons, we introduce a new concept called a periodic discrete bicycle path and study both its rigidity and integrals.

2. Bicycle (n, k)-paths

A bicycle (n, k)-gon is an equilateral n-gon whose k diagonals are of equal length [Tabachnikov 2006]. We consider the following analog.

Definition 1. Define $P = \{V_i \in \mathbb{R}^2 : i \in \mathbb{Z}\}$ (for brevity, $V_0 V_1 \cdots V_{n-1}$) to be a *discrete periodic bicycle* (n, k)*-path* (or discrete (n, k)-path) if the following conditions hold:

(i) $V_{n+i} = V_i + e_1$ for all i, where $e_1 = (1, 0)$ and $V_0 = (0, 0)$ (periodicity condition).

(ii) $|V_i V_{i+1}| = |V_j V_{j+1}|$ for all i, j (equilateralness).

(iii) $|V_i V_{i+k}| = |V_j V_{j+k}|$ for all i, j (equality of k-diagonals).

Definition 1 is meant to model the motion of a bicycle whose trajectory is spatially periodic. The condition that $|V_j V_{j+1}|$ is independent of j prescribes a constant speed for the motion of the bike. The condition that $|V_j V_{j+k}|$ is independent of j

represents the ambiguity of the direction in which the bicycle went (see [Tabachnikov 2006] for details).

Some natural questions regarding periodic (n, k)-paths are for which pairs (n, k) they exist, how many there are and whether they are rigid or flexible. We consider these questions in Section 3. A simple example of a bicycle (n, k)-path, analogous to the regular (n, k)-polygon, is $V_i = (i/n, 0)$, i.e., when all vertices lie at equal intervals on the line. We call this the *regular path*. Since bicycle (n, k)-paths are discretized bicycle paths, it is also interesting to see if there are any integrals of motion. We show that this is indeed the case in Section 4, by showing that area is an integral of motion.

3. Rigidity

The following two lemmas will be helpful when analyzing the rigidity of discrete bicycle paths.

Lemma 2. *Let $n \in \mathbb{N}$, $\chi_i \in \{-1, 1\}$ for every $i \in \mathbb{Z}/n\mathbb{Z}$ and let*

$$S = \{(x_0, x_1, \dots, x_{n-1}) \in \mathbb{R}^n : (x_{i+1} - x_i)^2 = (x_{j+1} - x_j)^2 \ \textit{for all } i, j \in \mathbb{Z}/n\mathbb{Z}\}.$$

Then

$$S = \Bigl\{(x_0, x_1, \dots, x_{n-1}) : x_{i+1} = x_i + \chi_i r \ \textit{ for } i \in \mathbb{Z}/n\mathbb{Z}, \sum_{i=0}^{n-1} r\chi_i = 0 \textit{ and } r \geq 0\Bigr\}.$$

In particular, if n is odd, then $S = \{(x_0, x_1, \dots, x_{n-1}) : x_i = x_j$ for all $i, j \in \mathbb{Z}/n\mathbb{Z}\}$.

Proof. First note that the candidate set is well-defined since

$$x_{j+n} = x_j + \sum_{i=j}^{j+n-1} r\chi_i = x_j + \sum_{i=0}^{n-1} r\chi_i = x_j.$$

Let $(x_0, x_1, \dots, x_{n-1}) \in S$. Recall that

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0, \\ -1 & \text{if } x < 0, \end{cases}$$

and that $\operatorname{sgn}(x)|x| = x$. Set $r := |x_{i+1} - x_i|$ and $\chi_i = \operatorname{sgn}(x_{i+1} - x_i) + (1 - \operatorname{sgn}(r))$. Then

$$x_{i+1} = x_i + \chi_i r,$$

and

$$\sum_{i=0}^{n-1} r \operatorname{sgn}(x_{i+1} - x_i) = 0.$$

It follows that any n-tuple in S satisfies the conditions $x_{i+1} = x_i + \chi_i r$, $\sum_{i=0}^{n-1} r\chi_i = 0$ and $r \geq 0$. The opposite inclusion is clear. □

Lemma 3. *Let $x_i \in \mathbb{R}$ for every $i \in \mathbb{Z}$ with $x_0 = 0$ and let k and n be coprime integers. Assume that $x_{i+k} - x_i = x_i - x_{i-k}$ for each i and that $x_{i+n} = 1 + x_i$. Then $x_i = i/n$ for each i.*

Proof. Define z_i via $x_i = z_i + i/n$. The hypothesis $x_{i+n} = 1 + x_i$ implies that

$$z_{i+n} = z_i.$$

The difference

$$\Delta z := z_{i+k} - z_i$$

is independent of i due to the assumption that $x_{i+k} - x_i = x_i - x_{i-k}$ and because k and n are coprime. This implies that

$$0 = z_{i+n} - z_i = z_{i+nk} - z_i = n\Delta z.$$

It follows that $z_i = 0$ for every i. □

The following theorem gives a classification of a family of periodic (n, k)-paths.

Theorem 4. *The discrete $(n, dn-1)$-paths $V_i = (x_i, y_i)$, $i \in \mathbb{N}$ with $d \neq 0$ are exactly those paths which satisfy*

$$x_j = \frac{j}{n}$$

and

$$y_{j+1} = y_j + \chi_j r \quad \text{for } j \in \mathbb{Z}/n\mathbb{Z} \text{ with } \sum_{i=0}^{n-1} r\chi_i = 0 \text{ and } r \geq 0$$

for each j. In particular, if n is odd then a discrete $(n, dn-1)$-path must be regular.

Proof. For every i,

$$|V_i V_{i+1}| = |V_{i+dn-1} V_{i+dn}|,$$
$$|V_i V_{i+dn-1}| = |V_{i+1} V_{i+dn}|.$$

Therefore

$$(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 = (x_i - x_{i-1})^2 + (y_i - y_{i-1})^2,$$
$$(d + x_{i-1} - x_i)^2 + (y_{i-1} - y_i)^2 = (d + x_i - x_{i+1})^2 + (y_i - y_{i+1})^2.$$

It follows that

$$d(x_{i+1} - x_i) = d(x_i - x_{i-1}).$$

Since $d \neq 0$,

$$x_{i+1} - x_i = x_i - x_{i-1}.$$

By Lemma 3, $x_j = j/n$ for each j. Now equation $|V_i V_{i+1}| = |V_j V_{j+1}|$ for all i, j implies that

$$(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 = (x_{j+1} - x_j)^2 + (y_{j+1} - y_j)^2 \quad \text{for all } i, j,$$

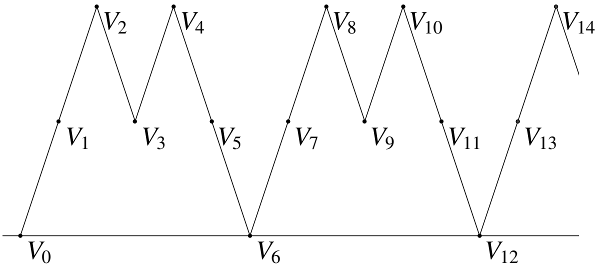


Figure 2. An example of a discrete (6, 5)-path.

so that

$$(y_{i+1}-y_i)^2=(y_{j+1}-y_j)^2.$$

By Lemma 2, we are done. □

Theorem 5. *The discrete* $(n, dn+1)$*-paths* $V_i=(x_i, y_i), i\in\mathbb{N}$ *with* $d\neq 0$ *are exactly those paths which satisfy*

$$x_j=\frac{j}{n}$$

and

$$y_{j+1}=y_j+\chi_j r \quad \textit{for } j\in\mathbb{Z}/n\mathbb{Z} \textit{ with } \sum_{i=0}^{n-1} r\chi_i=0 \textit{ and } r\geq 0$$

for each j*. In particular, if* n *is odd then a discrete* $(n, dn+1)$*-path must be regular.*

Proof. Set $C_1=|V_iV_{i+dn+1}|^2$ and $C_2=|V_iV_{i+1}|^2$. Then

$$(d+x_{i+1}-x_i)^2+(y_{i+1}-y_i)^2=C_1,$$
$$(x_{i+1}-x_i)^2+(y_{i+1}-y_i)^2=C_2.$$

Substituting, we get

$$d^2+2d(x_{i+1}-x_i)+C_2=C_1,$$

so that $x_{i+1}-x_i$ is constant. By Lemma 3, $x_i=i/n$. It follows that $(y_{i+1}-y_i)^2$ is constant, so by Lemma 3 we are done. □

Corollary 6. *Any* $(n, dn+1)$*-path is an* $(n, dn-1)$*-path and vice versa.*

For an example, see Figure 2.

4. Darboux transformation and integrals

It is important to make a distinction between infinitesimal "trapezoidal" movement and infinitesimal "parallelogram" movement of the bicycle. Consider a pair of conjoined bikes, sharing a back wheel and facing in opposite directions. Since this bicycle moves in such a way that the distance between the turnable wheels is constant, at each moment of time the turnable wheels must enclose equal angles with the line of the frame. When the two turnable wheels are parallel, the trike is gliding, but then the common back wheel of the bikes is slipping, which is not allowed. That is why we exclude parallelogram movements from consideration for the remainder of this paper.

Definition 7 (trapezoidal condition). We will say that a discrete (n, k)-path satisfies the trapezoidal condition if $V_i V_{i+k+1}$ and $V_{i+1} V_{i+k}$ are parallel for each $i \in \mathbb{Z}$.

As an illustration of these concepts, consider Figure 2: $V_0 V_1 V_5 V_6$ is a trapezoidal motion, while $V_1 V_2 V_6 V_7$ is a parallelogram motion. Consequently, the bicycle path in the figure does not satisfy the trapezoidal condition.

Assuming the trapezoidal condition, we may view bicycle paths in terms of an important construction in discrete differential geometry called the Darboux transformation [Bobenko and Suris 2008; Tsuruga 2010].

Definition 8 (Darboux transform). We say that two polygons $P = P_1 P_2 \cdots$ and $Q = Q_1 Q_2 \cdots$ are in Darboux correspondence if for each $i = 1, 2, \ldots$, we have that Q_{i+1} is the reflection of P_i in the perpendicular bisector of the segment $P_{i+1} Q_i$.

If segment $P_1 Q_1$ is of length ℓ then for each i, $P_i Q_i$ is of length ℓ. We then say that P and Q are in Darboux correspondence with parameter ℓ. We also note that each quadrilateral $P_i Q_i P_{i+1} Q_{i+1}$ is an isosceles trapezoid.

We denote the map taking vertex P_i to Q_i by $\mathfrak{D}$. We will also refer to the map of polygons $\mathfrak{D}(P) = Q$ by the same letter, since no confusion ought to occur.

Consider a polygonal line P with vertices $V_0, V_1, \ldots, V_{n-1}$. Let v_0 be a vector with its origin at V_0. Having a vector v_i at vertex V_i, we obtain a vertex v_{i+1} of the same length at V_{i+1} via the trapezoidal condition. For example, in Figure 3, $v_1 = P_1 Q_1$ and $v_2 = P_2 Q_2$. For a fixed length of v_0, we may view the map taking v_0 to v_j as a self-map of the circle of radius $|v_0| = |v_j|$ by identifying the circle at V_0 with circle at V_j via parallel translation.

Definition 9 (monodromy map of the Darboux transformation). The monodromy map is the map acting on the identified circles at V_0 and V_n which takes v_0 to v_n.

It is known that the monodromy map is a cross-ratio preserving transformation (in terms of affine coordinates, a fractional linear transformation) on a circle of fixed radius after we identify the circle with the real projective line via stereographic projection [Tabachnikov and Tsukerman 2013]. We will assume throughout, unless otherwise stated, that the monodromy map is acting on a fixed point; in other words, we will assume that the Darboux transform has been chosen so that the initial vector v_0 is equal to the vector v_n, where n is the period. This is analogous to applying the Darboux transform to a closed polygon and requiring that its image is closed also.

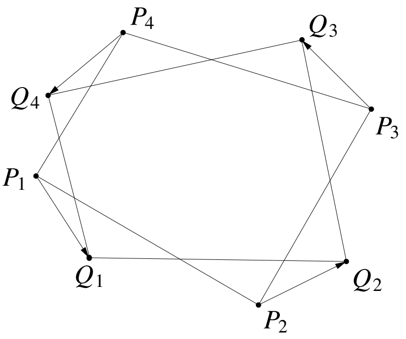


Figure 3. Two polygons in Darboux correspondence.

We mention in passing that in the case of closed polygons, Darboux correspondence implies that the monodromies of the two polygons are conjugated to each other. The invariants of the conjugacy class of the monodromy, viewed as functions of the length parameter, are consequently integrals of the Darboux correspondence [Tabachnikov and Tsukerman 2013].

Connection between Darboux transformation and discrete (n, k)-paths. A discrete (n, k)-path satisfying the trapezoidal condition may be interpreted in terms of the Darboux transform. Indeed, given such a path, we consider the periodic equilateral linkages $L_i = \cdots V_{0+i} V_{k+i} V_{2k+i} \cdots$ for $i = 0, 1, \ldots, k-1$. The trapezoidal condition implies that there is a Darboux correspondence $\mathfrak{D}(L_i) = L_{i+1}$ of the same parameter (since the (n, k)-path is equilateral) for consecutive linkages (see Figure 4).

The Darboux transformation also preserves the area of periodic paths. More precisely, let $y = -c$ for $c > 0$ sufficiently large so that the periodic path P and its Darboux transformation P' lie completely above $y = -c$. We define an area function as follows. Let $\check{V}_i = (x(V_i), -c)$. We define the area of P to be the signed area of the polygon $\check{V}_0 V_0 V_1 \cdots V_n \check{V}_n$ and denote it by $|P|$. We show that this area is preserved under Darboux transformation (see Figure 5). In particular, it will follow that the area of $V_0 V_k \cdots V_{nk}$ is equal to the area of $V_m V_{k+m} \cdots V_{nk+m}$ for every $m \in \mathbb{Z}$.

Theorem 10. *The Darboux transformation is area-preserving on periodic polygonal paths.*

Proof. Let P and P' be two periodic polygonal paths in Darboux correspondence. We show that the difference of the areas of P and P' is zero. We denote the vertex of P' which corresponds via the Darboux transformation to the vertex V_i in P by V_i' for each i. We have

$$|P| = \sum_{i=0}^{n-1} |\check{V}_i V_i V_{i+1} \check{V}_{i+1}|,$$

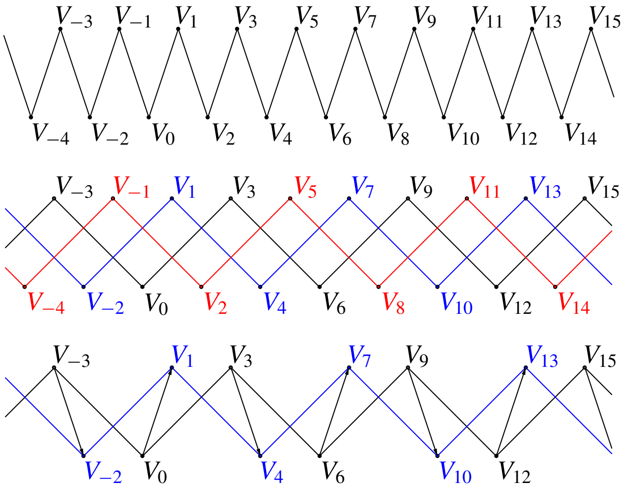


Figure 4. Viewing a discrete (n,k)-path satisfying the trapezoidal condition (top) in terms of the Darboux transformation. The path is decomposed into equilateral linkages (middle). Any two consecutive linkages are in Darboux correspondence (bottom).

and similarly for P'. Therefore

$$|P|-|P'|=\sum_{i=0}^{n-1}|\check{V}_i V_i V_{i+1}\check{V}_{i+1}|-|\check{V}_i' V_i' V_{i+1}'\check{V}_{i+1}'|.$$

From the isosceles trapezoids,

$$|V_i V_{i+1} V_{i+1}'|=|V_i' V_{i+1}' V_i|. \tag{4-1}$$

Also,

$$|\check{V}_i V_i V_{i+1}\check{V}_{i+1}|=|\check{V}_i V_i V_{i+1}'\check{V}_{i+1}'|+|\check{V}_{i+1}' V_{i+1}' V_{i+1}\check{V}_{i+1}|+|V_i V_{i+1} V_{i+1}'|.$$

Similarly,

$$|\check{V}_i' V_i' V_{i+1}'\check{V}_{i+1}'|=|\check{V}_i' V_i' V_i\check{V}_i|+|\check{V}_i V_i V_{i+1}'\check{V}_{i+1}'|+|V_i' V_{i+1}' V_i|.$$

Using (4-1),

$$|\check{V}_i V_i V_{i+1}\check{V}_{i+1}|-|\check{V}_i' V_i' V_{i+1}'\check{V}_{i+1}'|=|\check{V}_{i+1}' V_{i+1}' V_{i+1}\check{V}_{i+1}|-|\check{V}_i' V_i' V_i\check{V}_i|.$$

It follows that

$$|P|-|P'|=\sum_{i=0}^{n-1}|\check{V}_{i+1}' V_{i+1}' V_{i+1}\check{V}_{i+1}|-|\check{V}_i' V_i' V_i\check{V}_i|,$$

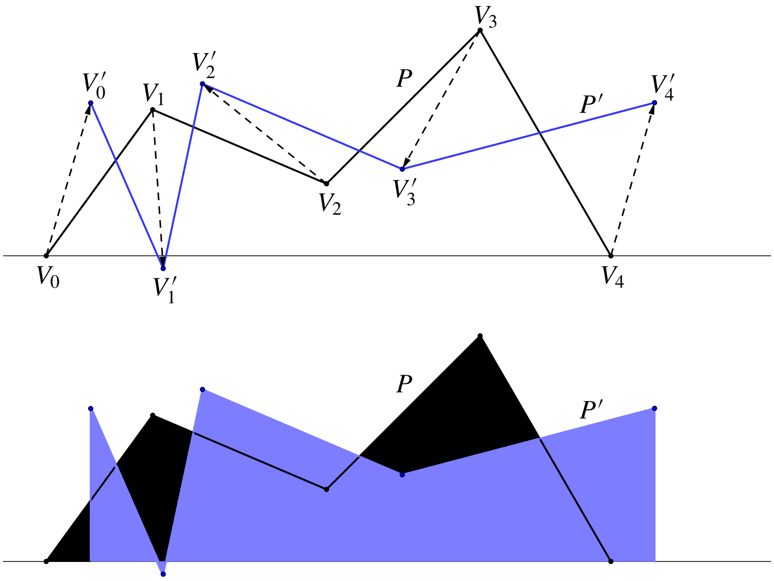


Figure 5. Two periodic paths P and P' in Darboux correspondence. By Theorem 10, the two paths have equal areas under the curve.

which telescopes to

$$|P| - |P'| = |\check{V}_n' V_n' V_n \check{V}_n| - |\check{V}_0' V_0' V_0 \check{V}_0|.$$

Since $V_n' = V_0' + e_1$ and $V_n = V_0 + e_1$, it follows that $\overrightarrow{V_n V_n'} = \overrightarrow{V_0 V_0'}$ and $|\check{V}_n' V_n' V_n \check{V}_n| = |\check{V}_0' V_0' V_0 \check{V}_0|$, so that $|P| = |P'|$. □

5. Questions

We end our discussion with some research topics and questions of interest concerning bicycle (n, k)-paths.

(1) Construct interesting families of bicycle (n, k)-paths. For example, ones for which the condition $x_j = j/n$ does not hold.

(2) What is the m-th order ($m \in \mathbb{N}$) infinitesimal rigidity theory of bicycle (n, k)-paths like?

(3) For closed bicycle polygons, there are many integrals of motion [Tabachnikov and Tsukerman 2013]. For example, a geometric center called the circumcenter of mass [Tabachnikov and Tsukerman 2014] is invariant under Darboux transformation for closed polygons. Are there other integrals of motion for bicycle (n, k)-paths?

Acknowledgments

It is a pleasure to acknowledge the valuable help and discussion with S. Tabachnikov. This project originated during the program Summer@ICERM 2012; we are grateful to ICERM for the support. We are also thankful to the anonymous referee for valuable suggestions.

References

[Adler 1993] V. È. Adler, "Recuttings of polygons", *Funktsional. Anal. i Prilozhen.* **27**:2 (1993), 79–82. In Russian; translated in *Funct. Anal. Appl.* **27**:2 (1993), 141–143. MR 94j:58072 Zbl 0812.58072

[Adler 1995] V. È. Adler, "Integrable deformations of a polygon", *Phys. D* **87**:1–4 (1995), 52–57. MR 96m:58100 Zbl 1194.35353

[Bobenko and Suris 2008] A. I. Bobenko and Y. B. Suris, *Discrete differential geometry: Integrable structure*, Graduate Studies in Mathematics **98**, American Mathematical Society, Providence, RI, 2008. MR 2010f:37125 Zbl 1158.53001

[Connelly and Csikós 2009] R. Connelly and B. Csikós, "Classification of first-order flexible regular bicycle polygons", *Studia Sci. Math. Hungar.* **46**:1 (2009), 37–46. MR 2011d:52042 Zbl 1240.11057

[Csikós 2007] B. Csikós, "On the rigidity of regular bicycle (n, k)-gons", *Contrib. Discrete Math.* **2**:1 (2007), 93–106. MR 2007m:37144 Zbl 1191.52001

[Cyr 2012] V. Cyr, "A number theoretic question arising in the geometry of plane curves and in billiard dynamics", *Proc. Amer. Math. Soc.* **140**:9 (2012), 3035–3040. MR 2917076 Zbl 1282.37023

[Tabachnikov 2006] S. Tabachnikov, "Tire track geometry: variations on a theme", *Israel J. Math.* **151** (2006), 1–28. MR 2007d:37091 Zbl 1124.52005

[Tabachnikov and Tsukerman 2013] S. Tabachnikov and E. Tsukerman, "On the discrete bicycle transformation", *Publ. Mat. Urug.* **14** (2013), 201–219. MR 3235356 Zbl 1317.51027

[Tabachnikov and Tsukerman 2014] S. Tabachnikov and E. Tsukerman, "Circumcenter of mass and generalized Euler line", *Discrete Comput. Geom.* **51** (2014), 815–836. MR 3216665 Zbl 1301.51023

[Tsuruga 2010] M. Tsuruga, "Discrete Differential Geometry", Lecture notes, Freie Universität, 2010, available at http://boolesrings.org/matsguru/files/2012/11/DDG1.pdf.

[Wegner 2007] F. Wegner, "Floating bodies of equilibrium in 2D, the tire track problem and electrons in a parabolic magnetic field", preprint, 2007. arXiv physics/0701241

Received: 2013-09-07 Revised: 2014-09-01 Accepted: 2014-12-08

ian_alevy@brown.edu *Division of Applied Mathematics, Brown University,*
Providence, RI 02912, United States

e.tsukerman@berkeley.edu *Department of Mathematics, University of California, Berkeley,*
Berkeley, CA 94720, United States

INVOLVE 9:1 (2016)
dx.doi.org/10.2140/involve.2016.9.67

Local well-posedness of a nonlocal Burgers' equation

Sam Goodchild and Hang Yang

(Communicated by Martin Bohner)

In this paper, we explore a nonlocal inviscid Burgers' equation. Fixing a parameter h, we prove existence and uniqueness of the local solution of the equation $u_t + \big(u(x+h,t) \pm u(x-h,t)\big)u_x = 0$ with given periodic initial condition $u(x,0) = u_0(x)$. We also explore the blow-up properties of the solutions to this Cauchy problem, and show that there exist initial data that lead to finite-time-blow-up solutions and others to globally regular solutions. This contrasts with the classical inviscid Burgers' equation, for which all nonconstant smooth periodic initial data lead to finite-time blow-up. Finally, we present results of simulations to illustrate our findings.

1. Introduction

Burgers' equation is a common equation that arises naturally in the study of fluid mechanics, traffic, and other fields. It is a relatively simple partial differential equation that has been extensively studied. In finite time, solutions to the inviscid Burgers' equation are known to develop shock waves and rarefactions for smooth initial data. It also serves as a basic example of conservation laws. Many different closed forms, series approximations, and numerical solutions are known for particular sets of boundary conditions.

The more general form of dissipative Burgers' equation is

$$\frac{\partial u}{\partial t} + u \cdot \nabla u = \gamma \Delta u, \tag{1-1}$$

where $u(x,t)$ represents the velocity at point $(x,t) \in \mathbb{R}^d \times \mathbb{R}^+$, $\gamma \in \mathbb{R}^+$, and the term on the right-hand side is the viscosity term which induces diffusion properties. For the inviscid one-dimensional case, Burgers' equation reduces to

$$\frac{\partial u}{\partial t} + u \cdot \frac{\partial u}{\partial x} = 0. \tag{1-2}$$

MSC2010: 35F20.
Keywords: nonlocal Burgers' equation, finite-time blow-up, global regularity.

The equation that we will be studying is

$$\frac{\partial u}{\partial t}(x,t)+\big(u(x+h,t)\pm u(x-h,t)\big)\frac{\partial u}{\partial x}(x,t)=0, \tag{1-3}$$

with $h \geq 0$. As we can see in the equation, which is a generalized form of the usual one-dimensional Burgers' equation, it includes nonlocal factors. Unlike the local Burgers' equation, analytical solutions are extremely hard to discover for this kind of nonlocal equation. Also, the existence of solutions cannot be easily derived from the method of characteristics. If we look at the characteristics, which are defined by $dx/dt = u(x+h,t) \pm u(x-h,t)$, they are hard to analyze due to the nonlocality.

In Section 2, we prove the following two theorems, illustrating respectively the existence and uniqueness of classical local solutions for periodic initial data $u(x,0) = u_0(x)$. First we introduce the norm which in the following part of the paper will facilitate our proof

Define the Sobolev norm as follows:

Definition 1.1 (Sobolev norm). Let $u(x,t) \in C^\infty(\mathbb{T})$ for some $m \in \mathbb{Z}^+$. Then the Sobolev norm is defined as

$$\begin{aligned}\|u(\cdot,t)\|^2_{H^m([0,L])} &= \int_0^L u(x,t)\big((-\partial_{xx})^m u(x,t)\big)\,\mathrm{d}x\\ &= \int_0^L |\partial_x^m u(x,t)|^2\,\mathrm{d}x.\end{aligned}$$

Remark 1.2. Without loss of generality we can assume that the functions defined on torus have period L. The Sobolev space $H^m([0,L])$ is the closure of $C^\infty([0,L])$ with respect to this norm. Observe that we will work with what is usually called the homogeneous Sobolev space $\dot{H}^m$.

Theorem 1.3 (local existence). *Suppose $u_0 \in C^\infty(\mathbb{T})$. Then there exists a classical local solution $u(x,t)$ to* (2-1) *for $0 \leq t \leq T(u_0)$ for some $T(u_0) > 0$.*

Theorem 1.4 (uniqueness). *The solution $u(x,t)$ to* (2-1) *which is in $C^1([0,T], H^r)$ for large enough r is unique.*

We resort to functional analysis skills in Sobolev spaces. Basically, we use the original equation to generate a recursive sequence of functions and prove that in appropriately chosen Sobolev spaces, the sequence admits a unique limit that converges to a classical local solution, which turns out to be regular by the topological structure of the Sobolev spaces. In Section 3 we look at blow-up and non-blow-up of solutions in finite time, presenting examples of both cases and contrasting with the local Burgers' equation. Interestingly, owing to the nonlocality factors introduced, the blow-up behaviors of (1-3) vary greatly from the local Burgers' equation (1-2). Finally, we use graphics to show simulations run on our equation in Section 4 to illustrate our results.

2. Existence and uniqueness of solution

Let us now consider the following nonlocal variation of Burgers' equation:

$$u_t + \big(u(x+h,t) \pm u(x-h,t)\big)u_x = 0. \tag{2-1}$$

We will prove Theorem 1.3 by justifying Proposition 2.2 and Lemma 2.7 below. To do this, we construct a sequence of functions $u_n(x,t)$ and show that $u_n(x,t)$ will be uniformly bounded in $C([0,T], H^m)$ with large m, while du_n/dt are also uniformly controlled. Thus, by a well-known compactness criterion, there exists a limit which we show solves the equation.

Remark 2.1. Throughout the rest of the paper, we will denote any universal constant by C, which does not depend on $u(x,t)$ and may vary from line to line.

Proposition 2.2. *Define a recursive sequence of functions* $\{u_n\}$ *as*

$$\partial_t u_n + \mathcal{L}u_{n-1}\,\partial_x u_n = 0, \quad u_n(x,0) = u_0(x) \in C^\infty(\mathbb{T}), \tag{2-2}$$

where $u_n = u_n(x,t)$ *for* $n \ge 1$, $\mathcal{L}u_n = u_n(x+h,t) \pm u_n(x-h,t)$ *is a shorthand notation, and* $u_0(x,t) = u_0(x)$ *is smooth. Then for all sufficiently large* $m \in \mathbb{Z}^+$, *there exists* $T(\|u_0\|_{H^m})$ *such that* $\|u_n(\cdot,t)\|_{C([0,T],H^m)} < C_1(T)$ *and* $\|du_n/dt\|_{C([0,T],H^{m-1})} \le C_2(T)$ *for all* $0 < t < T$. *Moreover, there exists a subsequence* n_j *such that* $u_{n_j}(x,t)$ *converges to* $u(x,t)$ *in* $C([0,T], H^r)$ *for any* $r < m$.

Remark 2.3. We should notice that (2-2) has unique solution in $C^\infty(\mathbb{T})$ for every n. To see this we apply an inductive argument to the method of characteristics. Since $u_0 \in C^\infty(\mathbb{T})$, we inductively assume that $u_{n-1} \in C^\infty(\mathbb{T})$. In this case, denote $\mathcal{L}u_{n-1}$ by $f_h(x,t)$. The characteristics system is

$$\begin{cases} \dfrac{dt}{dr}(r,s) = 1, & t(s,0) = 0, \\ \dfrac{dx}{dr}(r,s) = f_h(x,t), & x(s,0) = s, \\ \dfrac{dz}{dr}(r,s) = 0, & z(s,0) = u_0(s). \end{cases}$$

Solving the first we have $t = r$. Thus the second is nothing but $dx/dr = f_h(x,r)$. But $f_h(x,r)$ is smooth, which implies by ODE theory that we have a solution $x = g_h(r,s)$, where g_h is implicit and again smooth. Then the implicit function theorem suggests that we can write $s = k_h(x,r) = k_h(x,t)$. Solving the third, we get $u_n = u_0(s) = u_0(k_h(x,t))$. By the smoothness of both u_0 and k_h, the smoothness of u_n is obtained. The uniqueness of each u_n is guaranteed by the method of characteristics. For more details about the method of characteristics, see [Evans 1998]. Next, since u_0 has period L, an inductive argument will also show that u_n has period L for all n.

Then we move on to prove Proposition 2.2; notice that the above remark will justify the integration by parts in the following proof.

Proof. Let us multiply (2-2) by $\partial_x^{2m} u_n$ and integrate with respect to x from 0 to L:

$$\int_0^L \partial_t u_n \, \partial_x^{2m} u_n \, \mathrm{d}x = -\int_0^L \partial_x^{2m} u_n \, \mathcal{L}u_{n-1} \, \partial_x u_n \, \mathrm{d}x.$$

We can then integrate by parts m times and pull out the partial derivative with respect to time from the left-hand side:

$$\frac{\mathrm{d}}{\mathrm{d}t} \|u_n(\cdot, t)\|_{H^m}^2 = -\int_0^L \partial_x^{2m} u_n \, \mathcal{L}u_{n-1} \, \partial_x u_n \, \mathrm{d}x \le \left| \int_0^L \partial_x^{2m} u_n \, \mathcal{L}u_{n-1} \, \partial_x u_n \, \mathrm{d}x \right|.$$

Integrating by parts m times on the right-hand side and noting that all of the boundary terms vanish due to periodicity, we get

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \|u_n(\cdot, t)\|_{H^m}^2 &\le \left| \int_0^L \partial_x^m (\mathcal{L}u_{n-1} \, \partial_x u_n) \partial_x^m u_n \, \mathrm{d}x \right| \\ &\le \left| \int_0^L \sum_{l=0}^m \binom{m}{l} \partial_x^l (\mathcal{L}u_{n-1}) \, \partial_x^{m-l+1} u_n \, \partial_x^m u_n \, \mathrm{d}x \right| \\ &\le \sum_{l=0}^m \binom{m}{l} \left| \int_0^L \partial_x^l (\mathcal{L}u_{n-1}) \partial_x^{m-l+1} u_n \, \partial_x^m u_n \, \mathrm{d}x \right|. \end{aligned} \tag{2-3}$$

Lemma 2.4. *For all* $0 \le l \le m$ *and* $m > 3/2$,

$$\left| \int_0^L \partial_x^l (\mathcal{L}u_{n-1}) \partial_x^{m-l+1} u_n \, \partial_x^m u_n \, \mathrm{d}x \right| \le C \|u_{n-1}\|_{H^m} \|u_n\|_{H^m}^2.$$

Proof. For the $l = 0$ case, we can reduce this to the $l = 1$ case using integration by parts:

$$\left| \int_0^L \mathcal{L}u_{n-1} \partial_x^{m+1} u_n \, \partial_x^m u_n \, \mathrm{d}x \right| = C \left| \int_0^L \partial_x (\mathcal{L}u_{n-1}) (\partial_x^m u_n)^2 \, \mathrm{d}x \right|.$$

When $l = 1$, it is not hard to see that

$$\begin{aligned}\left|\int_0^L \partial_x(\mathcal{L}u_{n-1})(\partial_x^m u_n)^2\,\mathrm{d}x\right| &\le \|\partial_x(\mathcal{L}u_{n-1})\|_{L^\infty}\cdot\left|\int_0^L (\partial_x^m u_n)^2\,\mathrm{d}x\right|\\ &\le \|\partial_x(\mathcal{L}u_{n-1})\|_{L^\infty}\cdot\int_0^L |\partial_x^m u_n|^2\,\mathrm{d}x\\ &= \|\partial_x(\mathcal{L}u_{n-1})\|_{L^\infty}\cdot\|u_n\|_{H^m}^2.\end{aligned}$$

Applying the Sobolev embedding theorem, we have that for $m > 3/2$,

$$\|\partial_x(\mathcal{L}u_{n-1})\|_{L^\infty} \le C\|\partial_x(\mathcal{L}u_{n-1})\|_{H^{m-1}} \le C\|\mathcal{L}u_{n-1}\|_{H^m},$$

$$\begin{aligned}\|\mathcal{L}u_{n-1}\|_{H^m} &= \|u_{n-1}(x+h,t) \pm u_{n-1}(x-h,t)\|_{H^m}\\ &\le 2\|u_{n-1}\|_{H^m}.\end{aligned}$$

We can conclude

$$\left|\int_0^L \partial_x(\mathcal{L}u_{n-1})(\partial_x^m u_n)^2\,\mathrm{d}x\right| \le C\cdot\|u_{n-1}\|_{H^m}\cdot\|u_n\|_{H^m}^2 \quad \text{for } m > \tfrac{3}{2}.$$

In general, by Hölder's inequality, terms on the right-hand side of (2-3), for $l \ne 1$, are estimated by

$$\begin{aligned}&\left|\int_0^L \partial_x^l(\mathcal{L}u_{n-1})\partial_x^{m-l+1}u_n\,\partial_x^m u_n\,\mathrm{d}x\right|\\ &\qquad \le \|\partial_x^l(\mathcal{L}u_{n-1})\|_{L^{\frac{2(m-1)}{l-1}}}\cdot\|\partial_x^{m-l+1}u_n\|_{L^{\frac{2(m-1)}{m-l}}}\cdot\|\partial_x^m u_n\|_{L^2}.\end{aligned} \tag{2-4}$$

Recall that Gagliardo–Nirenberg inequality (see, e.g., [Doering and Gibbon 1995]) has the form

$$\|\partial_x^s f\|_{L^{2m/s}} \le C\|f\|_{L^\infty}^{1-s/m}\|f\|_{H^m}^{s/m} \quad \text{for all } 1 \le s \le m. \tag{2-5}$$

Now by applying (2-5) and the Sobolev embedding theorem, we can conclude the following two facts:

$$\begin{aligned}\|\partial_x^{m-l+1}u_n\|_{L^{\frac{2(m-1)}{m-l}}} &= \|\partial_x^{m-l}(\partial_x u_n)\|_{L^{\frac{2(m-1)}{m-l}}}\\ &\le C\cdot\|\partial_x u_n\|_{L^\infty}^{1-\frac{m-l}{m-1}}\cdot\|\partial_x^m u_n\|_{L^2}^{\frac{m-l}{m-1}}\\ &\le C\cdot\|\partial_x u_n\|_{H^{m-1}}^{1-\frac{m-l}{m-1}}\cdot\|\partial_x u_n\|_{H^{m-1}}^{\frac{m-l}{m-1}}\\ &= C\cdot\|\partial_x u_n\|_{H^{m-1}}\\ &= C\cdot\|u_n\|_{H^m},\end{aligned} \tag{2-6}$$

$$\begin{aligned}\|\partial_x^l(\mathcal{L}u_{n-1})\|_{L^{\frac{2(m-1)}{l-1}}} &= \|\partial_x^{l-1}(\partial_x(\mathcal{L}u_{n-1}))\|_{L^{\frac{2(m-1)}{l-1}}}\\ &\le C\cdot\|\partial_x(\mathcal{L}u_{n-1})\|_{L^\infty}^{1-\frac{l-1}{m-1}}\cdot\|\partial_x^m(\mathcal{L}u_{n-1})\|_{L^2}^{\frac{l-1}{m-1}}\\ &\le C\cdot\|\partial_x(\mathcal{L}u_{n-1})\|_{H^{m-1}}^{1-\frac{l-1}{m-1}}\cdot\|\partial_x^m(\mathcal{L}u_{n-1})\|_{H^{m-1}}^{\frac{l-1}{m-1}}\\ &= C\cdot\|\partial_x(\mathcal{L}u_{n-1})\|_{H^{m-1}}\\ &= C\cdot\|\mathcal{L}u_{n-1}\|_{H^m}\\ &\le C\cdot\|u_{n-1}\|_{H^m}. \end{aligned}\tag{2-7}$$

Plugging (2-6) and (2-7) into (2-4), we get

$$\left|\int_0^L \partial_x^l(\mathcal{L}u_{n-1})\partial_x^{m-l+1}u_n\,\partial_x^m u_n\,\mathrm{d}x\right| \le C\|u_{n-1}\|_{H^m}\cdot\|u_n\|_{H^m}^2,$$

with constant C which depends only on m. So we have proved the lemma. □

Now let

$$f_0(t) = f_n(0) = \|u_0\|_{H^m}^2.$$

Notice that

$$\|u_n(\cdot,0)\|_{H^m} = \|u_0\|_{H^m}.$$

Now we define $f_n(t)$ inductively by

$$f_n'(t) = C(m)\sqrt{f_{n-1}(t)}f_n(t),\tag{2-8}$$

where $C(m)$ is a constant depending only on m from proof above.

Observe that $f_1(t)\ge f_0(t) > 0$ for all $t\ge 0$ since the right-hand side of (2-8) is always positive. Then inductively, we can obtain that $f_n(t)\ge f_{n-1}(t)$ for all $t\ge 0$.

Also, given

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u_n(\cdot,t)\|_{H^m}^2 \le C(m)\|u_{n-1}(\cdot,t)\|_{H^m}\cdot\|u_n(\cdot,t)\|_{H^m}^2,$$

it follows that

$$f_n(t)\ge \|u_n(\cdot,t)\|_{H^m}^2.$$

Thus

$$f_n'(t) = C(m)\sqrt{f_{n-1}(t)}f_n(t)\le C(m)f_n^{3/2}(t).$$

Because $f_n(t)\ne 0$, we can divide by $f_n^{3/2}(t)$ to get

$$\frac{f_n'(t)}{f_n^{3/2}(t)}\le C(m).$$

We can then integrate from 0 to t, giving

$$\int_0^t \frac{f_n'(s)}{f_n^{3/2}(s)}\,\mathrm{d}s \le \int_0^t C(m)\,\mathrm{d}t,$$

$$-2f_n^{-1/2}(t) + 2(\|u_0\|_{H^m})^{-1/2} \le C(m)t,$$

$$f_n^{1/2}(t) \le \frac{1}{\|u_0\|_{H^m}^{-1/2} - C(m)t/2}.$$

If we let $T := \bigl(C(m)\sqrt{\|u_0\|_{H^m}}\bigr)^{-1}$, we can conclude that for any $0 \le t \le T$, $\{f_n(t)\}$ will be uniformly bounded by some constant $C_1(T)$. But we know that $f_n(t) \ge \|u_n(\cdot\,,t)\|_{H^m}^2$. Therefore

$$\sup_{t\in[0,T]} \|u_n(\cdot\,,t)\|_{H^m(\mathbb{R})} \le C_1(T).$$

Since u_n satisfies (2-2), and H^s in dimension one is an algebra for every $s > 1/2$, this bound also implies

$$\|\partial_t u_n(\cdot\,,t)\|_{H^{m-1}[0,L]} \le C_2(T),$$

if $m > 3/2$. Now standard arguments (see, e.g., [Majda and Bertozzi 2002]) yield existence of a subsequence u_{n_j} converging to a function $u(x,t)$ in $L^\infty([0,T], H^r)$ for any $r < m$. Namely, recall the following compactness criterion.

Proposition 2.5. *Define a Banach space*

$$Y = \bigl\{v \in L^{\alpha_0}([0,T], H^m),\ \partial_t v \in L^{\alpha_1}([0,T], H^s)\bigr\},$$

where $s \le m$, and $1 \le \alpha_{0,1} \le \infty$. Define the norm on the space Y by

$$\|v\|_Y = \|v\|_{L^{\alpha_0}([0,T],H^m)} + \|\partial_t v\|_{L^{\alpha_1}([0,T],H^s)}.$$

Then Y imbeds compactly into any $L^{\alpha_0}([0,T], H^r)$ with $r < s$.

Remark 2.6. This criterion can be found, for example, in [Temam 1977, page 184] (see also [Constantin and Foias 1988]).

It follows that for any $r < m$, we can find u_{n_j} converging to some u strongly in $L^\infty([0,T], H^r)$. This concludes the proof of Proposition 2.2. □

Lemma 2.7. *The function $u(x,t)$ from Proposition 2.2 is a classical solution of* (2-1) *and belongs to $C([0,T], H^r)$ for any $r < m$.*

Remark 2.8. Since so far u has been defined only up to sets of measure zero in time, what we mean is that it can be fixed, if necessary, on a set of times of measure zero so that the claim of the lemma holds.

Proof. Pick m large enough; $m > 7/2$ is sufficient for the argument below to work. Fix any $5/2 < l < m$. We have the recursive formula for u_n in (2-2) and we proved in Proposition 2.2 that a subsequence u_{n_j} (which we will for simplicity denote u_n) converges to u in $L^\infty([0,T], H^l)$. Take some s such that $l-1 > s > 3/2$. We have

$$\begin{aligned}\|\mathcal{L}u_{n-1}\,\partial_x u_n - \mathcal{L}u\,\partial_x u\|_{H^s} &\le \|(\mathcal{L}u_{n-1}-\mathcal{L}u)\partial_x u_n\|_{H^s} + \|\mathcal{L}u(\partial_x u_n - \partial_x u)\|_{H^s}\\ &\le \|\mathcal{L}u_{n-1}-\mathcal{L}u\|_{H^s}\|\partial_x u_n\|_{H^s} + \|\mathcal{L}u\|_{H^s}\|\partial_x u_n - \partial_x u\|_{H^s}\\ &\le C\|u_{n-1}-u_n\|_{H^s}\|\partial_x u_n\|_{H^s} + C\|\mathcal{L}u\|_{H^s}\|u_n - u\|_{H^{s+1}}.\end{aligned}$$

By our choice of l and s, we have

$$\begin{aligned}\|u_{n-1}-u_n\|_{H^s} &\to 0 \text{ uniformly in } t \in [0,T] \text{ as } n \to \infty,\\ \|u_n - u\|_{H^{s+1}} &\to 0 \text{ uniformly in } t \in [0,T] \text{ as } n \to \infty.\end{aligned}$$

Thus

$$\|\mathcal{L}u_{n-1}\,\partial_x u_n - \mathcal{L}u\,\partial_x u\|_{H^s} \to 0 \text{ uniformly in } t \in [0,T] \text{ as } n \to \infty.$$

Now, integrating (2-2) from 0 to t, we have

$$u_n(x,t) = u_n(x,0) - \int_0^t \mathcal{L}u_{n-1}\,\partial_x u_n\,\mathrm{d}s = u_0(x) - \int_0^t \mathcal{L}u_{n-1}\,\partial_x u_n\,\mathrm{d}s. \tag{2-9}$$

Note that by our choice of l and s the H^l- or H^s-convergence implies pointwise convergence, so $u_n \to u$, $\mathcal{L}u_{n-1}\,\partial_x u_n \to \mathcal{L}u\,\partial_x u$ pointwise for almost every t. Then from (2-9), as proved in Proposition 2.2, we conclude that for almost every t,

$$u(x,t) = u_0(x) - \int_0^t \mathcal{L}u\,\partial_x u\,\mathrm{d}s.$$

This means that $u(x,t)$ is Lipschitz in time with values in H^s (up to fixing it on a measure-zero set of times). We also have that $u \in L^\infty([0,T], H^m)$ since the approximating sequence satisfies uniform bound in this space. But then for every $s < r < m$, we have

$$\|u(\cdot,t_2) - u(\cdot,t_1)\|_{H^r} \le \|u(\cdot,t_2)-u(\cdot,t_1)\|_{H^m}^{\frac{r-s}{m-s}}\,\|u(\cdot,t_2)-u(\cdot,t_1)\|_{H^s}^{\frac{m-r}{m-s}},$$

and so we obtain that $u \in C([0,T], H^r)$. □

We have therefore proved that there exists a solution to our equation, (2-1). We now prove uniqueness by considering two different solutions of our equation, $\theta(x,t)$ and $\varphi(x,t)$, and showing that their difference $w(x,t) = \theta(x,t) - \varphi(x,t)$ is zero for all t and x.

Next, we prove that the classical solution is also unique, which is indicated in Theorem 1.4.

Proof. Let θ and φ be solutions to (2-1) with initial data $u(x,0) = u_0(x)$. Then

$$\theta_t + \mathcal{L}\theta\,\theta_x = 0, \tag{2-10}$$

$$\varphi_t + \mathcal{L}\varphi\,\varphi_x = 0. \tag{2-11}$$

Let $w = \theta - \varphi$. Subtracting (2-11) from (2-10), we get

$$\begin{aligned}\partial_t w &= -(\mathcal{L}\theta\,\theta_x - \mathcal{L}\varphi\,\varphi_x)\\ &= -(\mathcal{L}\theta\,\theta_x - \mathcal{L}\varphi\,\varphi_x) + \mathcal{L}\theta\,\varphi_x - \mathcal{L}\theta\,\varphi_x = -\mathcal{L}\theta w_x - \mathcal{L}w\,\varphi_x.\end{aligned}$$

We multiply by $(-1)^r\partial_x^{2r}w$, integrate from 0 to L, and integrate the left-hand side by parts r times, giving

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^L (\partial_x^r w)^2\,\mathrm{d}x = (-1)^{r+1}\int_0^L \partial_x^{2r} w\,\mathcal{L}\theta\,\partial_x w\,\mathrm{d}x + (-1)^{r+1}\int_0^L \partial_x^{2r} w\,\mathcal{L}w\,\partial_x\varphi\,\mathrm{d}x,$$

so

$$\frac{\mathrm{d}}{\mathrm{d}t}\|w\|_{H^r}^2 \le \underbrace{\left|\int_0^L \partial_x^{2r} w\,\mathcal{L}\theta\,\partial_x w\,\mathrm{d}x\right|}_{I_1} + \underbrace{\left|\int_0^L \partial_x^{2r} w\,\mathcal{L}w\,\partial_x\varphi\,\mathrm{d}x\right|}_{I_2}. \tag{2-12}$$

Integrating I_1 by parts r times gives

$$\left|\int_0^L \partial_x^{2r} w\,\mathcal{L}\theta\,\partial_x w\,\mathrm{d}x\right| \le \sum_{l=0}^{r}\binom{m}{l}\left|\int_0^L \partial_x^l(\mathcal{L}\theta)\,\partial_x^{r-l+1} w\,\partial_x^r w\,\mathrm{d}x\right|.$$

Again, when $l = 0$, we can reduce this to the $l = 1$ case using integration by parts. When $l = 1$,

$$\begin{aligned}I_1 = \left|\int_0^L \partial_x^l(\mathcal{L}\theta)\partial_x^{r-l+1} w\,\partial_x^r w\,\mathrm{d}x\right| &= \left|\int_0^L \partial_x(\mathcal{L}\theta)\,\partial_x^r w\,\partial_x^r w\,\mathrm{d}x\right|\\ &= \left|\int_0^L \partial_x(\mathcal{L}\theta)(\partial_x^r w)^2\,\mathrm{d}x\right|\\ &\le \|\partial_x(\mathcal{L}\theta)\|_{L^\infty}\cdot\int_0^L |\partial_x^r w|^2\,\mathrm{d}x\\ &\le C\cdot\|\partial_x(\mathcal{L}\theta)\|_{L^\infty}\cdot\|w\|_{H^r}^2\\ &\le C\cdot\|\theta\|_{H^r}\cdot\|w\|_{H^r}^2\end{aligned}$$

if $r-1>1/2$. When $l \neq 1$,

$$I_1=\left|\int_0^L \partial_x^l(\mathcal{L}\theta)\partial_x^{r-l+1}w\,\partial_x^r w\,\mathrm{d}x\right| \le \|\partial_x^l(\mathcal{L}\theta)\|_{L^{\frac{2(r-1)}{l-1}}}\cdot\|\partial_x^{r-l+1}w\|_{L^{\frac{2(r-1)}{r-l}}}\cdot\|\partial_x^r w\|_{L^2}$$
$$\le C\|\theta\|_{H^r}\cdot\|w\|_{H^r}^2$$

as before. We can therefore conclude that

$$I_1=\left|\int_0^L \partial_x^{2r}w\,\mathcal{L}\theta\,\partial_x w\,\mathrm{d}x\right| \le C\|\theta\|_{H^r}\cdot\|w\|_{H^r}^2.$$

The same process can be done to I_2 to determine a bound for the integral, giving the result

$$I_2=\left|\int_0^L \partial_x^{2r}w\,\mathcal{L}w\,\partial_x\varphi\,\mathrm{d}x\right| \le C\|\varphi\|_{H^r}\cdot\|w\|_{H^r}^2.$$

Thus, (2-12) becomes

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|w\|_{H^r}^2 &\le C\|\theta\|_{H^r}\cdot\|w\|_{H^r}^2 + C\|\varphi\|_{H^r}\cdot\|w\|_{H^r}^2\\ &= \|w\|_{H^r}^2(C\|\theta\|_{H^r}+C\|\varphi\|_{H^r}).\end{aligned}$$

Then by Grönwall's inequality, we have

$$\|w(\cdot,t)\|_{H^r}\le\|w(\cdot,0)\|_{H^r}\exp\biggl(\int_0^t (C\|\theta(\cdot,s)\|_{H^r}+C\|\varphi(\cdot,s)\|_{H^r})\,\mathrm{d}s\biggr),$$

but $\|w(\cdot,0)\|_{H^r}=0$ because θ and φ are solutions to the same Cauchy problem. Therefore, the difference $w=\theta-\varphi$ is zero a.e. Since θ and φ are sufficiently smooth, they must be equal everywhere. □

3. Blow-up and non-blow-up properties

Let us consider the following two subcases of equation (2-1), where they both have initial data $u_0(x)$ of period L:

$$u_t+\bigl(u(x+h,t)+u(x-h,t)\bigr)u_x=0, \tag{3-1}$$
$$u_t+\bigl(u(x+h,t)-u(x-h,t)\bigr)u_x=0. \tag{3-2}$$

Remark 3.1. Let us introduce the following notation: denote $u^h(x,t)$ to be the solution of an equation with spatial shift h. Looking at (3-2), it can be shown using symmetry and uniqueness that if the smooth initial condition $u_0(x)$ is even, the solution, while it remains smooth, will stay even in x. Also, $u^h(x,t)=u^{L-h}(x,t)$

for all periodic initial data. Now consider (3-1). If $u_0(x)$ is odd, the solution will stay odd in x. Also, $u^h(x,t) = u^{L-h}(x,t)$ will hold for all even initial data $u_0(x)$.

These facts are deduced from the existence and uniqueness of solutions, definitions of evenness and oddness, and periodicity applied to our equation.

We now state the existence of solutions that blow up in finite time.

Theorem 3.2 (Existence of blow-up). *There exists initial data $u_0 \in C^\infty(\mathbb{R})$ such that the solution $u(x,t)$ to* (2-1) *blows up in finite time.*

We prove this result in Section 3. We first derive some properties of the solution.

Lemma 3.3. *Suppose $u(x,t)$ is a periodic solution of* (2-1) *with period $L = 2h$. Let $u(0,0) = u(h,0) = 0$; then $u(0,t) = u(h,t) = 0$, for all $t > 0$.*

We can prove this by considering both the plus and minus cases as follows:

Proof. Let us first consider the plus sign case, (3-1). Plugging $x = 0, h$ into to the recursive formula (2-2) for the plus case, we get

$$\partial_t u_n(0,t) = -2u_{n-1}(h,t)\partial_x u_n(0,t),$$
$$\partial_t u_n(h,t) = -2u_{n-1}(0,t)\partial_x u_n(h,t).$$

Since $u(0,0) = u_0(0) = u(h,0) = u_0(h) = 0$, we easily see that $\partial_t u_1(0,t) = \partial_t u_1(h,t) = 0$; therefore u_1 is constant at $x = 0, h$. But $u_1(0,0) = u_0(0) = 0$ and $u_1(h,0) = u_0(h) = 0$, so we have $u_1(0,t) = u_1(h,t) = 0$. Then, inductively, assume $u_{n-1}(0,t) = u_{n-1}(h,t) = 0$. Then, $\partial_t u_n(0,t) = \partial_t u_n(h,t) = 0$ so they are both constant. By the same reasoning, $u_n(0,0) = u_n(h,0) = 0$; therefore they are identically zero for all time. But our solution is just the limit of a subsequence of u_n, so $u(0,t) = u(h,t) = 0$

Now let us consider the minus sign case, (3-2). Plugging $x = 0$ into (3-2), we get

$$u_t(0,t) = (u(h,t) - u(h,t))u_x(0,t) = 0,$$

because $u(-h,t) = u(h,t)$ due to the period $L = 2h$. So $u(0,t) = C$, independent of time. Therefore, if we choose $u(0,0) = 0$, then $u(0,t) = 0$ for all $t > 0$. The same may be done at $u(h,0)$ to show that if $u(h,0) = 0$, then $u(h,t) = 0$. □

Corollary 3.4. *Suppose $u_0(x) \in C^\infty(\mathbb{R})$ has period $L = kh$ for some $k \in \mathbb{Z}$ and $u_0(mh) = 0$ for all $0 \le m \le k$. Then the solution to* (2-1) *satisfies $u(mh,t) = 0$ for all $t \ge 0$ and $0 \le m \le k$.*

The proof is similar to that from Lemma 3.3 extended for more general integers.

Blow-up. Now we investigate the cases where $u_0(x)$ has period $L = 2h$ and $u_0(0) = u_0(h) = 0$, and derive the possibility of blow-up.

Lemma 3.5. *Consider the equation* $u_t + \big(u(x+h,t) + u(x-h,t)\big)u_x = 0$ *with* $u(x,0) = u_0(x) \in C^\infty(\mathbb{R})$, *period* $L = 2h$, *and* $u_0(0) = u_0(h) = 0$. *Assume* $u_x(0,0) < 0$ *and* $u_x(h,0) < 0$. *Then the solution* $u(x,t)$ *blows up in finite time.*

Proof. Note that in Proposition 2.2, we proved that if the initial data $u_0(x)$ has period $2h$, then $u(x,t)$ will also have period $L = 2h$. Also, in this case, by (3-1), $u(0,t) = u(h,t) = 0$.

Differentiating the equation with respect to x gives

$$u_{tx}(x,t) + \big(u_x(x+h,t) + u_x(x-h,t)\big)u_x(x,t) \\ + \big(u(x+h,t) + u(x-h,t)\big)u_{xx}(x,t) = 0. \quad (3\text{-}3)$$

Plug in $x = 0, h$ respectively and define $F_1(t) = u_x(0,t)$ and $F_2(t) = u_x(h,t)$. Noting that the last terms in both cases vanish, we get

$$F_1' + 2F_1F_2 = 0, \quad (3\text{-}4)$$

$$F_2' + 2F_1F_2 = 0. \quad (3\text{-}5)$$

It is easy to see that $F_1' - F_2' = 0$; thus $F_1 - F_2 = A$, where A is a constant. Since we assume $F_1 = F_2$, we get that $A = 0$. Plugging this into (3-4) gives

$$F_1' + 2F_1^2 = 0.$$

The solution to this differential equation is

$$F_1(t) = \frac{1}{\frac{1}{F_1(0)} + 2t}.$$

This blows up in finite time when

$$t = -\frac{1}{2F_1(0)} = -\frac{1}{2u_x(0,0)} > 0.$$

We can argue similarly for (3-5) to show that F_2 also blows up in finite time under the same conditions. □

Remark 3.6. For instance, we can take

$$u(x,0) = u_0(x) = x(x-h)(x-2h)\left(-\frac{1}{2h^2} + \frac{3}{h^3}x - \frac{3}{2h^4}x^2\right)$$

for $0 \le x \le 2h$. This satisfies our assumptions in Lemma 3.5 and thus the corresponding solution blows up in finite time.

Remark 3.7. There is an obvious case of blow-up for the plus sign equation when the period L is just h. Equation (3-1) reduces to

$$u_t + 2u \cdot u_x = 0.$$

This is the typical Burgers' equation, which is known to blow up in finite time for any nonconstant periodic initial condition $u_0(x)$ [McOwen 2003].

Lemma 3.8. *Suppose u_0 has period $L{=}6h$ and is even, and $u_0(kh){=}0$, $u_0'(3kh){=}0$ for all $k \in \mathbb{Z}$. Assume $u_x(2h,0) < 0$, $u_x(h,0) > 0$ and*

$$\frac{\ln u_x(h,0) - \ln(-u_x(2h,0))}{u_x(h,0) + u_x(2h,0)} > 0.$$

Then the solution $u(x,t)$ to the Cauchy problem,

$$\begin{gathered} u_t + \big(u(x+h,t) - u(x-h,t)\big)u_x = 0, \\ u(x,0) = u_0(x), \end{gathered}$$

blows up in finite time.

Proof. By Lemma 3.3 and Corollary 3.4, we have $u(kh,t) = 0$, for all $k \in \mathbb{Z}$ and $u(x,t)$ is even if $u_0(x)$ is even. Differentiating the equation with respect to x gives

$$\begin{multline} u_{tx}(x,t) + \big(u_x(x+h,t) - u_x(x-h,t)\big)u_x(x,t) \\ + \big(u(x+h,t) - u(x-h,t)\big)u_{xx}(x,t) = 0. \end{multline}$$

Observe that $u_x(3kh,t) = 0$ for all time by an argument similar to proof of Lemma 3.3. Plugging in $x = h, 2h$ gives

$$\begin{aligned} F_1'(t) + F_1(t)F_2(t) &= 0, \\ F_2'(t) - F_1(t)F_2(t) &= 0, \end{aligned}$$

where $F_1(t) = u_x(h,t)$ and $F_2(t) = u_x(2h,t)$. Solving this system of ordinary differential equations gives

$$\begin{gathered} F_1(t) + F_2(t) = F_1(0) + F_2(0) = A, \\ F_1'(t) = F_1^2(t) - AF_1(t) \end{gathered}$$

for some constant A. Thus

$$F_1(t) = \frac{A\exp(AB)}{\exp(AB) - \exp(At)},$$

where

$$B = \frac{\ln F_1(0) - \ln(-F_2(0))}{F_1(0) + F_2(0)}.$$

This blows up in finite time if $F_2(0) = u_x(2h,0) < 0$, $F_1(0) = u_x(h,0) > 0$ and $B > 0$. □

Remark 3.9. To give an example, take $h = 4/3$. Then we can take

$$u(x,0) = u_0(x) = \frac{16(x-4)^2(x+4)^2x^2(3x-8)(3x+8)(3x+4)(3x-4)}{3375(112+153x^2)}.$$

This satisfies our assumptions in Lemma 3.8 and thus blows up in finite time. This may not be a very nicely manufactured example, but our point is that functions specified by Lemma 3.8 do exist.

Non-blow-up. We will now look for stationary solutions by taking specific initial data to (3-2) and showing that it cannot blow up in finite time. Let $u(x,t) = \sin(\pi x k/h)$, where h is fixed and $k \in \mathbb{Z}$. Noting that $u_t = 0$ and $u(x+h,t) - u(x-h,t) = 0$ (by trigonometric identities), we have that $u(x,t)$ solves the equation and never blows up.

Similarly, for (3-1), we will take $u(x,t) = \sin\bigl(\pi x\bigl(k-\frac{1}{2}\bigr)/h\bigr)$, where h is fixed and $k \in \mathbb{Z}$. Once again, noting that $u_t = 0$ and $u(x+h,t)+u(x-h,t) = 0$, we have that $u(x,t)$ solves the equation. We also know that $u(x,t) = \sin\bigl(\pi x\bigl(k-\frac{1}{2}\bigr)/h\bigr)$ never blows up. So we have found stationary solutions for both equations (3-1) and (3-2) that never blow up in finite time. So the nonlocal models are different from Burgers' equation where any nonconstant solution blows up in finite time: there exists nontrivial initial data for which solutions are globally regular for the nonlocal equation.

We can also construct a stationary solution to (3-2) by setting the period L to be h. The nonlocal terms become $u(x+h,t) = u(x-h,t) = u(x,t)$, so (3-2) reduces to $u_t = 0$. This is constant in time. Therefore $u(x,t) = u_0(x)$ for all t, so given a smooth initial condition, $u(x,t)$ will not blow up.

4. Simulations

In this section, we compare our model with the well-known "local" Burgers' equation (1-2). We used Matlab v2013 to run all simulations, with a forward-in-time, centered-in-space scheme. We illustrate many of the results of this paper in the graphics we generate.

We first look at the "local" Burgers' equation, (1-2). We know that this leads to gradient catastrophe (i.e., blow-up in gradient) in finite time for all nonconstant smooth initial data. We use $u(x,0) = \sin(\pi x)$ to generate Figure 1 (left).

As we can see, the slope of the graph in Figure 1 (left) at $x = 0$ blows up in finite time. Now, considering our equation with the plus sign,

$$u_t + \bigl(u(x+h,t)+u(x-h,t)\bigr)u_x = 0,$$

notice that there is a translation parameter h in our equation which affects the location of blow-up. As we can see in Figure 1 (right) with $h = L/8$, where L is the period of the initial data, blow-up does not occur at the origin, and two peaks

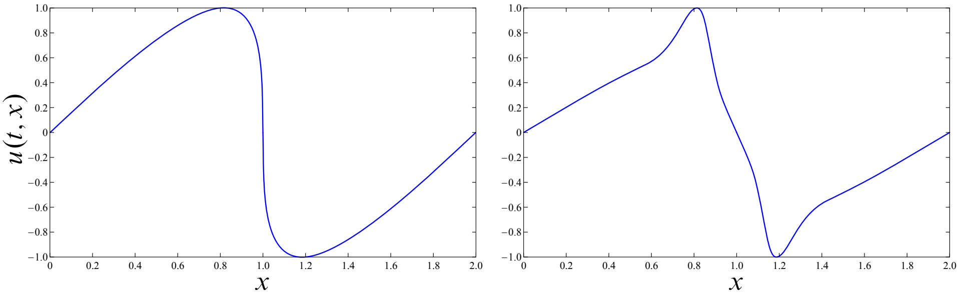


Figure 1. Local Burgers' equation with $h = 0$ (left) and nonlocal Burgers' equation with $h = L/8$ (right).

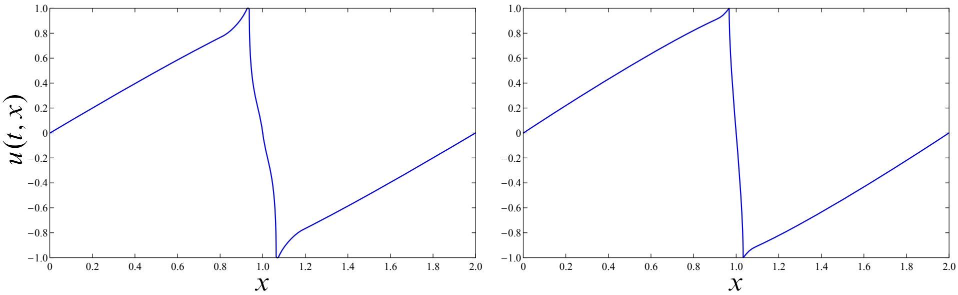


Figure 2. Nonlocal Burgers' equations with $h = L/16$ (left) and $h = L/32$ (right).

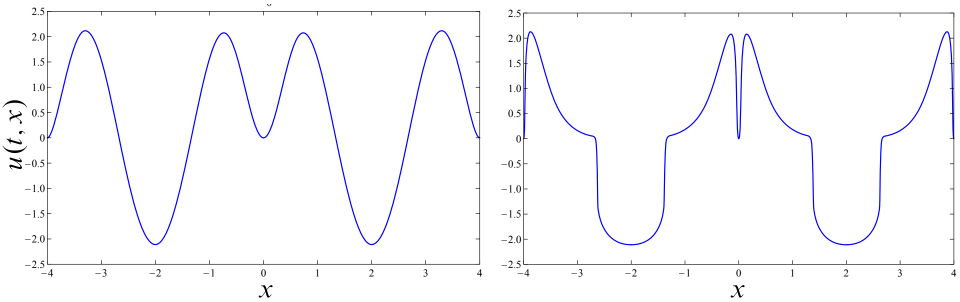


Figure 3. Nonlocal Burgers' equation minus case initial condition (left) and in finite time (right).

form instead of the usual one. We then varied the value of h to be $L/16$ and $L/32$ in Figure 2, which gives blow-up closer and closer to the origin.

Now we constructed initial data to fit Lemma 3.8 to get intuition on how it will blow up at $x = \pm L/3, \pm 2L/3$ in the minus sign case. Figure 3 (left) shows the initial data for our equation

$$u_t + \big(u(x+h,t) - u(x-h,t)\big)u_x = 0.$$

Note how $u(x,0) = 0$ at $x = kh$, where period $L = 6h$. Now in Figure 3 (right), we see that at $x = \pm L/3$, $\pm 2L/3$, vertical lines form, causing blow-up in slope.

Acknowledgements

Our research was done during the 2013 University of Wisconsin–Madison REU, sponsored by NSF grants DMS-1056327 and DMS-1147523. We would like to thank Professor Alexander Kiselev for his introduction of the topic and guidance throughout. We would also like to thank Kyudong Choi and Tam Do for their help with the many issues we ran into.

References

[Constantin and Foias 1988] P. Constantin and C. Foias, *Navier–Stokes equations*, Univ. of Chicago Press, 1988. MR 90b:35190 Zbl 0687.35071

[Doering and Gibbon 1995] C. R. Doering and J. D. Gibbon, *Applied analysis of the Navier–Stokes equations*, Cambridge Univ. Press, 1995. MR 96a:76024 Zbl 0838.76016

[Evans 1998] L. C. Evans, *Partial differential equations*, Graduate Studies in Mathematics **19**, Amer. Math. Soc., Providence, RI, 1998. MR 99e:35001 Zbl 0902.35002

[Majda and Bertozzi 2002] A. J. Majda and A. L. Bertozzi, *Vorticity and incompressible flow*, Cambridge Texts in Applied Mathematics **27**, Cambridge Univ. Press, 2002. MR 2003a:76002 Zbl 0983.76001

[McOwen 2003] R. C. McOwen, *Partial differential equations: Methods and applications*, 2nd ed., Prentice Hall, Upper Saddle River, NJ, 2003. Zbl 0849.35001

[Temam 1977] R. Temam, *Navier–Stokes equations: Theory and numerical analysis*, Studies in Mathematics and its Applications **2**, North-Holland, Amsterdam, 1977. MR 58 #29439 Zbl 0383.35057

Received: 2013-09-16 Revised: 2014-06-06 Accepted: 2014-06-08

sgoodchild11692@gmail.com *University of Wisconsin–Madison, Madison, WI 53706,*
United States

hy18@rice.edu *Department of Mathematics, Rice University,*
Houston, TX 77005, United States

INVOLVE 9:1 (2016)
dx.doi.org/10.2140/involve.2016.9.83

Investigating cholera using an SIR model with age-class structure and optimal control

K. Renee Fister, Holly Gaff, Elsa Schaefer,
Glenna Buford and Bryce C. Norris

(Communicated by Suzanne Lenhart)

The use of systems of differential equations in mathematical modeling in conjunction with epidemiology continues to be an area of focused research. This paper briefly acquaints readers with epidemiology, cholera, and the need for effective control strategies; discusses cholera dynamics through a variation on the SIR epidemiological model in which two separate age classes exist in a population; finds the numeric value for R_0 to be approximately 1.54 using estimated parameters for Bangladesh; and employs an optimal control resulting in a suggestion that a protection control be implemented at the end of the monsoon season.

1. Definition of epidemiology and cholera

The World Health Organization defines epidemiology as: “the study of the distribution and determinants of health-related states or events (including disease), and the application of this study to the control of diseases and other health problems” [WHO 2012c]. Notice, this definition addresses two foundational questions regarding any disease. The first question is simply, “How does this disease work?” Once that question is adequately answered, epidemiologists ask the natural follow-up question, “How can we control this disease?”

Diarrheal disease is the fifth most deadly disease category in the world claiming more lives annually than HIV and the deadliest of cancers [WHO 2012d]. Cholera is a fast-acting diarrheal disease capable of causing death within hours of the onset of symptoms [WHO 2012b]. Again, we reference the expertise of the World Health Organization to provide an excellent definition for cholera:

MSC2010: 35L45, 35L50, 92D30.

Keywords: endemic cholera, SIR model, age class structure, ordinary differential equations, optimal control.

We would like to acknowledge that this research has been made possible through a collaborator grant from the National Science Foundation, DMS-0813563.

> Cholera is an acute intestinal infection caused by ingestion of food or water contaminated with the bacterium *Vibrio cholerae*. It has a short incubation period, from less than one day to five days, and produces an enterotoxin that causes a copious, painless, watery diarrhoea that can quickly lead to severe dehydration and death if treatment is not promptly given. Vomiting also occurs in most patients. [WHO 2012a]

Having been in existence since the time of Christ and still having no cure, cholera is an excellent candidate for one of the longest-standing diseases in human history [Barua and Greenough 1992]. Cholera and similar diseases have been virtually eliminated in modernized nations through rigorous sanitation and waste treatment infrastructure [WHO 2012b]. Treatment of cholera continues to be challenging for those nations unable to effectively implement these more cost-prohibitive strategies.

2. Introduction to the two-compartment model

Among the most common epidemiological models is the standard SIR model, so named for the classifications within the model. S represents the susceptible class defined to be all of those people in a population that are not transmitting and have no immunity to the disease. I represents the infected class and is comprised of those who are transmitting the disease. Infecteds may or may not exhibit symptoms. R represents the recovered class of people who have some immunity to the disease. See Figure 1 for a depiction of this model.

One connection between the biological study of diseases and mathematics comes through differential equations. One may recall the foundational idea of calculus to be the derivative. Knowing about the derivative and how it relates to the original equation allows us to determine a representation of the original equation. Thus, if we can design a set of equations around the interaction of classes and how they are changing at any given time, we may use methods of solving differential equations to determine the number of individuals in each class at any given time.

In an attempt to analyze the effects of age on cholera dynamics, we begin with the two-compartment SIR model. This is a variation of the standard SIR model

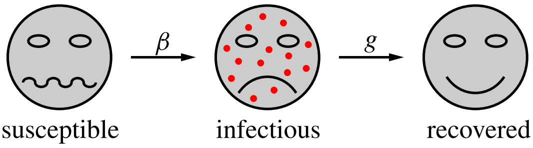


Figure 1. A pictorial representation of the standard SIR model. One may observe the nervousness of the susceptible class in anticipation of infection.

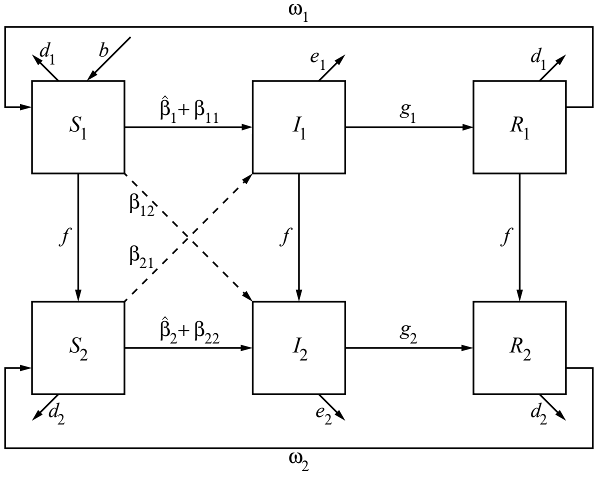


Figure 2. A pictorial representation of our two-compartment model. Dashed lines indicate an interaction between compartments, whereas solid lines indicate movement to and from compartments with $N = S_1 + S_2 + I_1 + I_2 + R_1 + R_2$.

common to biological epidemiology in which we assign classes based on age. Children under five are assigned to classes with a subscript of one. Population members at or above five years of age — henceforth referred to as matured — are assigned to the remaining classes with a subscript of two (the significance of age five is covered later in the paper). We assume that both age classes move through distinct SIR models independently except in regards to social interaction and aging. See Figure 2 for a pictorial representation of this model.

Susceptibles (denoted S_i) are defined as members of the population capable of contracting the disease (becoming infected) which implies that members of this class have no immunity to the disease. It is worth noting the underlying assumption that prior to the introduction of the disease, the entire host population is susceptible.

We define infected population members (denoted I_j) to be those capable of transmitting the disease — not necessarily those who show symptoms. Logically, these classes are assumed to be empty before the introduction of the disease.

Finally, the recovered class (denoted R_k) represents the population that has obtained, by some means, immunity to the disease. The movement of population members between these classes is discussed in the following section.

3. Two-compartment model differential equations

Our goal in studying this model is to discuss the dynamics and control of age-structured models of cholera. To this end we will translate our conceptual model from picture to ordinary differential equations as follows. A discussion of the meaning of the terms occurs in Section 3.1.

$$\frac{dS_1}{dt} = bN - \frac{\beta_{11}S_1I_1}{N} - \frac{\beta_{12}S_1I_2}{N} - \hat{\beta}_1S_1 - fS_1 - d_1S_1 + \omega_1R_1, \tag{3-1}$$

$$\frac{dS_2}{dt} = fS_1 - \frac{\beta_{21}S_2I_1}{N} - \frac{\beta_{22}S_2I_2}{N} - \hat{\beta}_2S_2 - d_2S_2 + \omega_2R_2, \tag{3-2}$$

$$\frac{dI_1}{dt} = \frac{\beta_{11}S_1I_1}{N} + \frac{\beta_{12}S_1I_2}{N} + \hat{\beta}_1S_1 - fI_1 - g_1I_1 - e_1I_1, \tag{3-3}$$

$$\frac{dI_2}{dt} = \frac{\beta_{21}S_2I_1}{N} + \frac{\beta_{22}S_2I_2}{N} + \hat{\beta}_2S_2 + fI_1 - g_2I_2 - e_2I_2, \tag{3-4}$$

$$\frac{dR_1}{dt} = g_1I_1 - fR_1 - \omega_1R_1 - d_1R_1, \tag{3-5}$$

$$\frac{dR_2}{dt} = g_2I_2 + fR_1 - \omega_2R_2 - d_2R_2, \tag{3-6}$$

subject to initial conditions $S_1(0) = S_{10}$, $S_2(0) = S_{20}$, $I_1(0) = I_{10}$, $I_2(0) = I_{20}$, $R_1(0) = R_{10}$, $R_2(0) = R_{20}$.

3.1. ***Equation descriptions.*** The equations are generally easy to generate by interpreting the pictorial model as a flow chart. As intuition would suggest, positive terms indicate entrance into a class, and negative terms indicate removal. It is worth noting that every term that enters a class must, at some point, be removed from another class (with the exception of the population growth term b). It would seem logical to assume the converse, but the introduction of death rates limits this assumption as each class has a death-rate term that does not enter any other class.

The $\beta_{ij}S_iI_j$ terms — the subscript of β is derived from the subscripts of the S and I classes respectively between whom the interaction is taking place — have a denominator of N, which is defined to be the total number of people in the system. These "mass action" (or frequency dependent) transmission terms come from the research of Keeling and Rohani [2008] and have been introduced to model the heterogeneous interaction tendencies of human populations. Every other class transfer rate is dependent solely upon the class from which it originates. Thus, every other term in the equations consists of a rate multiplied by its associated class. With this in mind, let us now consider our individual equations.

From the model, we expect that the only inputs (positive terms) for the S_1 class will come from population growth rate (represented by b) multiplied by the population size (recall that this term is defined as N) and, less intuitively, the loss of a recovered child's immunity to cholera over the course of time shown by rate ω_1.

Turning our attention to (3-1), we expect that some proportion (β_{1j}) of interactions of susceptible children with infecteds — two distinct terms representing interactions with both matured infecteds I_2 and children I_1 — will produce a new infected child. Data for Bangladesh, where cholera is endemic, suggests that vibrios may be reintroduced to a system periodically by environmental factors [Ryan and Charles 2011]. To accommodate this information, an environmental forcing term ($\hat{\beta}_1$) is introduced. We also must account for a natural death rate (d_1) of children in our model population being removed from our S_1 class by inserting a term $d_1 S_1$. Lastly, we assume that children become matured at rate f, and thus move from S_1 to S_2.

Let us discuss (3-2) by collecting the positive terms first. We assume that the S_2 class has only two inputs: one from the advancement of children to matured individuals (rate f) and the second from matured recovereds losing immunity to the disease (rate ω_2). Looking at outputs, we should expect to see similar interaction terms as appeared in the dS_1/dt equation. One infection term accounts for new infections resulting from the interaction of matured individuals with children, a second term accounts for those resulting from interactions among matured individuals, and a final term models infection of matured population due to environmental factors. The introduction of a natural death rate for adults in our population (rate d_2) provides the mortality term for our equation.

When considering the infected classes, we expect to see positive terms representing interactions between susceptibles and infecteds as well as terms representing environmental factors that result in new infections. Because cholera has no direct effect on the process of aging, we still expect children to become adults in the infected classes at the same rate (f) as in the susceptible classes. A certain proportion of those infected with cholera will recover and retain some immunity to the disease (at rates g_1 and g_2) moving them out of the infected class into the recovered class. We introduce a term for the death rate of infected individuals (rates e_1 and e_2) that includes the risk of cholera-related death.

In this model, the only way in which children may enter the recovered class R_1 is to survive the disease — which we have defined to happen at rate g_1. This is a limiting assumption that ignores the possibility of inoculation. Field research seems to indicate that surviving cholera infection does confer some degree of immunity; however, this immunity does not persist for the survivor's lifetime. For our purposes, the loss of immunity is modeled by a waning immunity rate (ω_1) that moves the population out of the recovered class back into the susceptible class. In addition, King et al. [2008] have found a variety of estimates of the time for which immunity persists.

The death rate for the recovered population is assumed to be the natural rate (d_1), and children continue to move into the matured class at the rate f. This means that the matured recovered population may be increased by either mature population members surviving the disease (rate g_2) or children surviving the disease and maturing during the period in which they have some degree of immunity (represented by the familiar age-class-advancement term f). We assume that recovered individuals die at the natural rate (d_2) and that they lose immunity and return to the susceptible population at a rate of ω_2.

3.2. ***Parameter estimates.*** The following subsections contain a discussion of our parameter estimates for the selected population in Bangladesh. During our research, we became particularly interested in endemic cholera, and found data from Bangladesh readily available. However, it is worth noting that the appropriate selection of parameters would allow this system to model cholera for two age classes in any population.

After exploration of the Wolfram Alpha information database [2010], we concluded that a natural death rate for Bangladesh should be 0.0092 people per person per year, and the birth rate should be 0.0247 people per person per year. Because our model analyzes the system in days, it is expedient for us to ensure that the units on all of our rates are also in days. The result of these and the remaining parameter calculations may be seen in Table 1.

The dissertation work of Peng Zhong [2011] seems to indicate a relatively low death rate as compared to the rate of deaths caused by cholera — .0014 people per person per year. This estimation would include a number of control tactics actually being used in the field. As our intention is to develop a system that accurately models real-world data in Bangladesh, the use of this parameter estimate is warranted. However, we must again convert into the proper unit of measure and account for naturally occurring deaths.

The research of Harris et al. [2008] among others, [Ryan 2011; Ryan and Charles 2011], posits statistical significance of increased risk of infection for children under five. Thus, we have chosen five years as our significant age for class advancement. Calculating a rate of advancement of the first age class to the second seems to be a simple matter of arithmetic, dividing the birth rate by the number of days in five years to find an appropriate rate.

Again, Harris et al. [2008] gave us a significant clue as to what parameter values to use in modeling the interaction between infected and susceptible classes. Their conclusion was that twenty-one percent of household contacts "...develop definite *V. cholerae* infections..." and "...children 5 years of age or younger ...were 2.7 times..." as likely to develop infections as were "older individuals". Thus, a bit of algebra gives us estimates for our interaction parameters.

The forcing term used is a randomized Heaviside function, which we sought to use to model the monsoon season in Bangladesh and east India where cholera outbreaks are common. We used Bangladesh's monsoon data [Sack et al. 2003], and concluded that the monsoon season lasts from the beginning of June through the end of September. With this information, we "turned on" our Heaviside function on the 152nd day of the year (corresponding with June 1) and "turned off" our Heaviside function on the 254th day of the year (corresponding with September 30). During monsoon season, the Heaviside function assigns $\hat{\beta}_i$, a random proportion, for each day. This allows us to account for the randomness of nature in that the

symbol	description	rate
b	birth rate of the entire population [1]	6.8×10^{-5} day^{-1}
d_1	natural death rate of children [1]	2.5×10^{-5} day^{-1}
d_2	natural death rate of matured individuals [1]	2.5×10^{-5} day^{-1}
f	proportion of children maturing [2]	7.8×10^{-7} day^{-1}
β_{11}	child infections resulting from interactions between susceptible children and infected children [2]	1.5×10^{-1} day^{-1}
β_{12}	child infections resulting from interactions between susceptible children and matured infecteds [2]	1.5×10^{-1} day^{-1}
β_{21}	matured infections resulting from interactions between matured susceptibles and infected children [2]	5.7×10^{-2} day^{-1}
β_{22}	matured infections resulting from interactions between matured susceptibles and matured infecteds [2]	5.7×10^{-2} day^{-1}
$\hat{\beta}_1$	environmental forcing term for children	Heaviside
$\hat{\beta}_2$	environmental forcing term for mature individuals	Heaviside
e_1	death rate of child-aged infected population [1]	2.9×10^{-5} day^{-1}
e_2	death rate of matured infected population [1]	2.9×10^{-5} day^{-1}
g_1	infected children to recovered children transition [3,4]	6.7×10^{-2} day^{-1}
g_2	infected adults to recovered adults transition [3,4]	6.7×10^{-2} day^{-1}
ω_1	waning immunity rate of children [5]	2.2×10^{-3} day^{-1}
ω_2	waning immunity rate of matured individuals [5]	2.2×10^{-3} day^{-1}

Table 1. Two-compartment parameter estimates. Key: [1] = [Wolfram Alpha 2010]; [2] = [Harris et al. 2008]; [3] = [Nelson et al. 2009]; [4] = [WHO 2012b]; [5] = [Zhong 2011].

effects of rain and floods could vary significantly day by day. It is important to note that, although the forcing terms obtain new values for each time step, they are considered constant for each time step.

Here we define

$$\hat{\beta}_i = \begin{cases} \texttt{rand} & \text{if } 152 \le t \bmod 365 \le 254, \\ 0 & \text{elsewhere,} \end{cases}$$

where $i = 1, 2$ and `rand` is the MATLAB function reference that generates random numbers whose elements are uniformly distributed in the interval (0, 1).

As noted previously, the rate at which recovered members of the population lose immunity is difficult to accurately estimate. We elected to use estimates from [Zhong 2011] in our model; however, we wish to note that in [Nelson et al. 2009], they observe that these rates assume almost entirely asymptomatic infections within the population which is not supported by recent studies.

We determined an appropriate estimation of the length of a cholera infection to be fifteen days. As such, we are easily able to calculate the rate of recovery of the infected classes.

Table 1 contains estimates for parameters used in the two-compartment model, short descriptions of the parameters, and the source references for our estimates.

4. R_0 calculation

R_0, or the basic reproductive ratio, is commonly defined to be the mean number of secondary cases resulting from a single primary infection within a population. This is a useful measure of the epidemicity, or speed of spread, of a disease. In calculating R_0, we used the methods outlined by van den Driessche and Watmough [2002]. In order to carefully calculate R_0, we recognize that if there is no disease present initially, then the population remains disease-free for all time. The $\hat{\beta}_i$ coefficients for $i = 1, 2$ cause a challenge. Therefore, we analyze the system with the $\hat{\beta}_i$ values for $i = 1, 2$ equal to zero. If the disease is spread in this case, it would also spread with nonzero $\hat{\beta}_i$ coefficients. Using this assumption within the system (3-1)–(3-6), we use the next generation matrix technique as follows.

$\mathcal{F}$ is formed by compiling the terms that bring new infections into an infected class. For the sake of convenience and clarity, we rearranged the differential equations in such a way that the infected classes are at the top (i.e., I_1, I_2, S_1, S_2, R_1, R_2). This operation is permitted by common linear algebra operations; however, its application must be consistent through the calculation. The vector $\mathcal{V}$ is constructed by compiling the additive inverse of the remaining terms such that our state equations can be determined by taking $\mathcal{F} - \mathcal{V}$.

It is important to note that only new infections are considered in $\mathcal{F}$. Movement between infected classes is shown in $\mathcal{V}$:

$$\mathcal{F} = \begin{bmatrix} \beta_{11} S_1 I_1 N^{-1} + \beta_{12} S_1 I_2 N^{-1} \\ \beta_{21} S_2 I_1 N^{-1} + \beta_{22} S_2 I_2 N^{-1} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{4-1}$$

$$\mathcal{V} = \begin{bmatrix} f I_1 + g_1 I_1 + e_1 I_1 \\ -f I_1 + g_2 I_2 + e_2 I_2 \\ -bN + \beta_{11} S_1 I_1 N^{-1} + \beta_{12} S_1 I_2 N^{-1} + \bar{\beta}_1 S_1 + f S_1 + d_1 S_1 - \omega_1 R_1 \\ -f S_1 + \beta_{21} S_2 I_1 N^{-1} + \beta_{22} S_2 I_2 N^{-1} + \bar{\beta}_2 S_2 + d_2 S_2 - \omega_2 R_2 \\ -g_1 + f R_1 + \omega_1 R_1 + d_1 R_1 \\ -g_2 I_2 - f R_1 + \omega_2 R_2 + d_2 R_2 \end{bmatrix}. \tag{4-2}$$

We then calculate the disease-free equilibrium (henceforth DFE) — mathematically denoted $\vec{x}_0$ — by setting the state equations equal to zero and $I_1 = I_2 = 0 = R_1 = R_2$. This comes from the assumption that the DFE is the state of our population before the introduction of the disease. From this assumption, we get

$$\vec{x}_0 = \big(0, 0, S_1(0), S_2(0), 0, 0\big). \tag{4-3}$$

Following [van den Driessche and Watmough 2002], the square matrix F is calculated by taking the partial derivatives of $\mathcal{F}$ with respect to I_1 and I_2, evaluating at the DFE, and placing the resulting column vector into corresponding columns of F. Symbolically, we have

$$F = \begin{bmatrix} \frac{\partial \mathcal{F}}{\partial I_1}(\vec{x}_0) & \frac{\partial \mathcal{F}}{\partial I_2}(\vec{x}_0) \end{bmatrix} = \begin{bmatrix} \frac{\beta_{11}S_{10}}{S_{10}+S_{20}} & \frac{\beta_{12}S_{10}}{S_{10}+S_{20}} \\ \frac{\beta_{21}S_{20}}{S_{10}+S_{20}} & \frac{\beta_{22}S_{20}}{S_{10}+S_{20}} \end{bmatrix}, \tag{4-4}$$

where S_{10} and S_{20} are the initial populations of the S_1 and S_2 classes respectively. A similar process is used in calculating the square matrix V:

$$V = \begin{bmatrix} \frac{\partial \mathcal{V}}{\partial I_1}(\vec{x}_0) & \frac{\partial \mathcal{V}}{\partial I_2}(\vec{x}_0) \end{bmatrix} = \begin{bmatrix} f+g_1+e_1 & 0 \\ -f & e_2+g_2 \end{bmatrix} \quad \Rightarrow \quad V^{-1} = \frac{1}{\alpha\gamma}\begin{bmatrix} \gamma & 0 \\ f & \alpha \end{bmatrix},$$

where $\alpha = f + e_1 + g_1$ and $\gamma = e_2 + g_2$.

Generally, $R_0 \equiv \rho(FV^{-1})$, where ρ is the spectral radius, or maximum magnitude of the spectrum of a square matrix [van den Driessche and Watmough 2002]. Thus, R_0 is calculated by finding the spectrum of FV^{-1}, namely $\operatorname{eig}(FV^{-1})$, and determining the largest value in terms of absolute value. For our model,

$$\begin{aligned} FV^{-1} &= \frac{1}{\alpha\gamma(S_{10}+S_{20})}\begin{bmatrix} \beta_{11}S_{10} & \beta_{12}S_{10} \\ \beta_{21}S_{20} & \beta_{22}S_{20} \end{bmatrix}\begin{bmatrix} \gamma & 0 \\ f & \alpha \end{bmatrix} \\ &= \frac{1}{\alpha\gamma(S_{10}+S_{20})}\begin{bmatrix} \beta_{11}S_{10}\gamma + \beta_{12}S_{10}f & \beta_{12}S_{10}\alpha \\ \beta_{21}S_{20}\gamma + \beta_{22}S_{20}f & \beta_{22}S_{20}\alpha \end{bmatrix}. \end{aligned}$$

We consider

$$\lambda I - FV^{-1} = \begin{bmatrix} \lambda - \frac{\beta_{11}S_{10}\gamma+\beta_{12}S_{10}f}{\alpha\gamma(S_{10}+S_{20})} & -\frac{\beta_{12}S_{10}\alpha}{\alpha\gamma(S_{10}+S_{20})} \\ -\frac{\beta_{21}S_{20}\gamma+\beta_{22}S_{20}f}{\alpha\gamma(S_{10}+S_{20})} & \lambda - \frac{\beta_{22}S_{20}\alpha}{\alpha\gamma(S_{10}+S_{20})} \end{bmatrix}.$$

We find the determinant of $\lambda I - FV^{-1}$ and obtain

$$0 = \lambda^2 - \lambda\left(\frac{\beta_{11}S_{10}\gamma + \beta_{12}S_{10}f + \beta_{22}S_{20}\alpha}{\alpha\gamma(S_{10}+S_{20})}\right) + \frac{\beta_{22}S_{20}\alpha(\beta_{11}S_{10}\gamma + \beta_{12}S_{10}f) - \beta_{12}S_{10}\alpha(\beta_{21}S_{20}\gamma + \beta_{22}S_{20}f)}{(\alpha\gamma(S_{10}+S_{20}))^2},$$

$$0 = (\alpha\gamma(S_{10}+S_{20}))^2\lambda^2 - \lambda(\alpha\gamma(S_{10}+S_{20}))(\beta_{11}S_{10}\gamma + \beta_{12}S_{10}f + \beta_{22}S_{20}\alpha) + \alpha\gamma S_{10}S_{20}(\beta_{22}\beta_{11} - \beta_{12}\beta_{21}).$$

Let

$$\begin{aligned} \omega &= \alpha\gamma(S_{10}+S_{20}), \\ \eta &= \beta_{11}S_{10}\gamma + \beta_{12}S_{10}f + \beta_{22}S_{20}\alpha, \\ \psi &= \alpha\gamma S_{10}S_{20}(\beta_{11}\beta_{22} - \beta_{12}\beta_{21}). \end{aligned}$$

Then we have

$$0 = \omega^2\lambda^2 - \omega\eta\lambda + \psi \quad \text{or} \quad \lambda = \frac{\eta \pm \sqrt{\eta^2 - 4\psi}}{2\omega}.$$

Here, our parameters allow some useful simplification. Because $\beta_{11} = \beta_{12}$ and $\beta_{21} = \beta_{22}$, we can say $\psi = \alpha\gamma S_{10}S_{20}(\beta_{11}\beta_{22} - \beta_{12}\beta_{21}) = 0$. This gives us an abbreviated representation of our basic reproductive ratio,

$$\lambda = 0 \quad \text{or} \quad \lambda = \frac{\eta}{\omega}.$$

We are looking for the largest absolute value; therefore,

$$R_0 \equiv \frac{\beta_{11}S_{10}\gamma + \beta_{12}S_{10}f + \beta_{22}S_{20}\alpha}{\alpha\gamma(S_{10}+S_{20})}. \tag{4-5}$$

From the general equation for R_0 and the estimates listed previously, we find the numeric value of R_0 to be approximately 1.54. Since the R_0 value is greater than 1, the disease spreads. Consequently, with the disease spreading in this case, it would also spread with nonzero $\hat{\beta}_i$ for $i = 1, 2$. These coefficients do have an impact on the behavior of the model denoted in the graphics. With this R_0 value, this indicates that each primary cholera infection introduces the disease to a portion of the remaining susceptible population. In consideration of this fact, we employed a control discussed in Section 6.

5. Graphics with environmental forcing term

From the graph presented in Figure 3, we can see the disease-free equilibrium, or DFE, for our model is followed by a dramatic decrease in the susceptible population and a corresponding increase in the infected population (this may be clearer in Figure 4). The infected population quickly dwindles giving rise to a recovered population with some immunity. Perhaps unexpectedly, we see a

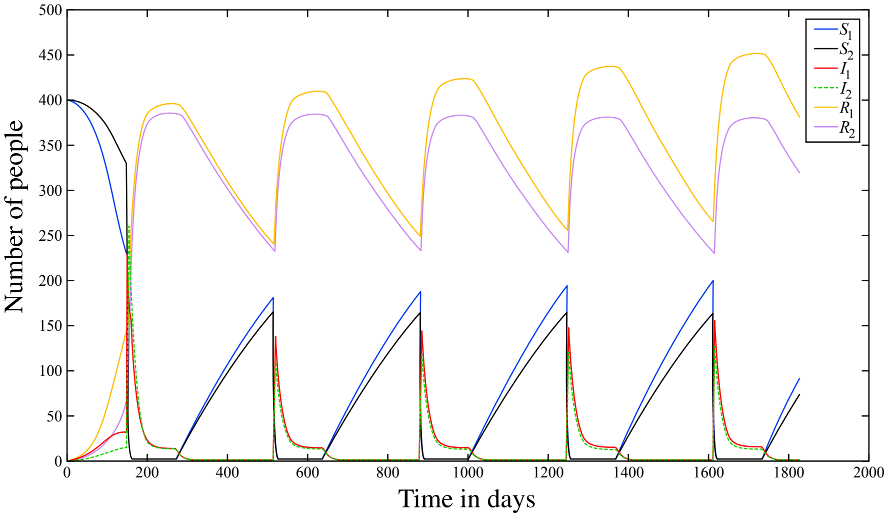


Figure 3. This represents a simulation over five years with initial data $S_1(0) = S_2(0) = 400$, $I_1(0) = I_2(0) = 1$ and $R_1(0) = R_2(0) = 0$.

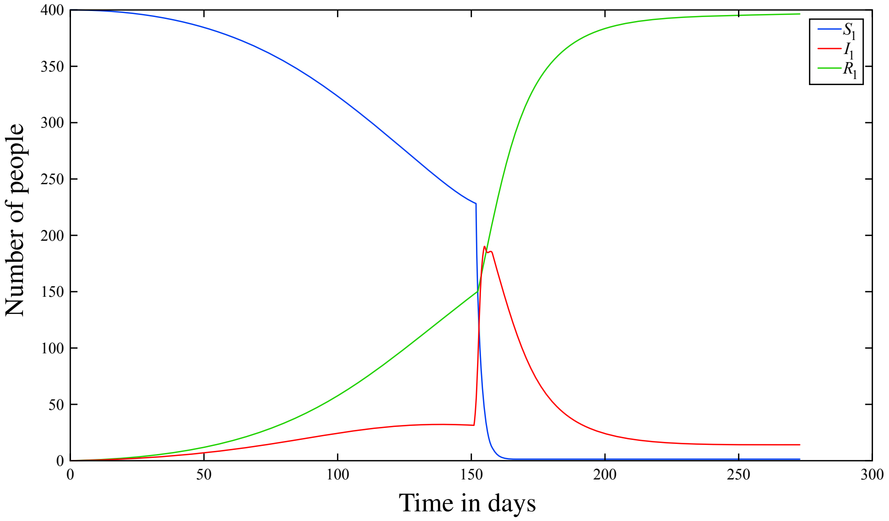


Figure 4. For the sake of clarity, this is the first 9 months of the previous simulation for children in Figure 3.

cyclical pattern representing seasonal spikes in infecteds and subsequent increases in recovered individuals. This is a result of the environmental forcing terms, $\hat{\beta}_i$, which were chosen to display the effects of the Bangladeshi monsoon season on cholera dynamics.

Notice, population growth still occurs over the course of time. Since this model does not assume a closed population, we would expect population growth to be a function of the country chosen for analysis. This conclusion seems to be in line with demographic data that indicates continued population growth despite the existence of endemic cholera in Bangladesh.

6. Optimal control

Before we consider control of the disease, we must establish the kind of control we will use. This selection has significant effect on the derivation of our new differential equations (those that include the dynamics of the control) and every subsequent step in the process of optimization. For the sake of this paper, we consider the effects of what may be called a protection control in which we limit the interaction of susceptible children with both infected classes. Our optimization process is an application addressed in the work of Lenhart and Workman [2007].

For this model we are interested in the effect of a simple control, one in which we limit the interactions of susceptible children. Symbolically, this is shown by inserting the control $(1-u(t))$ as a coefficient of the interaction terms $\beta_{11}S_1I_1/N$ and $\beta_{12}S_1I_2/N$. That is, the control allows only part of the I_1 and I_2 classes to interact with the S_1 class. Our system of differential equations becomes

$$\frac{dS_1}{dt}=bN-\frac{(1-u)\beta_{11}S_1I_1}{N}-\frac{(1-u)\beta_{12}S_1I_2}{N}-\hat{\beta}_1S_1-fS_1-d_1S_1+\omega_1R_1, \tag{6-1}$$

$$\frac{dS_2}{dt}=fS_1-\frac{\beta_{21}S_2I_1}{N}-\hat{\beta}_2S_2-\frac{\beta_{22}S_2I_2}{N}-d_2S_2+\omega_2R_2, \tag{6-2}$$

$$\frac{dI_1}{dt}=\frac{(1-u)\beta_{11}S_1I_1}{N}+\frac{(1-u)\beta_{12}S_1I_2}{N}+\hat{\beta}_1S_1-fI_1-g_1I_1-e_1I_1, \tag{6-3}$$

$$\frac{dI_2}{dt}=\frac{\beta_{21}S_2I_1}{N}+\frac{\beta_{22}S_2I_2}{N}+\hat{\beta}_2S_2+fI_1-g_2I_2-e_2I_2, \tag{6-4}$$

$$\frac{dR_1}{dt}=g_1I_1-fR_1-\omega_1R_1-d_1R_1, \tag{6-5}$$

$$\frac{dR_2}{dt}=g_2I_2+fR_1-\omega_2R_2-d_2R_2, \tag{6-6}$$

subject to initial conditions $S_1(0)=S_{10}$, $S_2(0)=S_{20}$, $I_1(0)=I_{10}$, $I_2(0)=I_{20}$, $R_1(0)=R_{10}$, $R_2(0)=R_{20}$.

We wish to minimize an objective functional that represents the members of each infected class and the cost of control implementation. Perhaps put more simply, we minimize the numbers of infected children and mature adults as well as the cost of the protection control. Having a goal and method in mind, we may write

$$J(u)=\int_0^T\big(AI_1(t)+BI_2(t)+Cu^2(t)\big)\,dt. \tag{6-7}$$

Note that if the protection control is at 0, then the cost to the population is the least in the objective functional. If the control $u(t)$ is low, then this means a small impact on transmission. Hence, this represents a small cost to the system.

We must define our control set U, which is the set of all possible control outcomes. This set is restricted to measurable functions and must be bounded. Thus,

$$U = \{u(t) \text{ measurable} \mid 0 \le u(t) \le u_{\max}\}. \tag{6-8}$$

In consideration of our goal, we will attempt to minimize $J(u)$ over the class of controls U subject to equations (6-1)–(6-6).

By determining the Hamiltonian for our system, we are able to determine the necessary conditions for optimality and transversality [Lenhart and Workman 2007]. This also allows us to determine the form of the adjoint equations by taking the negative derivative of the Hamiltonian with respect to each of the state variables. The Hamiltonian is

$$\begin{aligned}
H = {} & AI_1(t)+BI_2(t)+Cu^2(t)\\
& +\lambda_{S1}\left(bN-\frac{(1-u)\beta_{11}S_1I_1}{N}-\frac{(1-u)\beta_{12}S_1I_2}{N}-fS_1-d_1S_1+\omega_1R_1-\hat{\beta}_1S_1\right)\\
& +\lambda_{S2}\left(fS_1-\frac{\beta_{21}S_2I_1}{N}-\frac{\beta_{22}S_2I_2}{N}-d_2S_2+\omega_2R_2-\hat{\beta}_2S_2\right)\\
& +\lambda_{I1}\left(\frac{(1-u)\beta_{11}S_1I_1}{N}+\frac{(1-u)\beta_{12}S_1I_2}{N}-fI_1-g_1I_1-e_1I_1+\hat{\beta}_1S_1\right)\\
& +\lambda_{I2}\left(\frac{\beta_{21}S_2I_1}{N}+\frac{\beta_{22}S_2I_2}{N}+fI_1-g_2I_2-e_2I_2+\hat{\beta}_2S_2\right)\\
& +\lambda_{R1}(g_1I_1-fR_1-\omega_1R_1-d_1R_1)\\
& +\lambda_{R2}(g_2I_2+fR_1-\omega_2R_2-d_2R_2).
\end{aligned} \tag{6-9}$$

Because the state system is bounded, the work of Fleming and Rishel [1975] allows us to obtain the existence of an optimal control for our problem. Moreover, we can formulate the adjoint equations and the optimal control representation associated with the minimization of $J(u)$ subject to equations (6-1)–(6-6). We state the following theorem to do so and reference Lenhart and Workman [2007] for the details of the complementary proof.

Theorem. *Given an optimal control $u^* \in U$ and corresponding states*

$$(S_1^*, S_2^*, I_1^*, I_2^*, R_1^*, R_2^*),$$

there exist adjoint functions $(\lambda_{S_1}, \lambda_{S_2}, \lambda_{I_1}, \lambda_{I_2}, \lambda_{R_1}, \lambda_{R_2})$ satisfying

$$
\begin{aligned}
\frac{d\lambda_{S_1}}{dt} = &-\lambda_{S_1}\left(b-f-d_1+(1-u)\left(\frac{\beta_{11}S_1I_1}{N^2}-\frac{\beta_{11}I_1}{N}+\frac{\beta_{12}S_1I_2}{N^2}-\frac{\beta_{12}I_1}{N}\right)-\hat{\beta}_1\right)\\
&-\lambda_{S_2}\left(f+\frac{\beta_{21}S_2I_1}{N^2}+\frac{\beta_{22}S_2I_2}{N^2}\right)\\
&-\lambda_{I_1}\Bigg(-\frac{(1-u)\beta_{11}S_1I_1}{N^2}+\frac{(1-u)\beta_{11}I_1}{N}-\frac{(1-u)\beta_{12}S_1I_2}{N^2}\\
&\qquad\qquad+\frac{(1-u)\beta_{12}I_1}{N}+\hat{\beta}_1\Bigg)\\
&-\lambda_{I_2}\left(-\frac{\beta_{21}S_2I_1}{N^2}-\frac{\beta_{22}S_2I_2}{N^2}\right),
\end{aligned}
\tag{6-10}
$$

$$
\begin{aligned}
\frac{d\lambda_{S_2}}{dt} = &-\lambda_{S_1}\left(b+\frac{(1-u)\beta_{11}S_1I_1}{N^2}+\frac{(1-u)\beta_{12}S_1I_2}{N^2}\right)\\
&-\lambda_{S_2}\left(-d_2+\frac{\beta_{21}S_2I_1}{N^2}-\frac{\beta_{21}I_1}{N}+\frac{\beta_{22}S_2I_2}{N^2}-\frac{\beta_{22}I_2}{N}-\hat{\beta}_2\right)\\
&-\lambda_{I_1}\left(-\frac{(1-u)\beta_{11}S_1I_1}{N^2}-\frac{(1-u)\beta_{12}S_1I_2}{N^2}\right)\\
&-\lambda_{I_2}\left(-\frac{\beta_{21}S_2I_1}{N^2}+\frac{\beta_{21}I_1}{N}-\frac{\beta_{22}S_2I_2}{N^2}+\frac{\beta_{22}I_2}{N}+\hat{\beta}_2\right),
\end{aligned}
\tag{6-11}
$$

$$
\begin{aligned}
\frac{d\lambda_{I_1}}{dt} = &-\lambda_{S_1}\left(b+\frac{(1-u)\beta_{11}S_1I_1}{N^2}-\frac{(1-u)\beta_{11}S_1}{N}+\frac{(1-u)\beta_{12}S_1I_2}{N^2}\right)\\
&-\lambda_{S_2}\left(\frac{\beta_{21}S_2I_1}{N^2}-\frac{\beta_{21}S_2}{N}+\frac{\beta_{22}S_2I_2}{N^2}\right)\\
&-\lambda_{I_1}\left(-f-e_1-g_1-\frac{(1-u)\beta_{11}S_1I_1}{N^2}+\frac{(1-u)\beta_{11}S_1}{N}-\frac{(1-u)\beta_{12}S_1I_2}{N^2}\right)\\
&-\lambda_{I_2}\left(f-\frac{\beta_{21}S_2I_1}{N^2}+\frac{\beta_{21}S_2}{N}-\frac{\beta_{22}S_2I_2}{N^2}\right)-\lambda_{R_1}g_1,
\end{aligned}
\tag{6-12}
$$

$$
\begin{aligned}
\frac{d\lambda_{I_2}}{dt} = &-\lambda_{S_1}\left(b+\frac{(1-u)\beta_{11}S_1I_1}{N^2}+\frac{(1-u)\beta_{12}S_1I_2}{N^2}-\frac{(1-u)\beta_{12}S_1}{N}\right)\\
&-\lambda_{S_2}\left(\frac{\beta_{21}S_2I_1}{N^2}+\frac{\beta_{22}S_2I_2}{N^2}-\frac{\beta_{22}S_2}{N}\right)\\
&-\lambda_{I_1}\left(-\frac{(1-u)\beta_{11}S_1I_1}{N^2}-\frac{(1-u)\beta_{12}S_1I_2}{N^2}+\frac{(1-u)\beta_{12}S_1}{N}\right)\\
&-\lambda_{I_2}\left(-e_2-g_2-\frac{\beta_{21}S_2I_1}{N^2}-\frac{\beta_{22}S_2I_2}{N^2}+\frac{\beta_{22}S_2}{N}\right)-\lambda_{R_2}g_2,
\end{aligned}
\tag{6-13}
$$

$$\begin{aligned}\frac{d\lambda_{R_1}}{dt}&=-\lambda_{S_1}\Big(b+\frac{(1-u)\beta_{11}S_1I_1}{N^2}+\frac{(1-u)\beta_{12}S_1I_2}{N^2}+\omega_1\Big)\\&\quad-\lambda_{S_2}\Big(\frac{\beta_{21}S_2I_1}{N^2}+\frac{\beta_{22}S_2I_2}{N^2}\Big)\\&\quad-\lambda_{I_1}\Big(-\frac{(1-u)\beta_{11}S_1I_1}{N^2}-\frac{(1-u)\beta_{12}S_1I_2}{N^2}\Big)\\&\quad-\lambda_{I_2}\Big(-\frac{\beta_{21}S_2I_1}{N^2}-\frac{\beta_{22}S_2I_2}{N^2}\Big)-\lambda_{R_1}(-f-d_1-\omega_1)-f\lambda_{R_2},\end{aligned}\tag{6-14}$$

$$\begin{aligned}\frac{d\lambda_{R_2}}{dt}&=-\lambda_{S_1}\Big(b+\frac{(1-u)\beta_{11}S_1I_1}{N^2}+\frac{(1-u)\beta_{12}S_1I_2}{N^2}\Big)\\&\quad-\lambda_{S_2}\Big(\frac{\beta_{21}S_2I_1}{N^2}+\frac{\beta_{22}S_2I_2}{N^2}+\omega_2\Big)-\lambda_{I_1}\Big(-\frac{(1-u)\beta_{11}S_1I_1}{N^2}-\frac{(1-u)\beta_{12}S_1I_2}{N^2}\Big)\\&\quad-\lambda_{I_2}\Big(-\frac{\beta_{21}S_2I_1}{N^2}-\frac{\beta_{22}S_2I_2}{N^2}\Big)-\lambda_{R_2}(-d_2-\omega_2),\end{aligned}\tag{6-15}$$

with transversality conditions

$$\lambda_{S_1}(T)=\lambda_{S_2}(T)=\lambda_{I_1}(T)=\lambda_{I_2}(T)=\lambda_{R_1}(T)=\lambda_{R_2}(T)=0,$$

and the optimal control is characterized by

$$u^*=\min\Big(u_{\max},\max\Big(0,\frac{1}{2NC}(\lambda_{I_1}-\lambda_{S_1})(\beta_{11}S_1I_1+\beta_{12}S_1I_2)\Big)\Big).$$

6.1. ***Discussion of control and infected classes graphics.*** In an attempt to see the effects of various weights on our optimal control, we ran the model with A, B, C at various orders of magnitude. Also, due to possible imperfections of human protection implementation, we tested various control levels ranging from to 0.1 to 0.9. This was achieved reasonably quickly since our MATLAB code allows the bounds on the control to be specified as an input. Each of these maximum control variations displayed structurally similar results. It is worth noting that for the sake of clarity, Figures 5–7 have the control weights set to $A=1$, $B=1$, $C=10$, where the cost of the control receives ten times as much emphasis.

For a time span of four years (see Figure 5), we see that our treatment should start with maximum implementation decreasing around the start of the first monsoon season. Control potency decreases dramatically through the two-hundred day mark and spikes again around the same time that we would expect the monsoon season to return. This pattern seems to be repeated with milder and shorter implementation through the remainder of the treatment period. The resulting effects on the infected classes are shown in Figures 6 and 7 and are discussed below.

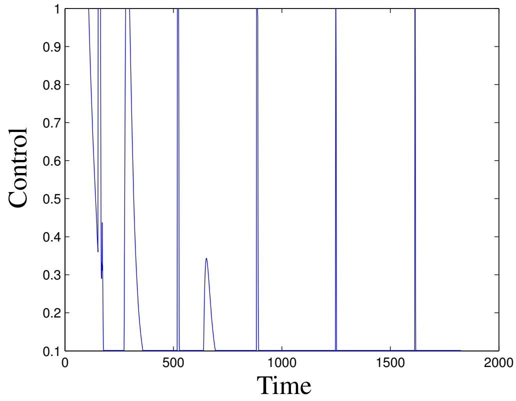


Figure 5. Our model with a forcing term and control implemented; this simulation covers a time span of 4 years with $S_1(0) = S_2(0) = 400$, $I_1(0) = I_2(0) = 1$ and $R_1(0) = R_2(0) = 0$.

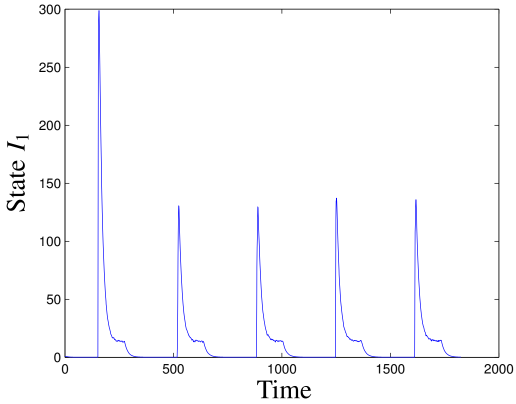


Figure 6. The effects of a five-year control on the I_1 class.

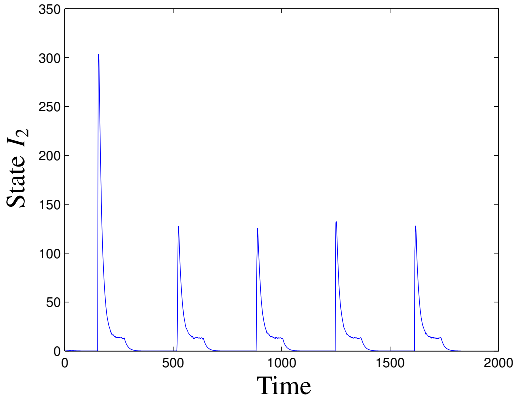


Figure 7. The effects of a five-year control on the I_2 class.

Toward the end of the first monsoon season, the model indicates that the protection control be implemented at approximately one-third efficacy. This increased protection control at the end of (or immediately following) the monsoon season, reflects data from Sack et al. [2003] which says the highest correlation seen between monsoon data and cholera outbreaks is a spike in outbreaks at the very end, immediately following the monsoon season.

In each infected class, we see a spike of infecteds at the beginning of the first monsoon season and decreasing immediately following. Notice that both classes continue to see yearly spikes, but the maximum number of infected individuals stabilizes and is limited to about 150 people per year.

7. Conclusions

Based on this specific model, it would seem advantageous to extend a protection control at the inception of each monsoon season. This would minimize the portion of the population in the infected classes over time. Additionally, it has the potential to be very cost-effective and practical. Lengthy protections would only be necessary the first year of the treatment process and this model could allow governments to schedule national protection or isolation days in advance, thus increasing the possibility of widespread compliance. One interesting insight offered by our data follows from the small increase in control around the end of the first monsoon season. It would seem that a small increase in protection control the first year is enough to disrupt the cycle of sickness and immunity loss in future years.

The results of our optimal control could have implications for consideration in future policy decisions in Bangladesh. With regard to the general strategy of analyzing treatments using nonhomogeneous age classes, benefits may be recognized through practical consideration. Social dynamics often vary greatly by age causing incongruencies between the assumptions of social homogeneity common to many epidemiological models and the actual practices of most cultures. For this reason, this work may offer clearer and more functional results for policy implementation.

References

[Barua and Greenough 1992] D. Barua and W. B. Greenough, III (editors), *Cholera*, Springer, New York, 1992.

[van den Driessche and Watmough 2002] P. van den Driessche and J. Watmough, “Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission”, *Math. Biosci.* **180** (2002), 29–48. MR 2003m:92071 Zbl 1015.92036

[Fleming and Rishel 1975] W. H. Fleming and R. W. Rishel, *Deterministic and stochastic optimal control*, Applications of Mathematics **1**, Springer, New York, 1975. MR 56 #13016 Zbl 0323.49001

[Harris et al. 2008] J. B. Harris, R. C. LaRorocque, F. Chowdhury, A. I. Khan, T. Logvinenko, A. S. G. Faruque, E. T. Ryan, F. Qadri, and S. B. Calderwood, “Susceptibility to *Vibrio cholerae* infection in a cohort of household contacts of patients with cholera in Bangladesh", *PLoS Negl. Trop. Dis.* **2**:4 (2008), e221.

[Keeling and Rohani 2008] M. J. Keeling and P. Rohani, *Modeling infectious diseases in humans and animals*, Princeton University Press, 2008. MR 2008f:92068 Zbl 1279.92038

[King et al. 2008] A. A. King, E. L. Ionides, M. Pascual, and M. J. Bouma, "Inapparent infections and cholera dynamics", *Nature* **454**:7206 (2008), 877–881.

[Lenhart and Workman 2007] S. Lenhart and J. T. Workman, *Optimal control applied to biological models*, Chapman & Hall/CRC, Boca Raton, FL, 2007. MR 2008f:49001 Zbl 1291.92010

[Nelson et al. 2009] E. J. Nelson, J. B. Harris, J. G. Morris, Jr., S. B. Calderwood, and A. Camilli, "Cholera transmission: the host, pathogen and bacteriophage dynamic", *Nat. Rev. Microbiol.* **7** (2009), 693–702.

[Ryan 2011] E. T. Ryan, "The cholera pandemic, still with us after half a century", *PLoS Negl. Trop. Dis.* **5**:1 (2011), e1003.

[Ryan and Charles 2011] E. T. Ryan and R. C. Charles, "Cholera in the 21st century", *Current Opinion in Infectious Diseases* **24**:5 (2011), 472–477.

[Sack et al. 2003] R. B. Sack, A. K. Siddique, J. Ira, M. Longini, A. Nizham, M. Yunus, M. S. Islam, J. J. G. Morris, A. Ali, A. Huq, G. B. Nair, F. Qadri, S. M. Faruque, D. A. Sack, and R. R. Colwell, "A 4-year study of the epidemiology of *Vibrio cholerae* in four rural areas of Bangladesh", *J. Infect. Dis.* **287**:1 (2003), 96–101.

[WHO 2012a] World Health Organization, "Cholera", website, 2012, http://www.who.int/topics/cholera/en/.

[WHO 2012b] World Health Organization, "Cholera fact sheet", website, 2012, http://www.who.int/mediacentre/factsheets/fs107/en/index.html.

[WHO 2012c] World Health Organization, "Epidemiology", website, 2012, http://www.who.int/topics/epidemiology/en/.

[WHO 2012d] World Health Organization, "The top 10 causes of death", website, 2012, http://www.who.int/mediacentre/factsheets/fs310/en/index.html.

[Wolfram Alpha 2010] Wolfram Alpha LLC, "Bangladesh annual births", website, 2010, http://www.wolframalpha.com/input/?i=bangladesh+births.

[Zhong 2011] P. Zhong, *Optimal Theory applied to integrodifference equation models in a cholera differential equation model*, Ph.D. thesis, University of Tennessee, 2011, http://trace.tennessee.edu/cgi/viewcontent.cgi?article=2287&context=utk_graddiss.

Received: 2014-01-07 Revised: 2014-12-07 Accepted: 2014-12-27

renee.fister@murraystate.edu *Department of Mathematics and Statistics,*
Murray State University, Murray, KY 42071, United States

hgaff@odu.edu *Department of Biological Sciences, Old Dominion University,*
Norfolk, VA 23529, United States

elsa.schaefer@marymount.edu *Department of Mathematics, Marymount University,*
Arlington, VA 22207, United States

glenna.buford@wooga.com *Engineer, Wooga GmbH, Berlin, Germany*

bryce.norris@gmail.com *Department of Mathematics and Statistics,*
Murray State University, Murray, KY 42071, United States

INVOLVE 9:1 (2016)
dx.doi.org/10.2140/involve.2016.9.101

Completions of reduced local rings with prescribed minimal prime ideals

Susan Loepp and Byron Perpetua

(Communicated by Scott T. Chapman)

Let T be a complete local ring of Krull dimension at least one, and let $\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_m$ each be countable sets of prime ideals of T. We find necessary and sufficient conditions for T to be the completion of a reduced local ring A such that A has exactly m minimal prime ideals $Q_1, Q_2, \dots, Q_m$, and such that, for every $i = 1, 2, \dots, m$, the set of maximal elements of $\{P \in \operatorname{Spec}(T) \mid P \cap A = Q_i\}$ is the set $\mathcal{C}_i$.

1. Introduction and preliminaries

As rings in general have a poorly understood structure but complete local rings are fully characterized, the relationship between local rings and their completions is an important area of study. Instead of beginning with a ring and examining its completion, we work backwards. In other words, we ask the question: when is a complete local ring T the completion of a local subring A if some restriction is placed on A? Certain restrictions have produced answers to this question. Notably, Lech [1986] gives necessary and sufficient conditions for T to be the completion of a local integral domain, and Heitmann [1993] does the same when A is required to be a local unique factorization domain.

Charters and Loepp [2004] address the question when A is a local integral domain whose generic formal fiber has finitely many maximal elements. For this paper, we define the generic formal fiber of an integral domain A to be the set $\{P \in \operatorname{Spec}(T) \mid P \cap A = (0)\}$, where T is the completion of A with respect to its maximal ideal. Charters and Loepp show that for any complete local ring T with maximal ideal M and collection G of prime ideals of T, where G has finitely many maximal elements, there exists a local integral domain A whose completion is T and whose generic formal fiber is precisely G if and only if G contains only the zero ideal and T is a field, or the following conditions are true:

MSC2010: 13B35, 13F25, 13J05, 13J10.
Keywords: completions of local rings, minimal prime ideals.

(1) $M \notin G$, and $\mathrm{Ass}(T) \subseteq G$.

(2) If $P \in \mathrm{Spec}(T)$ and $Q \in G$ with $P \subseteq Q$, then $P \in G$.

(3) If $Q \in G$, then the intersection of Q with the prime subring of T is $\langle 0 \rangle$.

The techniques employed in [Charters and Loepp 2004] also apply when A is required to be an excellent ring. In particular, the authors show that their main theorem holds for A excellent if two conditions are added to the three listed above: (1) T is equidimensional, and (2) for any P that is maximal in G, T_P is a regular local ring.

Suppose A is a local ring, T is the completion of A with respect to its maximal ideal, and Q is a prime ideal of A. We define the formal fiber of A at Q to be the set $\{P \in \mathrm{Spec}(T) \mid P \cap A = Q\}$. If the number of maximal elements of this set is finite, then we say that the formal fiber of A at Q is semilocal. In [Arnosti et al. 2012], the authors generalize the work of Charters and Loepp to the case where A is reduced with semilocal formal fibers at each of its minimal prime ideals, showing that such a ring A exists if T contains the rationals. Furthermore, they allow for control over the minimal prime ideals of A as follows: let $\mathcal{C}$ be a finite collection of prime ideals of T, and let $\mathcal{P} = \{\mathcal{C}_1, \dots, \mathcal{C}_m\}$ be a partition of $\mathcal{C}$ that simultaneously partitions all of the associated prime ideals of T (called a *feasible partition*). As A is constructed so that for any $\mathcal{C}_i$ and $P, P' \in \mathcal{C}_i$, we have $P \cap A = P' \cap A$, it is sensible to write $\mathcal{C}_i \cap A$ to denote $P \cap A$ for any $P \in \mathcal{C}_i$. Then the set of minimal prime ideals of A is precisely $\{\mathcal{C}_1 \cap A, \dots, \mathcal{C}_m \cap A\}$. Moreover, for each i, the formal fiber of A at $\mathcal{C}_i \cap A$ has maximal elements exactly the elements of $\mathcal{C}_i$. Defining Q_i as the intersection of all minimal prime ideals contained within any $P \in \mathcal{C}_i$, Arnosti et al. also show that A can be made excellent whenever

(1) T is reduced;

(2) for each Q_i and each $P \in \mathcal{C}_i$, $(T/Q_i)_{\overline{P}}$ is a regular local ring;

(3) for each Q_i, we have T/Q_i is equidimensional.

The method employed in [Arnosti et al. 2012] resembles that of [Loepp 2003] and [Heitmann 1994]. To ensure that T is the completion of A, Arnosti et al. construct A step by step so that it satisfies the conditions on R stated below:

Proposition 1.1 [Heitmann 1994, Proposition 1]. *Let T be a complete local ring with maximal ideal M. Suppose R is a quasilocal subring of T, $R \cap M$ is the maximal ideal of R, the map $R \to T/M^2$ is onto, and $IT \cap R = I$ for every finitely generated ideal I of R. Then R is Noetherian and the natural homomorphism $\widehat{R} \to T$ is an isomorphism.*

They begin with $\mathbb{Q}$, which is a subring of T by assumption, and repeatedly adjoin elements of T to build up A until it satisfies the conditions of Proposition 1.1. In order for the minimal prime ideals of A to have the desired properties, several constraints are enforced on each intermediate subring, the most important of which is that for any $\mathcal{C}_i$, it is the case that each prime ideal in $\mathcal{C}_i$, and every associated prime ideal contained within a prime ideal in $\mathcal{C}_i$, intersects identically with A, and that for each i this intersection is distinct. A subring satisfying these constraints is called an *intersection-preserving subring*, or IP subring (see Definition 1.4).

In this paper, we extend the work of Arnosti et al. by weakening several restrictions on T. First, we permit the collection $\mathcal{C}$ of prime ideals of T to be countably infinite, where previously it was required to be finite. This is a relatively simple task, as only one step in their construction needs to be fixed, and a lemma of Aiello, Loepp, and Vu [Aiello et al. 2015] appropriately extends the one lemma used in [Arnosti et al. 2012] that assumes $\mathcal{C}$ is finite. Second, we show that T need not contain the rationals. However, for the main theorem to hold in this case, T must satisfy a set of new conditions, detailed in our main theorem (Theorem 2.14), which we state here:

Theorem 2.14. *Let T be a complete local ring of dimension at least one, M be the maximal ideal of T, and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^{m})$ be a feasible partition. Then T is the completion of a reduced local subring A such that* $\operatorname{Min}(A) = \{\mathcal{C}_1 \cap A, \dots, \mathcal{C}_m \cap A\}$ *and the formal fiber of A at each $\mathcal{C}_i \cap A$ has countably many maximal elements, which are precisely the elements of $\mathcal{C}_i$, if and only if T has either zero or prime characteristic and at least one of the following is true*:

(1) $\operatorname{char}(T) \neq 0$.

(2) $\operatorname{char}(T) = 0$ *and* $M \cap \mathbb{Z} = \langle 0 \rangle$.

(3) $\operatorname{char}(T) = 0$ *and, for all $P \in \mathcal{C}$, we have $M \cap \mathbb{Z} \not\subseteq P$.*

(4) $\operatorname{char}(T) = 0$, $M \cap \mathbb{Z} = \langle p \rangle$ *for some prime integer p, and the following three conditions hold*:

   (a) *For each $P \in \mathcal{C}$ and for each $Q \in$* $\operatorname{Ass}(T)$ *with $Q \subseteq P$, we have $p \in Q$ whenever $p \in P$.*

   (b) *For each subcollection $\mathcal{C}_i$ and for any $P, P' \in \mathcal{C}_i$, we have $p \in P$ if and only if $p \in P'$.*

   (c) *For each $Q \in$* $\operatorname{Ass}(T)$, *if $p \in Q$, then* $\operatorname{Ann}_T(p) \not\subseteq Q$.

Furthermore, when one of the above four conditions is true, if J is an ideal of T such that $J \not\subseteq P$ for every $P \in \mathcal{C}$, A can be constructed so that the natural map $A \to T/J$ is onto.

We additionally prove that these conditions are necessary for the desired subring A of T to exist. While the work of [Arnosti et al. 2012] focuses on the excellent case, ours does not, as we suspect that it is very difficult to construct A to be excellent unless T contains the rationals. Only two lemmas in [loc. cit.] depend on T containing the rationals: Lemma 3.7, which allows A to be built up while remaining an intersection-preserving subring, and Lemma 3.11, which establishes the existence of an initial IP subring of T. Our approach is therefore to extend these two lemmas; our main result, Theorem 2.14, then follows.

We assume throughout this paper that all rings are commutative rings with unity and that local rings are Noetherian. The term *quasilocal ring*, on the other hand, denotes a ring that has one maximal ideal but is not necessarily Noetherian. The notation (R, M) indicates a quasilocal ring R with maximal ideal M.

For any complete local ring T containing the rationals and set of prime ideals $\mathcal{C} \subseteq \operatorname{Spec}(T)$, Arnosti et al. obtain a high degree of control over the minimal prime ideals of the reduced local subring A whose completion is T and over the formal fibers of $\operatorname{Min}(A)$. To achieve this control, $\mathcal{C}$ must be partitioned in such a way that the set of associated prime ideals of T is partitioned as well. The following definition formalizes this notion; we change it slightly so that $\mathcal{C}$ is allowed to be countable.

Definition 1.2 [Arnosti et al. 2012, Definition 2.2]. Let T be a complete local ring. Let $\mathcal{C} = \{P_i\}_{i=1}^{\infty}$ be a countable collection of incomparable nonmaximal prime ideals of T, and let $\mathcal{C}$ be partitioned into $m \geq 2$ subcollections $\mathcal{C}_1, \dots, \mathcal{C}_m$. We call $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^{m})$ a *feasible partition on* $\mathcal{C}$ (or simply a *feasible partition*) if, for each Q in $\operatorname{Ass}(T)$, $\mathcal{P}$ satisfies the following conditions:

(1) $Q \subseteq P_i$ for at least one $P_i \in \mathcal{C}$.

(2) There exists exactly one ℓ such that whenever $Q \subseteq P_i$, we have $P_i \in \mathcal{C}_\ell$.

Example 1.3. Let

$$T = \frac{\mathbb{Q}[[x, y, z]]}{\langle xyz \rangle},$$

and let $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_1, \mathcal{C}_2\})$, where $\mathcal{C}_1 = \{\langle x \rangle\}$ and $\mathcal{C}_2 = \{\langle y, z \rangle\}$. Then $\mathcal{P}$ is a feasible partition because each element of $\operatorname{Ass}(T) = \{\langle x \rangle, \langle y \rangle, \langle z \rangle\}$ is a subset of some $P \in \mathcal{C}$ and no element of $\operatorname{Ass}(T)$ is contained within more than one subcollection $\mathcal{C}_i$.

Central to the construction of A in [Arnosti et al. 2012] is the concept of an *intersection-preserving subring*, or IP subring. Following the approach of Heitmann [1994] and Loepp [2003], Arnosti et al. establish the existence of an IP subring of T, and then adjoin elements of T to build the ring A. Since A is constructed according to the feasible partition $\mathcal{P}$, it is essential that, for any i, the intersection of A with any $P \in \mathcal{C}_i$, or any associated prime ideal contained in any such P, is the same. Moreover, Arnosti et al. ensure that for every prime ideal P of T not contained within any prime ideal in $\mathcal{C}$, $P \cap A$ does not consist only of zerodivisors, so that only those prime ideals of T in $\mathcal{C}$ may be in the formal fiber of A at any minimal prime ideal of A. These requirements inspire their definition of an IP subring. We reproduce this definition below, with the only significant alteration again being that $\mathcal{C}$ may be countable.

Definition 1.4 [Arnosti et al. 2012, Definition 2.6]. Let (T, M) be a complete local ring, $\mathcal{C}$ a countable set of incomparable nonmaximal prime ideals of T and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ a feasible partition on $\mathcal{C}$. A quasilocal subring $(R, M \cap R)$ of T is called an *intersection-preserving subring* (IP subring) if the following conditions hold:

(1) R is infinite.

(2) For any $P \in \mathcal{C}$, we have $P \cap R = Q \cap R$ for any $Q \in \operatorname{Ass}(T)$ satisfying $Q \subseteq P$.

(3) For $P, P' \in \mathcal{C}$, we have $P, P' \in \mathcal{C}_i$ if and only if $P \cap R = P' \cap R$.

(4) For each $P \in \mathcal{C}$, we have $r \in P \cap R$ implies $\operatorname{Ann}_T(r) \not\subseteq P$.

The ring R is called *small intersection preserving* (abbreviated SIP) if, additionally, $|R| < |T|$.

We note here that in [Arnosti et al. 2012, Definition 2.6], $\operatorname{Min}(T)$ is used in part (2) instead of $\operatorname{Ass}(T)$. Using $\operatorname{Ass}(T)$ simply helps us keep track of the zerodivisors of T, and it is a minor change in the definition.

The following result, based on [Lee et al. 2001, Lemma 5], implies that an IP subring is reduced. Consequently, since the ring we construct in our main theorem is an IP subring, we know it is reduced. In [Arnosti et al. 2012, Lemma 2.8], more conditions were assumed, but were not needed in their proof. So we state the result here with only the needed conditions.

Lemma 1.5 [Arnosti et al. 2012, Lemma 2.8]. *Let T be a ring, $\mathcal{C}$ be a countable set of incomparable nonmaximal prime ideals of T, and $\mathcal{P}$ be a feasible partition on $\mathcal{C}$. Let R be a subring of T such that, for each $P \in \mathcal{C}$, if $r \in P \cap R$, then $\operatorname{Ann}_T(r) \not\subseteq P$. Then R is reduced.*

Throughout their paper, Arnosti et al. use the assumptions stated in the following remark.

Remark 1.6. Let (T, M) be a complete local ring of dimension at least one which contains the rationals. Let $\mathcal{C}$ be a finite set of incomparable nonmaximal prime ideals of T. Let $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ be a feasible partition, and let R be an IP subring of T. Let $P \in \mathcal{C}$; then $P \cap R$ is a prime ideal of R, and $P \in \mathcal{C}_i$ for some i. Abusing notation, we denote $P \cap R$ by $\mathcal{C}_i \cap R$. This abuse of notation makes sense because if $P, P' \in \mathcal{C}_i$, then $P \cap R = P' \cap R$.

Our assumptions closely follow those printed above, but with two substantial changes. First, $\mathcal{C}$ is permitted to be countable, not only finite. Second, we allow T to not contain the rationals, but add conditions that are necessary for the construction of A to be possible. If T contains the rationals, then every integer is a unit, so condition (2) of the following remark holds.

Remark 1.7. Hereafter, let (T, M) be a complete local ring of dimension at least one such that T has either zero or prime characteristic. Assume that at least one of the following is true:

(1) $\operatorname{char}(T) \neq 0$.

(2) $\operatorname{char}(T) = 0$ and $M \cap \mathbb{Z} = \langle 0 \rangle$.

(3) $\operatorname{char}(T) = 0$ and, for all $P \in \mathcal{C}$, we have $M \cap \mathbb{Z} \not\subseteq P$.

(4) $\operatorname{char}(T) = 0$, $M \cap \mathbb{Z} = p\mathbb{Z}$ for some prime p, and the following three conditions hold:

   (a) For each $P \in \mathcal{C}$ and for every $Q \in \operatorname{Ass}(T)$ with $Q \subseteq P$, we have $p \in Q$ whenever $p \in P$.

   (b) For each subcollection $\mathcal{C}_i$ and any $P, P' \in \mathcal{C}_i$, we have $p \in P$ if and only if $p \in P'$.

   (c) For each $Q \in \operatorname{Ass}(T)$, if $p \in Q$, then $\operatorname{Ann}_T(p) \not\subseteq Q$.

Let $\mathcal{C}$ be a countable set of incomparable nonmaximal prime ideals of T, and let $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ be a feasible partition. Suppose R is an IP subring of T. If $P \in \mathcal{C}_i$, then we denote $P \cap R$ by $\mathcal{C}_i \cap R$, a reasonable abuse of notation as we have guaranteed $P \cap R = P' \cap R$ for any $P, P' \in \mathcal{C}_i$.

2. Results

In [Arnosti et al. 2012, Definition 2.2], the collection $\mathcal{C}$ of nonmaximal prime ideals of T is required to be finite. Only Lemma 3.5 of their paper, however, uses the fact that $\mathcal{C}$ is finite and not merely countable. We therefore make a small modification to this lemma using the result below. This result generalizes [Arnosti et al. 2012, Lemma 3.4] (and, in fact, generalizes the prime avoidance theorem for complete local rings).

Lemma 2.1 [Aiello et al. 2015, Lemma 2.7]. *Let (T, M) be a complete local ring such that $\dim(T) \geq 1$, let $\mathcal{C}$ be a countable set of incomparable nonmaximal prime ideals of T, and let D be a subset of T such that $|D| < |T|$. Let I be an ideal of T such that $I \not\subseteq P$ for all $P \in \mathcal{C}$. Then $I \not\subseteq \bigcup\{r + P \mid r \in D, P \in \mathcal{C}\}$.*

The following result (both the statement and the proof) is taken almost exactly from [Arnosti et al. 2012, Lemma 3.5]. The statement of the result, however, has two minor changes. First, we state and prove, with the use of Lemma 2.1, that $\mathcal{C}$ can be countable instead of just finite. Second, for [Arnosti et al. 2012, Lemma 3.5], it is assumed that T contains the rationals. As that is not a necessary assumption in the proof, we need not assume T contains the rationals for our Lemma 2.2.

Lemma 2.2. *Let (T, M) and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ be as in Remark 1.7. Let R be an infinite subring of T such that $|R| < |T|$. Let J be an ideal of T such that*

$J \not\subseteq P$ for every $P \in \mathcal{C}$. Let $t, q \in T$. Then there exists an element $t' \in J$ such that for every $P \in \mathcal{C}$ with $q \notin P$, we have that $t + qt' + P \in T/P$ is transcendental over $R/(P \cap R)$. If, in addition, $Q \in \operatorname{Ass}(T)$, $P \in \mathcal{C}$ with $Q \subseteq P$, $q \notin P$, and $P \cap R = Q \cap R$, then $t + qt' + Q \in T/Q$ is transcendental over $R/(Q \cap R)$.

Proof. Let $\mathcal{G} = \{P \in \mathcal{C} \mid q \notin P\}$. Then $\mathcal{G}$ is a countable set of incomparable nonmaximal prime ideals of T. Suppose that $t + qt' + P = t + qs' + P$ for some $P \in \mathcal{G}$ and some $t', s' \in T$. Then $(t + qt') - (t + qs') = q(t' - s') \in P$. But $q \notin P$, so $(t' - s') \in P$. These steps are reversible, so $t + qt' + P = t + qs' + P$ if and only if $t' + P = s' + P$.

For each $P \in \mathcal{G}$, let $D_{(P)}$ be a full set of coset representatives of the cosets $t' + P$ that make $t + qt' + P \in T/P$ algebraic over $R/(P \cap R)$. Let $D = \bigcup_{P \in \mathcal{G}} D_{(P)}$. Then $|D_{(P)}| = |R/(P \cap R)| \leq |R| < |T|$ for every $P \in \mathcal{G}$, and noting that D is the countable union of sets with cardinality no greater than $|R|$, we have $|D| < |T|$. Now use Lemma 2.1 with $I = J$ and $\mathcal{C} = \mathcal{G}$ to conclude that there exists an element $t' \in J$ such that $t + qt' + P \in T/P$ is transcendental over $R/(P \cap R)$ for every $P \in \mathcal{G}$. Then we have that, for every $P \in \mathcal{C}$ with $q \notin P$, $t + qt' + P \in T/P$ is transcendental over $R/(P \cap R)$. Now suppose $Q \in \operatorname{Ass}(T)$, $P \in \mathcal{C}$ with $Q \subseteq P$, $q \notin P$, and $P \cap R = Q \cap R$. Then, $t + qt' + P \in T/P$ is transcendental over $R/(P \cap R)$. Since $P \cap R = Q \cap R$, we have $t + qt' + Q \in T/Q$ is transcendental over $R/(Q \cap R)$ as well. □

We now work to show that T need not contain the rationals. In [Arnosti et al. 2012], only Lemma 3.7, which allows the step-by-step construction of the IP subring A whose completion is T, and Lemma 3.11, which proves the existence of an initial IP subring of T, rely on T containing the rationals. Therefore, only these lemmas need to be modified in order for their results to hold when $\mathbb{Q} \not\subseteq T$. We handle these modifications in our Lemmas 2.8 and 2.13.

The following three technical lemmas are presented without proof.

Lemma 2.3 [Arnosti et al. 2012, Lemma 3.1]. *Let (T, M) and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ be as in Remark 1.7. Let B be a well-ordered index set and let R_β, $\beta \in B$ be a family of SIP subrings such that if $\beta, \gamma \in B$ such that $\beta < \gamma$, then $R_\beta \subseteq R_\gamma$. Then $R = \bigcup_{\beta \in B} R_\beta$ is an IP subring. Moreover, if there exists some $\lambda < |T|$ such that $|R_\beta| \leq \lambda$ for all β, and $|B| < |T|$, then $|R| \leq \max\{\lambda, |B|\}$, and R is an SIP subring.*

Lemma 2.4 [Arnosti et al. 2012, Lemma 3.2]. *Let (T, M) and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ be as in Remark 1.6. Let R be a subring of T satisfying all conditions for an IP subring except that it need not be quasilocal. Then $R_{(M \cap R)}$ is an IP subring of T with $|R_{(M \cap R)}| = |R|$. Additionally, if $|R| < |T|$, then $R_{(M \cap R)}$ is an SIP subring of T.*

Lemma 2.5 [Arnosti et al. 2012, Lemma 3.3]. *Let R be a subring of a complete local ring T. Let P_1, P_2 be prime ideals of T such that $P_1 \cap R = P_2 \cap R$. Suppose that, for $i = 1, 2$, we have that $u + P_i \in T/P_i$ is transcendental over $R/(P_i \cap R)$.*

Then $P_1 \cap R[u] = P_2 \cap R[u]$. *Furthermore, if* $\operatorname{Ann}_T(p) \not\subseteq P_1$ *for all* $p \in P_1 \cap R$, *then* $\operatorname{Ann}_T(p) \not\subseteq P_1$ *for all* $p \in P_1 \cap R[u]$.

To apply Proposition 1.1, we need to construct a subring A of T such that the map $A \to T/M^2$ is onto. The following lemma is a starting point for that. In addition, we will use it in the proof of Lemma 2.8. The statement and proof are taken almost exactly from [Arnosti et al. 2012, Corollary 3.6], but the proof is short and so we include it here.

Lemma 2.6 [Arnosti et al. 2012, Corollary 3.6]. *Let* (T, M) *and* $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ *be as in Remark 1.7, and let* J *be an ideal of* T *such that* $J \not\subseteq P$ *for every* $P \in \mathcal{C}$. *Let* R *be an SIP subring of* T *and* $t + J \in T/J$. *Then there exists an SIP subring* S *of* T *such that* $R \subseteq S \subset T$, $t + J$ *is in the image of the map* $S \to T/J$, *and* $|S| = |R|$. *Moreover, if* $t \in J$, *then* $S \cap J$ *contains a non-zerodivisor of* T.

Proof. Apply Lemma 2.2 with $q = 1$. Then $q \notin P$ for every $P \in \operatorname{Spec}(T)$, and so it is possible to choose $t' \in J$ such that $t + t' + P \in T/P$ is transcendental over $R/(R \cap P)$ for every $P \in \mathcal{C} \cup \operatorname{Ass}(T)$. Consider the ring $S = R[t + t']_{(M \cap R[t+t'])}$. By Lemma 2.5, $R[t + t']$ satisfies conditions (2), (3), and (4) of Definition 1.4. Further, $t + t' \in S$ and $(t + t') + J = t + J$, and so $t + J$ is in the image of the map $S \to T/J$.

Suppose $t \in J$ and $t + t'$ is a zerodivisor. Then $t + t' \in Q$ for some $Q \in \operatorname{Ass}(T)$. However, $Q \subseteq P$ for some $P \in \mathcal{C}$, and so $(t + t') + P = 0 + P$. Hence, $t + t' + P \in T/P$ is algebraic over $R/(R \cap P)$, a contradiction. Thus, $t + t'$ is a non-zerodivisor contained in $S \cap J$. □

Lemma 2.7 is an elementary result that will help us prove Lemma 2.8, a key lemma in this paper.

Lemma 2.7. *Let* R *be a ring and let* P *be a prime ideal of* R. *Let* $a, b \in R$, *and suppose* $a \notin P$. *If* ℓ *and* ℓ' *are units such that* $b + \ell a \in P$ *and* $b + \ell' a \in P$ *then* $\ell + P = \ell' + P$.

Proof. Suppose ℓ and ℓ' are units such that $b + \ell a \in P$ and $b + \ell' a \in P$. Then $(b + \ell a) - (b + \ell' a) = (\ell - \ell')a \in P$, and since $a \notin P$, we have $\ell - \ell' \in P$. That is, $\ell + P = \ell' + P$, completing the proof. □

The next lemma is analogous to [Arnosti et al. 2012, Lemma 3.7], which, given an SIP subring R, demonstrates the existence of an SIP subring S such that S contains R and, if I is a finitely generated ideal of R and c is an element of $IT \cap R$ then $c \in IS$. The proof is by induction on the number of generators of the ideal I of R. The only section requiring major modification for our result is the second part of the inductive step, in which it is proven that it is always possible to find generators for I satisfying the condition $(*)$ defined in the proof of the lemma. The portion of the proof of the lemma preceding the symbol $\diamondsuit$ is therefore quoted almost verbatim from the proof of [Arnosti et al. 2012, Lemma 3.7].

Lemma 2.8. *Let (T, M) and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ be as in Remark 1.7. Let R be an SIP subring of T. Then, for any finitely generated ideal I of R and any $c \in IT \cap R$, there exists a subring S of T with the following properties*:

(1) $R \subseteq S$.

(2) *S is an SIP subring of T.*

(3) $|S| = |R|$.

(4) $c \in IS$.

Proof. We shall proceed inductively on the number of generators of I. First suppose $I = aR$. If $a = 0$, then $S = R$ is the desired subring. Assume $a \neq 0$, and let $c = at$ for some $t \in T$. Note that, because $a \in R$, a is in some $P \in \mathcal{C}_i$ if and only if a is in every $P \in \mathcal{C}_i$. If this is the case, then, abusing notation, we shall refer to a as being contained in $\mathcal{C}_i$.

By condition (4) of Definition 1.4, $\text{Ann}_T(a) \nsubseteq P$ for all $P \in \mathcal{C}$ such that $a \in P$. By Lemma 2.1 with $D = \{0\}$ and $I = \text{Ann}_T(a)$, this means that $\text{Ann}_T(a) \nsubseteq \bigcup_{a \in P, P \in \mathcal{C}} P$. Thus, we can choose some $q \in \text{Ann}_T(a)$ such that $q \notin P$ for all $P \in \mathcal{C}$ such that $a \in P$. If $a \notin P$ for every $P \in \mathcal{C}$, we let $q = 0$. By Lemma 2.2, there exists some $t' \in T$ such that, for each $P \in \mathcal{C}$ with $a \in P$, the coset $t + qt' + P \in T/P$ is transcendental over $R/(P \cap R)$. Let $u = t + qt'$. We claim that $S = R[u]_{(R[u] \cap M)}$ is the desired subring. By Lemma 2.4, it suffices to show that $R[u]$ satisfies conditions (1), (2), (3), and (4) of being an SIP subring, and that $|R[u]| = |R|$. Condition (1) of Definition 1.4 follows immediately. We now show that condition (3) holds for $R[u]$.

For any $\mathcal{C}_i$ containing a, if $P, P' \in \mathcal{C}_i$, then $P \cap R[u] = P' \cap R[u]$ by Lemma 2.5. Next, consider any $\mathcal{C}_i$ not containing a. Let $P, P' \in \mathcal{C}_i$, and $f \in P \cap R[u]$. Then

$$f = r_n u^n + \cdots + r_1 u + r_0$$

for some $r_i \in R$. Multiplying both sides by a^n, we get

$$a^n f = r_n c^n + \cdots + a^{n-1} r_1 c + a^n r_0 \in P \cap R$$

since $au = at = c \in R$. Because R is an SIP subring, $a^n f \in P \cap R$ implies $a^n f \in P'$. However, by hypothesis, $a \notin P'$ and so f must be in P'. Consequently $f \in P' \cap R[u]$. The reverse inclusion follows by a similar argument, and so $P \cap R[u] = P' \cap R[u]$. Condition (2) of Definition 1.4 follows for $R[u]$ by a similar argument.

We will now show that condition (4) holds for $R[u]$. For each $P \in \mathcal{C}$, consider $f \in P \cap R[u]$, so that $f = r_n u^n + \cdots + r_1 u + r_0$. If $a \in P$, $u + P \in T/P$ is transcendental over $R/(P \cap R)$, so each r_i is in $P \cap R$. By assumption, for each r_i there exists a $q_i \notin P$ such that $r_i q_i = 0$. Let $q = \prod q_i \notin P$, and note that $fq = 0$. Thus, $\text{Ann}_T(f) \nsubseteq P$. If $a \notin P$, recall that $a^n f \in P \cap R$. By assumption, there exists a $q \notin P$ such that $qa^n f = 0$. Note that $qa^n \notin P$, so $\text{Ann}_T(f) \nsubseteq P$, and condition (4) holds. Hence, $R[u]_{(M\cap R[u])}$ is an SIP subring. Finally, observe that $|R[u]_{(M\cap R[u])}| = |R|$ and $c \in aR[u]_{(M\cap R[u])}$, as desired, so the lemma holds if I is generated by a single element.

Continuing inductively, suppose that the lemma holds when I is generated by $k-1$ elements, where $k \geq 2$. Let $I = \langle a_1, \dots, a_k\rangle R$ and $c = a_1t_1 + a_2t_2 + \cdots + a_kt_k \in R$ for some $t_i \in T$. We will first show that the lemma follows in the case where

$$\{\mathcal{C}_i \mid a_1 \in \mathcal{C}i\} = \{\mathcal{C}_j \mid a_2 \in \mathcal{C}_j\}. \tag{$*$}$$

We will then prove that it is always possible to define a generating set for I such that $(*)$ holds, completing the proof.

Assume that $(*)$ holds. Taking $a = a_1$, define q as in the principal case, and note that $a_1q = 0$. Thus, c can be rewritten as

$$c = a_1(t_1 + qt' + a_2t'') + a_2(t_2 - a_1t'') + a_3t_3 + \cdots + a_kt_k$$

for any $t', t'' \in T$. Let $u = t_1 + qt' + a_2t''$. We will choose t', t'' such that $u + P \in T/P$ is transcendental over $R/(P \cap R)$ for all $P \in \mathcal{C}$, allowing us to create an SIP subring $R[u]_{(M\cap R[u])}$.

Use Lemma 2.2 to find t' such that, for each $P \in \mathcal{C}$ with $q \notin P$, $t_1 + qt' + P \in T/P$ is transcendental over $R/(P \cap R)$. If $q \in P$ for all $P \in \mathcal{C}$, let $t' = 0$. By our choice of q and the assumption that $(*)$ holds, each $P \in \mathcal{C}$ contains precisely one of q and a_2. Thus, if $P \in \mathcal{C}$ is such that $q \notin P$, then

$$u + P = t_1 + qt' + a_2t'' + P = t_1 + qt' + P \in T/P$$

is transcendental over $R/(P \cap R)$ regardless of the choice of t''. Now, if $P \in \mathcal{C}$ is such that $q \in P$, then $a_2 \notin P$, and so we can use Lemma 2.2 to find $t'' \in T$ such that $t_1 + a_2t'' + P$ is transcendental over $R/(P \cap R)$ for all $P \in \mathcal{C}$ satisfying $a_2 \notin P$. If $a_2 \in P$ for all $P \in \mathcal{C}$, then let $t'' = 0$. By our choice of t' and t'', we have that $u + P$ is transcendental over $R/(P \cap R)$ for all $P \in \mathcal{C}$. By Lemma 2.5, $R[u]$ satisfies condition (3) of Definition 1.4. Using an identical argument to the principal case, $R[u]$ satisfies condition (4). It clearly satisfies conditions (1) and (2), and $|R[u]| = |R|$. By Lemma 2.4, $R' = R[u]_{(M\cap R[u])}$ is an SIP subring of T with $|R'| = |R|$.

Now let $J = \langle a_2, a_3, \dots, a_k\rangle R'$ and

$$c^* = c - a_1u = a_2(t_2 - a_1t'') + a_3t_3 + \cdots + a_kt_k.$$

We have $c \in R \subseteq R'$ and $a_1u \in R'$, so $c^* \in JT \cap R'$. By our inductive hypothesis, there exists an SIP subring S of T containing R' such that $c^* \in JS$, and so $c^* = a_2s_2 + \cdots + a_ks_k$ for some $s_i \in S$. It follows that $c = a_1u + a_2s_2 + \cdots + a_ks_k \in IS$, so S is the desired SIP subring.

$\diamondsuit$ We will now show that, given a set of generators $\langle a_1, a_2, \dots, a_k\rangle$ for I, we can reduce to the case that I satisfies $(*)$.

We first use Lemma 2.6 with $J = M$ and $t = 0$ to find an SIP subring R_0 of T such that $R \subseteq R_0 \subset T$, $|R_0| = |R|$, and $R_0 \cap M$ contains a non-zerodivisor, which we call m_0, of T. By condition (2) of Definition 1.4, if $P \in \mathcal{C}$, then $R_0 \cap P$ contains only zerodivisors of T. It follows that for every i, $m_0 \notin \mathcal{C}_i \cap R_0$.

Next, for each $P \in \mathcal{C}$, let $D_{(P)}$ be a full set of coset representatives of the cosets $t + P \in T/P$ that are algebraic over $R_0/(R_0 \cap P)$. Let $D' = \bigcup_{P \in \mathcal{C}} D_{(P)}$. Use Lemma 2.1 with $I = M$ and $D = D' \cup \{m_0\}$ to find an element m_1 of M such that, for all $P \in \mathcal{C}$, we have $m_1 + P \neq m_0 + P$ and $m_1 + P \in T/P$ is transcendental over $R_0/(R_0 \cap P)$. If $Q \in \operatorname{Ass}(T)$, then $Q \subseteq P$ for some $P \in \mathcal{C}$. Since R_0 is an SIP subring, we have $P \cap R_0 = Q \cap R_0$. It follows that $m_1 + Q \in T/Q$ is transcendental over $R_0/(R_0 \cap Q)$. Let $R_1 = R_0[m_1]_{(R_0 \cap M)}$. By Lemmas 2.4 and 2.5, R_1 satisfies conditions (2), (3), and (4) of being an IP subring. Clearly R_1 is infinite, and $|R_1| = |R_0| < |T|$. Thus, R_1 is an SIP subring of T.

Now, repeat the above procedure with R_0 replaced by R_1 and D replaced by $D' \cup \{m_0, m_1\}$ to obtain an element m_2 of M and an SIP subring R_2 of T such that $R_1 \subseteq R_2$ and, for every $P \in \mathcal{C}$, we have $m_2 + P \neq m_0 + P$ and $m_2 + P \neq m_1 + P$. Continue so that for every $n \in \{1, 2, \ldots\}$, we find $m_n \in M$ and R_n such that $R_{n-1} \subseteq R_n$, $|R_n| = |R_{n-1}|$, R_n is an SIP subring of T, and, for every $P \in \mathcal{C}$ and every $i < n$, we have $m_n + P \neq m_i + P$. Let $R' = \bigcup_{i=1}^{\infty} R_i$. Then if $P \in \mathcal{C}$, we have $m_i + P = m_j + P$ if and only if $i = j$. In addition, by Lemma 2.3, R' is an SIP subring and $|R'| = |R|$. Since $m_0 \in R' \cap M$, and $m_0 \notin P$ for all $P \in \mathcal{C}$, we have $\mathcal{C}_i \cap R' \neq M \cap R'$ for all $i = 1, 2, \ldots, m$. Also note that, for every i, in the ring $R'/(\mathcal{C}_i \cap R')$, we have $m_k + (\mathcal{C}_i \cap R') = m_j + (\mathcal{C}_i \cap R')$ if and only if $k = j$. It follows that $(1 + m_k) + (\mathcal{C}_i \cap R') = (1 + m_j) + (\mathcal{C}_i \cap R')$ if and only if $k = j$.

Now, $m_0 \in M \cap R'$ is not a zerodivisor of T and $\mathcal{C}_i \cap R'$ only contains zerodivisors of T, and so $m_0 \notin \mathcal{C}_i \cap R'$ for every i. Since m_0 is a nonunit, $m_0 + 1$ is a unit. We will consider an ideal of R' of the form $\langle m_0 a_1 + u a_2, a_1 - u a_2, a_3, \ldots, a_k \rangle$, where u is a unit we will choose later so that $(*)$ holds. This ideal is equal to $\langle (m_0 + 1) a_1, (m_0 + 1) u a_2, a_3, \ldots, a_k \rangle R'$ and therefore also equal to IR'.

For each $\mathcal{C}_i$, we know that $\mathcal{C}_i \cap R'$ is a nonmaximal prime ideal of R'. Therefore, since neither m_0 nor u is in any $\mathcal{C}_i \cap R'$, we have that for each $\mathcal{C}_i$, $m_0 a_1 \in \mathcal{C}_i \cap R'$ if and only if $a_1 \in \mathcal{C}_i \cap R'$, and $u a_2 \in \mathcal{C}_i \cap R'$ if and only if $a_2 \in \mathcal{C}_i \cap R'$. It follows that if $a_1, a_2 \in \mathcal{C}_i \cap R'$, then $m_0 a_1 + u a_2, a_1 - u a_2 \in \mathcal{C}_i \cap R'$. On the other hand, if $a_1 \in \mathcal{C}_i \cap R'$ but $a_2 \notin \mathcal{C}_i \cap R'$, then $m_0 a_1 + u a_2, a_1 - u a_2 \notin \mathcal{C}_i \cap R'$. The same holds if $a_1 \notin \mathcal{C}_i \cap R'$ but $a_2 \in \mathcal{C}_i \cap R'$.

Finally, consider the case where $a_1, a_2 \notin \mathcal{C}_i \cap R'$. As $m_0 a_1 \notin \mathcal{C}_i \cap R'$, by Lemma 2.7, every unit $\ell \in R'$ such that $m_0 a_1 + \ell a_2 \in \mathcal{C}_i \cap R'$ is in the same coset of $R'/(\mathcal{C}_i \cap R')$. Similarly, every unit ℓ' such that $a_1 - \ell' a_2 \in \mathcal{C}_i \cap R'$ is in the same coset of $R'/(\mathcal{C}_i \cap R')$. For each i, let ℓ_{i_+} be a representative of the coset of $R'/(\mathcal{C}_i \cap R')$ containing all units ℓ such that $m_0 a_1 + \ell a_2 \in \mathcal{C}_i \cap R'$, and let ℓ_{i_-} be a representative of the coset containing all units ℓ' such that $a_1 - \ell' a_2 \in \mathcal{C}_i \cap R'$. Since $L = \bigcup_{i=1}^{m} \{\ell_{i_+}, \ell_{i_i}\}$, is a finite set of elements of R' and $\mathcal{G} = \{\mathcal{C}_i \cap R'\}_{i=1}^{m}$ is a finite set of prime ideals of R', the set $\{\ell + P \mid \ell \in L, P \in \mathcal{G}\}$ is a finite set. Suppose for some $r \neq k$, $\ell \in L$, and $P \in \mathcal{G}$, we have $m_r + 1 \in \ell + P$ and $m_k + 1 \in \ell + P$. Then $m_r + P = m_k + P$, a contradiction. As the set $\{m_i + 1\}_{i=1}^{\infty}$ is infinite, there must be a positive integer r such that $m_r + 1 \notin \ell + P$ for all $\ell \in L$ and all $P \in \mathcal{G}$. So there exists a unit $u = m_r + 1 \in R'$ such that $u \notin \bigcup_{\ell \in L, P \in \mathcal{G}} \{\ell + P\}$. It follows that, for all i, we have $m_0 a_1 + u a_2 \notin \mathcal{C}_i \cap R'$ and $a_1 - u a_2 \notin \mathcal{C}_i \cap R'$.

We have now shown that, for every i, either $m_0 a_1 + u a_2$ and $a_1 - u a_2$ are both in $\mathcal{C}_i \cap R'$ or $m_0 a_1 + u a_2$ and $a_1 - u a_2$ are both not in $\mathcal{C}_i \cap R'$. Thus

$$\{\mathcal{C}_i \mid m_0 a_1 + u a_2 \in \mathcal{C}_i\} = \{\mathcal{C}_j \mid a_1 - u a_2 \in \mathcal{C}_j\},$$

and so we have $(*)$, and the previous argument applies with R replaced by R' and I replaced by IR'. Hence we can find an SIP subring S of T containing R' so that $c \in (IR')S = IS$. □

We use the following lemma to construct our ring to satisfy the hypotheses of Proposition 1.1.

Lemma 2.9 [Arnosti et al. 2012, Lemma 3.8]. *Let (T, M) and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^{m})$ be as in Remark 1.7. Let J be an ideal of T such that $J \not\subseteq P$ for all $P \in \mathcal{C}$, and let $u + J \in T/J$. Suppose R is an SIP subring of T. Then there exists an SIP subring S of T with the following properties*:

(1) $R \subseteq S$.

(2) *If $u \in J$, then $S \cap J$ contains a non-zerodivisor.*

(3) *$u + J$ is in the image of the map $S \to T/J$.*

(4) $|S| = |R|$.

(5) *For every finitely generated ideal I of S, we have $IT \cap S = I$.*

Proof. The proof follows from the proof of [Arnosti et al. 2012, Lemma 3.8], using our Lemma 2.8 where they used their Lemma 3.7. □

Lemma 2.13 demonstrates the existence of an SIP subring of T that serves as the starting point for the construction of A. Before building an SIP subring, we first find a ring in which every condition of the definition of an SIP subring is satisfied, except that each $\mathcal{C}_i \cap R$ need not be distinct. Such a ring is called *semi-SIP* and is formally defined below. Lemma 2.12 establishes a semi-SIP subring of T, making use of Lemma 2.11 in the case where T has prime characteristic.

Definition 2.10 [Arnosti et al. 2012, Definition 3.9]. Let (T, M) and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^{m})$ be as in Remark 1.7. We say that a quasilocal subring $(R, M \cap R)$ is a *semi-SIP subring of T* if the following conditions hold:

(1) R is infinite.

(2) For any $P \in \mathcal{C}$, we have $P \cap R = Q \cap R$ for any $Q \in \operatorname{Ass}(T)$ satisfying $Q \subseteq P$.

(3) For any $\mathcal{C}_i$, if $P, P' \in \mathcal{C}_i$, then $P \cap R = P' \cap R$.

(4) For each $P \in \mathcal{C}$, we have $r \in P \cap R$ implies $\operatorname{Ann}_T(r) \not\subseteq P$.

(5) $|R| < |T|$.

The following lemma is based on [Arnosti et al. 2012, Lemma 3.10]. Since the statement of that lemma requires a semi-SIP subring R of T but their proof does not use the fact that R is infinite, we weaken their requirement accordingly, and our lemma holds using their original proof. It is important to note that in their proof, the ring $R[up_i]_{(R[up_i] \cap M)}$ is infinite regardless of whether or not R is infinite.

Lemma 2.11 [Arnosti et al. 2012, Lemma 3.10]. *Let (T, M) and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ be as in Remark 1.7, and fix $\mathcal{C}_i$. Let R be a semi-SIP subring of T except that R need not be infinite, and let $p_i \in T$ be given such that $p_i \in Q$ for every minimal prime ideal Q contained within some $P \in \mathcal{C}_i$, but $p_i \notin P$ for any $P \in \mathcal{C}_j$, where $j \neq i$. Suppose further that* $\operatorname{Ann}_T(p_i) \not\subseteq P$ *for any $P \in \mathcal{C}_i$. Then there exists a unit u in T such that $R[up_i]_{(R[up_i] \cap M)}$ is a semi-SIP subring of T.*

In the proof of [Arnosti et al. 2012, Lemma 3.10], the authors define $S = R[p_i]$, and they note that, in their case, $|S| = |R| < |T|$. If R is finite, then the equality $|S| = |R|$ may not hold, but S is at most countable. Since complete local rings of positive dimension have cardinality greater than or equal to the cardinality of the real numbers (by, for example, [Charters and Loepp 2004, Lemma 2.3]), the inequality $|S| < |T|$ still holds.

Lemma 2.12. *Let (T, M) and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ be as in Remark 1.7. Then there exists a semi-SIP subring of T.*

Proof. By the assumptions of Remark 1.7, the characteristic of T is either 0 or some prime $\hat{p}$. In both cases, we will construct a semi-SIP subring of T.

Characteristic 0: First assume that the characteristic of T is zero. We examine three subcases, each assuming that only one of conditions (2), (3), and (4) of Remark 1.7 hold.

To begin, assume only condition (2) of Remark 1.7 is true: that is, $M \cap \mathbb{Z} = \langle 0 \rangle$. Since M contains no integers, every integer in T is a unit, so T contains the rationals. By [Arnosti et al. 2012, Lemma 3.11], there exists a semi-SIP subring of T.

Now assume that only condition (3) holds. Here $M \cap \mathbb{Z} \neq \langle 0 \rangle$, and so $M \cap \mathbb{Z} = p\mathbb{Z}$ for some prime integer p, but p is not in any of the prime ideals $P \in \mathcal{C}$. Choose any $P \in \mathcal{C}$. Since $p \notin P$, we have $P \cap \mathbb{Z} \neq p\mathbb{Z}$, and as $P \cap \mathbb{Z} \subseteq M \cap \mathbb{Z}$, we have $P \cap \mathbb{Z} = \langle 0 \rangle$. Let $R_0 = \mathbb{Z}_{(p\mathbb{Z})}$, so that R_0 is a local subring of T. Then $P \cap R_0 = \langle 0 \rangle$. Furthermore, for any $Q \in \operatorname{Ass}(T)$, we know that $Q \subseteq P$ for some

$P \in \mathcal{C}$, and so $Q \cap R_0 = \langle 0 \rangle$. Conditions (2), (3), and (4) of Definition 2.10 follow easily from these results. As R_0 is countably infinite and T has cardinality greater than or equal to the cardinality of the real numbers (by [Charters and Loepp 2004, Lemma 2.3]), $|R_0| < |T|$ and so R_0 is a semi-SIP subring of T.

Finally, assume that only condition (4) of Remark 1.7 holds. Again, given $M \cap \mathbb{Z} = p\mathbb{Z}$, let $R_0 = \mathbb{Z}_{(p\mathbb{Z})}$, so that R_0 is an infinite local subring of T with $|R_0| < |T|$. For any prime ideal P of T, $P \cap \mathbb{Z}$ is either $p\mathbb{Z}$, if $p \in P$, or $\langle 0 \rangle$, if $p \notin P$. Conditions (2) and (3) of Definition 2.10 follow respectively from conditions (4a) and (4b) of Remark 1.7. It remains to show that condition (4c) of Remark 1.7 implies condition (4) of Definition 2.10.

Recall that condition (4c) of Remark 1.7 ensures that for each $Q \in \operatorname{Ass}(T)$, if $p \in Q$, then $\operatorname{Ann}_T(p) \not\subseteq Q$. For contradiction, suppose that condition (4) of Definition 2.10 does not hold, so that there exists some $P \in \mathcal{C}$ and $r \in P \cap R_0$ with $\operatorname{Ann}_T(r) \subseteq P$. It must be that $p \in P$; if $p \notin P$, then $P \cap R_0 = \langle 0 \rangle$ and clearly $\operatorname{Ann}_T(0) \not\subseteq P$. As $P \cap R_0 = p\mathbb{Z}$ and $\operatorname{Ann}_T(p) \subseteq \operatorname{Ann}_T(kp)$ for any integer k, we may assume that $r = p$ and $\operatorname{Ann}_T(p) \subseteq P$. The set of zerodivisors of T_P is equal to $\bigcup\{QT_P : Q \in \operatorname{Ass}(T), Q \subseteq P\}$. Since $\operatorname{Ann}_{T_P}(p)$ consists entirely of zerodivisors, by the prime avoidance theorem, $\operatorname{Ann}_{T_P}(p) \subseteq QT_P$ for some $Q \in \operatorname{Ass}(T)$ with $Q \subseteq P$. Choose any $a \in \operatorname{Ann}_T(p)$, so that $ap = 0$. In T_P, we have $(a/1)(p/1) = 0/1$; thus $a/1 \in \operatorname{Ann}_{T_P}(p)$ and by above $a/1 \in QT_P$. It follows that $a \in Q$ and so $\operatorname{Ann}_T(p) \subseteq Q$. By condition (4c) of Remark 1.7, $p \notin Q$, but by condition (4a) of Remark 1.7, because $p \in P$, we have $p \in Q$, a contradiction. Therefore, condition (4) of Definition 2.10 holds, and R_0 is a semi-SIP subring of T.

Characteristic $\hat{p}$: We now assume that the characteristic of T is $\hat{p}$ for some prime integer $\hat{p}$. In this case, let $R_0 = \mathbb{Z}_{\hat{p}}$. Since R_0 is a field, no prime ideal of T contains a nonzero element of R_0. Therefore,

$$P \cap R_0 = \langle 0 \rangle = P' \cap R_0 = Q \cap R_0$$

for any $P, P' \in \mathcal{C}$ and $Q \in \operatorname{Ass}(T)$. Furthermore, it is trivially true that $\operatorname{Ann}_T(r) \not\subseteq P$ for any $P \in \mathcal{C}$ and $r \in P \cap R_0$, and clearly $|R_0| < |T|$. Thus R_0 satisfies every condition of being semi-SIP except for being infinite. By adjoining a carefully chosen element to R_0, we will create a ring that is semi-SIP. We will choose this element using the following method from [Arnosti et al. 2012, Lemma 3.11].

Let $\operatorname{Min}(T) = \{Q_1, \dots, Q_n\}$. For each minimal prime ideal Q_i, by Lemma 2.1 we can pick some $q_i \in Q_i$ such that $q_i \notin \bigcup\{P \in \mathcal{C} \mid Q_i \not\subseteq P\}$. Let $q = \prod_{i=1}^n q_i$. As q is nilpotent, we may choose ℓ to be the smallest positive integer such that $q^\ell = 0$. Choose any $\mathcal{C}_k$, let $\mathcal{E}_k$ be the set of minimal prime ideals of T contained within $\mathcal{C}_k$, and let $p_k = \prod_{Q_i \in \mathcal{E}_k} q_i^\ell$ and $s_k = \prod_{Q_i \notin \mathcal{E}_k} q_i^\ell$. As $p_k s_k = \prod_{i=1}^n q_i^\ell = \left(\prod_{i=1}^n q_i\right)^\ell = q^\ell = 0$, we have $\operatorname{Ann}_T(p_k) \not\subseteq P$ for every $P \in \mathcal{C}_k$.

Apply Lemma 2.11, setting $R = R_0$ and $p_i = p_k$ as chosen above, to find a unit $u \in T$ such that $R_0[up_k]_{(R_0[up_k] \cap M)}$ is the desired semi-SIP subring. □

Lemma 2.13. *Let (T, M) and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ be as in Remark 1.7. Then there exists an SIP subring of T. On the other hand, if* char(T) *is neither zero nor a prime p, or if none of the conditions in Remark 1.7 holds, then there does not exist an IP subring of T whose completion is T.*

Proof. By Lemma 2.12, there exists a semi-SIP subring of T, which we call R_0. Now use the proof of [Arnosti et al. 2012, Lemma 3.11], in which elements $p_1, \dots, p_m$ are adjoined to R_0 in such a way that the resulting ring is an SIP subring of T. The process is similar to the method we used in the characteristic $\hat{p}$ part of the proof of Lemma 2.12. So if (T, M) and $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ are as in Remark 1.7, then there exists an SIP subring of T.

Next, we show that the characteristic of T must be either zero or prime in order to construct an IP subring. We claim that if this condition does not hold, then char(T) must be a prime power p^k, where $k > 1$. Suppose otherwise, so that R has nonzero characteristic n, where n can be written as $n = ab$ for some relatively prime integers $1 < a, b < n$. Then a and b are zerodivisors and consequently nonunits, and we have $a, b \in M$. Therefore $\langle a, b\rangle \subseteq M$. By Bezout's identity, there exist integers r and s such that $ra + sb = 1$, implying $1 \in M$, a contradiction since M consists only of nonunits. Therefore a and b cannot be relatively prime, and char(T) must be a prime power. This implies that the prime subring of T is $\mathbb{Z}_{p^k}$ for some prime p and $k > 1$. Then, however, p is nilpotent in every subring of T, so no subring of T is reduced. As every IP subring is reduced, T has no IP subring.

Second, to prove that for the construction of an IP subring it is necessary to have at least one of the four conditions in Remark 1.7, we will show that if conditions (1), (2), and (3) do not hold and any one of conditions (4a), (4b), and (4c) does not hold, then an IP subring cannot exist. As A must be an IP subring in order for us to control its minimal prime ideals, the conditions of Remark 1.7 are necessary for Theorem 2.14.

Suppose that an IP subring R of T exists when conditions (1), (2), (3), and (4a) fail. Thus char$(T) = 0$, $M \cap \mathbb{Z} = p\mathbb{Z}$ for some prime p, and there exists some $P \in \mathcal{C}$ and some $Q \in \mathrm{Ass}(T)$ contained in P such that $p \in P$ but $p \notin Q$. By condition (2) of Definition 1.4, $P \cap R = Q \cap R$, so $p \notin R$, but this is impossible as any subring of T must contain the integers. Thus condition (4a) of Remark 1.7 is necessary in the absence of conditions (1), (2), and (3). Now suppose that there exists an IP subring R of T and conditions (1), (2), (3), and (4b) of Remark 1.7 do not hold, so $P \cap \mathbb{Z} \neq P' \cap \mathbb{Z}$ for some $\mathcal{C}_i$ and $P, P' \in \mathcal{C}_i$. As R contains the integers, $P \cap R \neq P' \cap R$, but this contradicts condition (3) of Definition 1.4, and we conclude that condition (4b) of Remark 1.7 is also necessary if we do not have conditions

(1), (2), and (3). Finally, suppose condition (4c) of Remark 1.7 fails, so for some $Q \in \operatorname{Ass}(T)$, $p \in Q$ but $\operatorname{Ann}_T(p) \subseteq Q$ (where again $M \cap \mathbb{Z} = p\mathbb{Z}$). Then, by [Lee et al. 2001, Theorem 1], T is not the completion of a reduced local subring. □

The following theorem extends the main result, Theorem 3.12, of [Arnosti et al. 2012].

Theorem 2.14. *Let (T, M) be a complete local ring of dimension at least one and let $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_i\}_{i=1}^m)$ be a feasible partition. Then T is the completion of a reduced local subring A such that* $\operatorname{Min}(A) = \{\mathcal{C}_1 \cap A, \dots, \mathcal{C}_m \cap A\}$ *and the formal fiber of A at each $\mathcal{C}_i \cap A$ has countably many maximal elements, which are precisely the elements of $\mathcal{C}_i$, if and only if T has either zero or prime characteristic and at least one of the following is true*:

(1) $\operatorname{char}(T) \neq 0$.

(2) $\operatorname{char}(T) = 0$ *and* $M \cap \mathbb{Z} = \langle 0 \rangle$.

(3) $\operatorname{char}(T) = 0$ *and for all* $P \in \mathcal{C}$, $M \cap \mathbb{Z} \not\subseteq P$.

(4) $\operatorname{char}(T) = 0$, *we have* $M \cap \mathbb{Z} = \langle p \rangle$ *for some prime* p, *and the following three conditions hold*:

  (a) *For each* $P \in \mathcal{C}$ *and for each* $Q \in \operatorname{Ass}(T)$ *with* $Q \subseteq P$, *we have* $p \in Q$ *whenever* $p \in P$.

  (b) *For each subcollection* $\mathcal{C}_i$ *and for any* $P, P' \in \mathcal{C}_i$, *we have* $p \in P$ *if and only if* $p \in P'$.

  (c) *For each* $Q \in \operatorname{Ass}(T)$, *if* $p \in Q$, *then* $\operatorname{Ann}_T(p) \not\subseteq Q$.

Furthermore, when one of these conditions is true, if J is an ideal of T such that $J \not\subseteq P$ for every $P \in \mathcal{C}$, then the natural map $A \to T/J$ is onto.

Proof. The proof is taken almost exactly from the proof of [Arnosti et al. 2012, Theorem 3.12]. Let

$$\Omega = \{u + J \mid u \in T, J \not\subseteq P \text{ for all } P \in \mathcal{C}\}$$

equipped with a well-ordering $<$, such that every element has strictly fewer than $|\Omega|$ predecessors. Note that

$$\bigl|\{J \mid J \text{ is an ideal of } T \text{ with } J \not\subseteq P \text{ for every } P \in \mathcal{C}\}\bigr| \leq |T|.$$

For each $\alpha \in \Omega$, we let $|\alpha| = |\{\beta \in \Omega \mid \beta \leq \alpha\}|$, by abuse of notation.

Let 0 denote the first element of Ω, and let R_0 be the SIP subring of T constructed in Lemma 2.13.

For each $\lambda \in \Omega$ after the first, we define R_λ recursively as follows: assume R_β is defined for all $\beta < \lambda$ such that R_β is an SIP subring, and $|R_\beta| \leq |\beta||R_0|$ for all $\beta < \lambda$. Let $\gamma(\lambda) = u + J$ denote the least upper bound of the set of predecessors of λ. If $\gamma(\lambda) < \lambda$, we use Lemma 2.9 with $R = R_{\gamma(\lambda)}$ to find an SIP subring R_λ such that

(1) $R_{\gamma(\lambda)} \subseteq R_\lambda \subseteq T$;

(2) if $u \in J$, then $J \cap R_\lambda$ contains a non-zerodivisor:

(3) the coset $u + J$ is in the image of the map $R_\lambda \to T/J$;

(4) for all finitely generated ideals I of R_λ, $IT \cap R_\lambda = I$.

In this case,

$$\begin{aligned}|R_\lambda| = |R_{\gamma(\lambda)}| &\le |\gamma(\lambda)|\,|R_0| \\ &\le |\lambda|\,|R_0|.\end{aligned}$$

On the other hand, if $\gamma(\lambda) = \lambda$, we let $R_\lambda = \bigcup_{\beta<\lambda} R_\beta$. We note that for all $\beta < \lambda$,

$$\begin{aligned}|R_\beta| &\le |\beta|\,|R_0| \\ &\le |\lambda|\,|R_0| \\ &< |T|.\end{aligned}$$

The last inequality holds since $|\lambda| < |\Omega| = |T|$ and $|R_0| < |T|$. By Lemma 2.3, it follows that R_λ is an SIP subring of T and $|R_\lambda| \le |\lambda|\,|R_0|$.

Let $A' = \bigcup_{\alpha\in\Omega} R_\alpha$, and define $A = A'_{(A'\cap M)}$. Then A is an IP subring of T.

Note that $M^2 \not\subseteq P$ for every $P \in \mathcal{C}$ so, since every $\alpha = u + M^2 \in \Omega$ has a successor λ (where $\gamma(\lambda) = \alpha$), our construction guarantees that α is in the image of $A \to T/M^2$. Hence this map is onto. Next, let $I = (a_1, \dots, a_n)A$ be a finitely generated ideal of A and $c \in IT \cap A$. Then for some $\delta \in \Omega$, with $\gamma(\delta) < \delta$, $\{c, a_1, \dots, a_n\} \subset R_\delta$. It follows that $c \in IT \cap R_\delta = IR_\delta \subseteq I$. Hence $IT \cap A = I$ for all finitely generated ideals I of A. Since A is a quasilocal subring of T, Proposition 1.1 implies that A is Noetherian and $\widehat{A} = T$.

Now, since T is faithfully flat over A, the ideals $C_i \cap A$ are the minimal prime ideals of A, so that $\operatorname{Min}(A)$ has m elements. Furthermore, by our construction, the natural map $A \to T/J$ is onto for any ideal J such that $J \not\subseteq P$ for all $P \in \mathcal{C}$. By construction, the formal fiber of $C_i \cap A$ has maximal ideals precisely the elements of C_i.

By Lemma 2.13, unless the conditions of the above theorem are satisfied, no IP subring of T exists whose completion is T. As A must be an IP subring in order for us to control its minimal prime ideals, if the conditions do not hold, then neither does this theorem. □

Arnosti et al. [2012] note that, for their ring A, not only are the formal fibers of the minimal prime ideals of A known, but so are the formal fibers at *all* prime ideals of A. The same holds for our A of Theorem 2.14. Specifically, let A be as in Theorem 2.14, and suppose that P is any nonminimal prime ideal of A. Then by the theorem, $A \to T/PT$ is an onto map. As T is the completion of A, T is a faithfully flat extension of A and so $PT \cap A = P$. It follows that $A/P \cong T/PT$, and so the only element in the formal fiber of A at P is PT.

To demonstrate that this result is not trivial, we examine a complete ring T for which this theorem, but not previous work, allows us to find the desired subring A.

Example 2.15. Let

$$T = \frac{\widehat{\mathbb{Z}_{\langle 5\rangle}}[[x, y, z]]}{\langle 5xy\rangle},$$

and let $\mathcal{P} = (\mathcal{C}, \{\mathcal{C}_1, \mathcal{C}_2\})$, where $\mathcal{C}_1 = \{\langle 5\rangle\}$ and $\mathcal{C}_2 = \{\langle x, y\rangle\}$. Then T is a complete local ring with dimension at least one and characteristic zero, and $\mathcal{P}$ is a feasible partition. Furthermore, T satisfies every subcondition of condition (4) of Theorem 2.14. Therefore, by that theorem, there exists a reduced local subring A of T having the desired properties. As T does not contain the rationals and is not of one of the previously characterized categories of complete rings, Theorem 2.14 is necessary to establish the existence of A.

References

[Aiello et al. 2015] D. Aiello, S. Loepp, and P. Vu, "Formal fibers with countably many maximal elements", *Rocky Mountain J. Math.* **45**:2 (2015), 371–388. MR 3356620

[Arnosti et al. 2012] N. Arnosti, R. Karpman, C. Leverson, J. Levinson, and S. Loepp, "Semi-local formal fibers of minimal prime ideals of excellent reduced local rings", *J. Commut. Algebra* **4**:1 (2012), 29–56. MR 2913526 Zbl 1239.13034

[Charters and Loepp 2004] P. Charters and S. Loepp, "Semilocal generic formal fibers", *J. Algebra* **278**:1 (2004), 370–382. MR 2005e:13008 Zbl 1093.13023

[Heitmann 1993] R. C. Heitmann, "Characterization of completions of unique factorization domains", *Trans. Amer. Math. Soc.* **337**:1 (1993), 379–387. MR 93g:13006 Zbl 0792.13011

[Heitmann 1994] R. C. Heitmann, "Completions of local rings with an isolated singularity", *J. Algebra* **163**:2 (1994), 538–567. MR 95f:13032 Zbl 0798.13009

[Lech 1986] C. Lech, "A method for constructing bad Noetherian local rings", pp. 241–247 in *Algebra, algebraic topology and their interactions* (Stockholm, 1983), Lecture Notes in Math. **1183**, Springer, Berlin, 1986. MR 87m:13010a Zbl 0589.13006

[Lee et al. 2001] D. Lee, L. Leer, S. Pilch, and Y. Yasufuku, "Characterization of completions of reduced local rings", *Proc. Amer. Math. Soc.* **129**:11 (2001), 3193–3200. MR 2002e:13021 Zbl 0971.13022

[Loepp 2003] S. Loepp, "Characterization of completions of excellent domains of characteristic zero", *J. Algebra* **265**:1 (2003), 221–228. MR 2004e:13033 Zbl 1083.13508

Received: 2014-08-21 Revised: 2014-10-27 Accepted: 2015-01-09

sloepp@williams.edu

Department of Mathematics and Statistics, Williams College,
18 Hoxsey Street, Bronfman Science Center,
Williamstown, MA 01267, United States

byronperpetua@gmail.com

Department of Mathematics and Statistics, Williams College,
c/o Susan Loepp, Bronfman Science Center,
18 Hoxsey Street, Williamstown,MA 01267, United States

INVOLVE 9:1 (2016)
dx.doi.org/10.2140/involve.2016.9.119

Global regularity of chemotaxis equations with advection

Saad Khan, Jay Johnson, Elliot Cartee and Yao Yao

(Communicated by Behrouz Emamizadeh)

We study the Patlak–Keller–Segel (PKS) equations in 2D that describe chemotaxis with an additional advection term. We show that solutions are globally regular for smooth initial data with subcritical mass as long as the flow has nonpositive divergence. For initial data with supercritical mass, numerical simulations suggest that blow-up might be prevented by imposing some strong incompressible advection term.

1. Introduction

In this paper, we study the effect of adding an advective flow term in the Patlak–Keller–Segel (PKS) equations that model chemotaxis. Chemotaxis is the means by which small organisms such as bacteria and somatic cells direct their movements towards or against the gradient of some chemical concentration.

For such organisms, which often swim in low Reynold's numbers settings, movement is a huge challenge for many reasons. The organism's small size relative to the fluid it inhabits means that it has to overcome the effects of diffusion. Furthermore, organisms at this level typically do not have any sort of neural system, and thus cannot process the large amounts of information that would be needed to purposefully move from one place to another. The available information is also limited to whatever chemicals can bind to ligands on the cell membranes. For these reasons, chemotaxis is a very simple mechanism. The organism simply measures the gradient of a relevant chemical concentration (such as glucose, cAMP, bicoid, etc.) and moves in that direction. Its motion can be modeled by combining the diffusive and chemotactic components of its trajectory into an evolution equation

$$\rho_t = \Delta\rho + \nabla\cdot(\rho\nabla c), \tag{1-1}$$

MSC2010: 35A01.

Keywords: chemotaxis, partial differential equations, PDE, symmetric decreasing rearrangement, advection, Keller–Segel.

where ρ is the density of the organism and c represents the concentration of the chemical. The PKS equation arises by assuming that the chemical is produced by the organism itself. In this case, the chemical diffuses like

$$\epsilon c_t - \Delta c = \rho.$$

In the physically relevant regime, the diffusion of the chemical is much faster than the diffusion of the organism. Hence the $\epsilon \to 0$ limit is often taken, and the equation above becomes

$$-\Delta c = \rho. \tag{1-2}$$

The PKS equation [Patlak 1953; Keller and Segel 1970; 1971] is then obtained by combining (1-1) and (1-2):

$$\rho_t = \Delta\rho + \nabla\cdot(\rho\nabla(\mathcal{N} * \rho)), \tag{1-3}$$

where $*$ denotes convolution and $\mathcal{N}(x) = 1/(2\pi)\log|x|$ is the Newtonian potential in 2D.

This equation has been studied extensively (see [Horstmann 2003] and the references therein). In particular, it is well known that solutions with different mass sizes exhibit different behaviors: For nonnegative initial data with L^1-norm greater than 8π (i.e., solutions with *supercritical mass*), solutions blow up in finite time [Patlak 1953; Perthame 2007]. On the other hand, if the L^1-norm is less than 8π (i.e., *subcritical mass*), the diffusive term dominates the dynamics of the equation, where the solutions are globally regular, and indeed the L^∞-norm goes to 0 as $t \to \infty$.

In this paper, we incorporate an extra advection term into the PKS equation, which then becomes

$$\rho_t = \Delta\rho + \nabla\cdot(\rho\nabla(\mathcal{N} * \rho)) - \nabla\cdot(\rho\vec{u}). \tag{1-4}$$

The motivation of this extra term is that the fluid medium the organisms inhabit may have its own current, which we denote by $\vec{u}$. Throughout this paper, we assume that the underlying flow $\vec{u}(x, y)$ is an a priori given velocity field and does not depend on ρ. We are particularly interested in the case where the flow $\vec{u}$ is incompressible, since this is the physically relevant case. The goal of this paper is to investigate the effect of the advection term on the behavior of the solution. In particular, we would like to answer the following questions:

(1) For smooth initial data with subcritical mass (i.e., $\|\rho_0\|_{L^1} < 8\pi$) and any incompressible flow $\vec{u}$, does the solution to (1-4) always have global regularity?

(2) For initial data with supercritical mass (i.e., $\|\rho_0\|_{L^1} > 8\pi$), is it possible to prevent a finite-time blow-up by imposing a strong incompressible advection term in (1-4)?

In Section 2, we give a positive answer to the first question. More precisely, we prove that as long as the velocity field $\vec{u}$ has nonpositive divergence (which includes incompressible velocity fields as a special case), any solution to (1-4) with subcritical mass remains regular for all time. The main ingredients of our proof are a comparison principle and symmetric decreasing rearrangements.

As an attempt to address the second question, we perform some numerical simulations in Section 3. Using a stochastic particle simulation based on [Haškovec and Schmeiser 2009], we experiment with a variety of flows on solutions with supercritical mass. Our numerical results suggest that the answer might be positive. Namely, some incompressible flows, such as the shear flow and the strain flow, seem to be good candidates for preventing solutions with supercritical mass from blowing up, as long as the flow strength is sufficiently large. A rigorous proof of this phenomenon remains a very interesting open question.

2. Global regularity for solutions with subcritical mass

In this section, we consider velocity fields $\vec{u}$ with nonpositive divergence, and our goal is to prove global regularity for (1-4) with subcritical initial data.

Radially symmetric case. We first deal with the PKS equation without advection (1-3), and we prove some results for radially symmetric solutions. Although these results are known (e.g., see [Perthame 2007]), we sketch their proofs below for the sake of completeness, since some of these techniques will be useful for the PKS equation with advection as well.

For convenience, we define the mass function, which will be used throughout this section, as follows.

Definition 2.1. For a function $f \in L^1(\mathbb{R}^2)$, we say that $M_f(r) := \int_{B_0(r)} f\,dx$ is the *mass function* associated with f.

Making use of the mass function, we identify all the radially symmetric steady state solutions for the PKS equation (without advection) in the next theorem.

Theorem 2.2. *Consider the PKS equation* (*without advection*)

$$\rho_t = \Delta\rho + \nabla\cdot(\rho\nabla(\mathcal{N} * \rho)). \tag{2-1}$$

All nonzero radially symmetric steady state solutions of (2-1) *are of the form*

$$\rho_\lambda(r) = \frac{8\lambda}{(\lambda + r^2)^2} \tag{2-2}$$

for some $\lambda > 0$.

Proof. When $\rho(x,t)$ satisfies the PKS equation, one can check by direct computation that the mass function $M_\rho(r,t)$ (which we denote by $M(r,t)$ for simplicity) solves the equation

$$M_t = M_{rr} - \frac{1}{r}M_r + \frac{MM_r}{2\pi r}, \tag{2-3}$$

and hence the mass function of a radial steady state solution satisfies

$$M_{rr} - \frac{1}{r} M_r + \frac{M M_r}{2\pi r} = 0. \tag{2-4}$$

Although this ODE is nonlinear, we can multiply it by r, then rewrite $r M_{rr}$ as $(r M_r)_r - M_r$ and $M M_r$ as $\frac{1}{2}(M^2)_r$ to obtain

$$(r M_r)_r - 2M_r + \frac{1}{4\pi}(M^2)_r = 0, \tag{2-5}$$

which yields

$$r M_r - 2M + \frac{M^2}{4\pi} = c \tag{2-6}$$

for some constant c. Since $M(r)$ is the mass contained in a disk of radius r, we have $M(0) = 0$ by definition, which implies that $c = 0$. Hence the equation above becomes separable and can be written as

$$\frac{M_r}{M(2 - M/(4\pi))} = \frac{1}{r}, \tag{2-7}$$

and by integrating it, we obtain the following family of steady state solutions labeled by a parameter $\lambda > 0$:

$$M_\lambda(r) = \frac{8\pi r^2}{\lambda + r^2}. \tag{2-8}$$

Lastly, observe that $\rho_\lambda(r) = M'_\lambda(r)/(2\pi r)$, which gives (2-2). □

Now we want to show that all radially symmetric solutions with subcritical mass are controlled by (2-2) for some λ. Even though there is no comparison principle for ρ in (2-1), a comparison principle does hold for the mass function M for (2-3), which we describe below.

Theorem 2.3. *Assume $\rho_1(x, t)$ and $\rho_2(x, t)$ are two radially symmetric solutions to* (2-1), *where they satisfy $M_2(r, 0) \geq M_1(r, 0)$ for all $r \geq 0$.* (*Here M_1 and M_2 are the mass functions for ρ_1 and ρ_2 respectively.*) *Then for every t, we have $M_2(r, t) \geq M_1(r, t)$.*

The proof can be found in [Kim and Yao 2012] and will be omitted here.

Making use of this comparison principle, we can now show global regularity for radially decreasing initial data. Here we say ρ_0 is *radially decreasing* if it is radially symmetric, and $\rho_0(r)$ is nonincreasing in r.

Corollary 2.4. *Let $\rho_0(r)$ be a nonnegative smooth radially decreasing function, with $\|\rho_0\|_{L^1} < 8\pi$. Let $\rho(x, t)$ be the solution to* (2-1) *with initial data ρ_0. Then $\rho(x, t)$ is globally regular and bounded.*

Proof. We find a sufficiently small $\lambda > 0$ such that $M_\lambda(r) > M_{\rho_0}(r)$ for all $r \geq 0$. Then, since $M_\lambda(r)$ is a steady state solution, the comparison principle in Theorem 2.3 ensures that $M_\rho(r,t) \leq M_\lambda(r)$ for all time during the existence of ρ, which implies that $\rho(0,t) \leq \rho_\lambda(0)$ for all time. Note that the radially decreasing property of ρ is preserved by the PKS equation (see Theorem 4.2 of [Kim and Yao 2012]), so its maximum occurs at the origin for all time. Combining these two facts, we have $\|\rho(\,\cdot\,,t)\|_{L^\infty} \leq \rho_\lambda(0)$ for all time during the existence of ρ. Once we have this global L^∞-bound, one can proceed as in [Kiselev and Ryzhik 2012] to show that the solution is indeed smooth for all time. □

General, advective case with nonpositive divergence. Using the previous results, we aim to study the regularity properties of the advective chemotaxis equation (1-4) with a nonpositive divergence flow, where the initial data are not necessarily radially symmetric. To do this we will utilize symmetric decreasing rearrangements, which can map arbitrary measurable functions to symmetric decreasing functions for which the above results hold. Using this transformation and some inequalities, we can show that the symmetric case is a "supersolution" of the general case in some sense.

The symmetric decreasing rearrangement of a function is defined as follows.

Definition 2.5. For a measurable set $\Omega \subset \mathbb{R}^2$, we define its symmetric rearrangement Ω^* as $\Omega^* := B(0,r)$, where r is chosen such that $|B(0,r)| = |\Omega|$.

For a nonnegative $f \in L^1(\mathbb{R}^2)$, its *symmetric decreasing rearrangement* f^* is given by

$$f^*(x) := \int_0^\infty \chi_{\{f>t\}^*}(x)\,dt, \tag{2-9}$$

where χ denotes the characteristic function.

Below we list a couple of properties on the symmetric decreasing rearrangement.

Lemma 2.6. *Let $\rho \in L^1(\mathbb{R}^2)$ be nonnegative and let $\Omega = \{\rho > s\}$ for some $s \geq 0$. Then the following hold*:

(1) $\int_\Omega \rho\,dx = \int_{\Omega^*} \rho^*\,dx$.

(2) $\int_\Omega \Delta\rho\,dx \leq \int_{\Omega^*} \Delta\rho^*\,dx$.

(3) $\int_\Omega f\,dx \leq \int_{\Omega^*} f^*\,dx$ *for any nonnegative* $f \in L^1(\mathbb{R}^2)$.

(4) $\int_\Omega \nabla\cdot(\rho\nabla(\mathcal{N} * f))\,dx = s\int_\Omega f\,dx$ *for any nonnegative* $f \in L^1(\mathbb{R}^2)$.

Proof. The proofs for (1)–(3) can be found in [Burchard 2009]. Next we will prove (4). Apply the divergence theorem to get

$$\int_\Omega \nabla\cdot(\rho\nabla(\mathcal{N} * f))\,dx = \int_{\partial\Omega} \rho\nabla(\mathcal{N} * f)\cdot\hat{n}\,dx, \tag{2-10}$$

where $\hat{n}$ is the unit normal vector of Ω. Observe that $\partial\Omega = \{\rho = s\}$, which, when applied to the right-hand side of (2-10), gives

$$\int_{\partial\Omega} \rho\nabla(\mathcal{N} * f)\cdot\hat{n}\,dx = s\int_{\Omega}\nabla\cdot\nabla(\mathcal{N} * f)\,dx = s\int_{\Omega} f\,dx, \tag{2-11}$$

where in the last equality we used the fact that $\mathcal{N} * f$ inverts the Laplacian. We thus obtain (4) by combining the above two equations. □

Definition 2.7. Given any two functions $f, g \in L^1(\mathbb{R}^2)$, we say $f \prec g$ if $M_f(r) \le M_g(r)$ for all $r \ge 0$.

With these tools in hand, we can show that solutions of (1-4) with subcritical mass are globally regular by showing that $\rho^*(\,\cdot\,,t) \prec \rho'$ for all time during the existence of ρ, where ρ' is some radially symmetric steady state in (2-2). This implies that ρ has an L^∞-bound that is uniform in time, which finally gives the global regularity of ρ. To do so, we proceed in stages, and construct a sequence converging to the appropriate solution. The method of this proof follows very closely the approach used in [Kim and Yao 2012].

Theorem 2.8. *Assume* $\nabla\cdot\vec{u} \le 0$. *Let* ρ *and* ρ' *solve*

$$\rho_t = \Delta\rho + \nabla\cdot(\rho\nabla(\mathcal{N} * f)) + \nabla\cdot(\vec{u}\rho), \tag{2-12}$$

$$\rho'_t = \Delta\rho' + \nabla\cdot(\rho'\nabla(\mathcal{N} * f')), \tag{2-13}$$

respectively, with $\rho(\,\cdot\,,0) = \rho_0$, $\rho'(\,\cdot\,,0) = \rho_0^*$, *and* $f \prec f'$. *Then we have* $\rho^*(\,\cdot\,,t) \prec \rho'(\,\cdot\,,t)$ *for all* $t \ge 0$.

Proof. We prove this theorem by first discretely approximating ρ_t as $(\rho_{n+1} - \rho_n)/h$. Set $g = \rho_n$. We will show that $\rho^*_{n+1} \prec \rho'_{n+1}$ for every positive h. Since the Laplacian is an m-accretive operator [Barbu 2010], we can then let h go to 0 and recover Theorem 2.8. Therefore it suffices to prove the following lemma for each discrete time step. □

Lemma 2.9. *Assume* $\nabla\cdot\vec{u} \le 0$. *Let* ρ *and* ρ' *solve*

$$\rho = h\Delta\rho + h\nabla\cdot(\rho\nabla(\mathcal{N} * f)) + \nabla\cdot(\rho\vec{u}) + g, \tag{2-14}$$

$$\rho' = h\Delta\rho' + h\nabla\cdot(\rho'\nabla(\mathcal{N} * f')) + g', \tag{2-15}$$

respectively, and assume that ρ', f' *and* g' *are all radially symmetric. If* $f^* \prec f'$ *and* $g^* \prec g'$, *then we have* $\rho^* \prec \rho'$.

Proof. For any $r > 0$, one can choose $s > 0$ such that $|\{\rho > s\}| = |B(0,r)|$. Thus, let $\Omega := \{\rho > s\}$ and it follows that $\Omega^* = B(0,r)$. Integrate (2-15) over $B(0,r)$, which gives

$$M_{\rho'}(r) = hM_{\Delta\rho'}(r) + h\rho'(r)M_{f'}(r) + M_{g'}(r). \tag{2-16}$$

We also integrate (2-14) over $\Omega = \{\rho > s\}$. Using Lemma 2.6 and Definition 2.1, we have

$$\begin{aligned} M_{\rho^*}(r) &= \int_\Omega \rho(x)\,dx \quad \text{(by Lemma 2.6(a))} \\ &= h\int_\Omega \Delta\rho\,dx + h\nabla\cdot(\rho\nabla(\mathcal{N} * f)) + \int_\Omega \nabla\cdot(\rho\vec{u})\,dx + \int_\Omega g\,dx \\ &\le hM_{\Delta\rho^*}(r) + h\rho(r)M_f^*(r) + M_g^*(r) \quad \text{(by Lemma 2.6(2)–(4)).} \end{aligned} \tag{2-17}$$

In the last inequality, we also used $\int_\Omega \nabla\cdot(\rho\vec{u})\,dx = \int_{\partial\Omega} \hat{n}\cdot(\rho\vec{u})\,d\sigma = s\int_\Omega \nabla\cdot\vec{u}\,dx \le 0$, where our assumption $\nabla\cdot\vec{u} \le 0$ is applied.

In order to show that $\rho^* \prec \rho'$, first notice that by sending $r \to \infty$ in (2-16) and (2-17) and using the assumption that $g^* \prec g'$, we have $\lim_{r\to\infty} M_{\rho^*}(r) - M_{\rho'}(r) \le 0$. Therefore, if $\rho^* \prec \rho'$ does not hold, there must exist some finite $r_0 > 0$, such that $M_{\rho^*}(r) - M_{\rho'}(r)$ attains a positive maximum at r_0. Note that we have $\rho'(r_0) = \rho^*(r_0)$ since $\partial_r(M_{\rho^*} - M_{\rho'})(r_0) = 0$.

Subtracting (2-17) and (2-16) at r_0 gives

$$\begin{aligned} M_{\rho^*-\rho'}(r_0) &\le hM_{\Delta(\rho^*-\rho')}(r_0) + h\rho'(r_0)M_{f^*-f'}(r_0) + M_{g^*-g'}(r_0) \\ &\le hM_{\Delta(\rho^*-\rho')}(r_0), \end{aligned} \tag{2-18}$$

where in the last step we use the assumptions that $f^* \prec f'$ and $g^* \prec g'$. Since ρ^* and ρ' are radially symmetric, we can simplify $M_{\Delta(\rho^*-\rho')}$ as

$$M_{\Delta(\rho^*-\rho')} = \partial_{rr}M_{\rho^*-\rho'} - \frac{\partial_r M_{\rho^*-\rho'}}{r}. \tag{2-19}$$

Since $M_{\rho^*-\rho'}$ achieves a maximum at r_0, it follows from the above expression that $M_{\Delta(\rho^*-\rho')}(r_0) \le 0$, and combining it with (2-18), we have $M_{\rho^*-\rho'}(r_0) \le 0$, leading to a contradiction. This yields $\rho^* \prec \rho'$. □

Once we have Theorem 2.8, our global regularity result follows from the same iteration argument as in [Kim and Yao 2012], which we sketch below.

Theorem 2.10. *Let ρ solve* (1-4) *with smooth, subcritical initial data ρ_0. Then ρ is globally regular and bounded for all time.*

Proof. Let $\rho_1(\,\cdot\,,t) := \rho^*(\,\cdot\,,t)$. For any $n \ge 1$, we iteratively define $\rho_{n+1}(\,\cdot\,,t)$ by

$$\partial_t\rho_{n+1} = \Delta\rho_{n+1} + \nabla\cdot(\rho_{n+1}\nabla(\mathcal{N} * \rho_n)), \tag{2-20}$$

with initial data ρ_0^*.

By Lemma 2.9, $\rho_1 \prec \rho_2$. Once we have this, one can then iteratively apply Lemma 2.9 (with zero velocity field) to obtain $\rho_n \prec \rho_{n+1}$ for all $n \ge 1$. Thus, we have an increasing sequence of mass functions. Finally, along a subsequence,

$\rho_n \to \bar{\rho}$ (hence $\rho^* = \rho_1 \prec \bar{\rho}$ too), where $\bar{\rho}$ solves

$$\partial_t \bar{\rho} = \Delta \bar{\rho} + \nabla \cdot (\bar{\rho} \nabla (\mathcal{N} * \bar{\rho})). \tag{2-21}$$

Details for this can be found in [Kim and Yao 2012]. Since $\bar{\rho}$ solves the PKS equation without advection, Corollary 2.4 gives that $\bar{\rho}(\cdot, t)$ is bounded and regular globally in time. Since $\rho(\cdot, t) \prec \bar{\rho}(\cdot, t)$ for all $t \geq 0$, the same holds for $\rho(\cdot, t)$ as desired. □

3. Numerical study of solutions with supercritical mass

Numerical methods. We follow the Euler–Maruyama stochastic particle approximation to the 2D Keller–Segel equation that is explained in detail in [Haškovec and Schmeiser 2009], henceforth abbreviated [HS]. The key benefit of this approach, as opposed to using finite element or finite volume, is that the system can still be analyzed after a singularity blow-up forms. It is also relatively simple to code, which was our main concern for this investigation. The only significant difference between our method and the one described in detail in [HS] is the addition of a deterministic advection step at every time interval. Below we lay out the basic aspects of the numerical method from [HS] and the added advection term.

For all of the simulations, we use a system of $N = 1000$ particles located at $x_1, \ldots, x_N$ with mass sizes $M_1, \ldots, M_N$ respectively. The "light" particles, with mass less than or equal to 8π, approximate the smooth part of the solution, and the "heavy" particles, with mass greater than 8π, approximate delta functions when the solution has blown up.

Step 1: Advection. To model the extra advection term at every time step, the particles move according to advection, which gives $\mathrm{d}x_n = \vec{u}(x_n)\, dt$.

Step 2: Aggregation. Using the stochastic particle approximation from [HS], the nonlocal, chemotactic interaction term of the PDE is

$$\mathrm{d}x_n = -\frac{1}{2\pi} \sum_{m \neq n} M_m \frac{x_n - x_m}{|x_n - x_m|^2}\, dt.$$

Then our set of particles undergoes the processes of collision and splitting. This part is the same as in [HS], which is explained further in the next subsection for the sake of completeness.

Step 3: Diffusion. After the aggregation step, each light particle undergoes a random-walk step, giving $\mathrm{d}x_n = \sqrt{2\, dt}\, \mathcal{N}_{(0,1)}$, where $\mathcal{N}_{(0,1)}$ denotes the Gaussian distribution with mean 0 and variance 1.

Particle collisions and splitting. In our method, particle collisions and splitting are handled in the same way as in [HS], which we describe below for the sake of completeness. In the aggregation step, it is easy to see that as the distance between two particles goes to zero, the interaction kernel grows to infinity. Hence we allow two sufficiently close particles $\{x_1, x_2\}$ to collide and form one new particle, with the new mass $M' = M_1 + M_2$. The criterion, from [HS], for two particles to collide during a time interval $(0, \Delta T)$ is given by

$$\|x_1 - x_2\|^2 \leq \frac{M_1 + M_2}{\pi} \Delta T.$$

At each discrete time step, this inequality is evaluated for each particle pair and the particles satisfying this condition are fused. This inevitably leads to single particles accumulating more and more mass as they pull in other particles towards them.

The main issue with fusing particles is that as particles collide, the effective grid spacing coarsens, as there are fewer and fewer particles in the domain. To compensate for the particles lost to collisions, particles are randomly split at each time step so that the total number of particles remains constant. This helps maintain a proper discretization of the space and helps more accurately model the desired effects. The exact method of particle collisions and splitting is explained in detail in [HS]. The random-walk step is evaluated after the splitting occurs so that when two particles are split, they remain at the same position but are then redistributed by Brownian motion.

Numerical results. We run numerical simulations for a variety of different flows $\vec{u}$ to find whether blow-up could be delayed or prevented with the presence of a flow. The flows we investigate are

$$\begin{aligned} \text{shear flow:} \quad & \vec{u}(x, y) = (e^{-y^2}, 0), && \text{(3-1)} \\ \text{strain flow:} \quad & \vec{u}(x, y) = (-x, y), && \text{(3-2)} \\ \text{diverging flow:} \quad & \vec{u}(r, \theta) \ = (1/r, \theta), && \text{(3-3)} \end{aligned}$$

which are compared against the no-flow case. Note that the shear flow and strain flow are both incompressible. Although the diverging flow is not incompressible, we also test it and it turns out that it is effective at preventing blow-up with a relatively weak flow strength. Velocity fields for these flows are illustrated in Figure 1.

We use 1000 particles in our simulations, which are initially randomly distributed in a disk $B(0, \sqrt{0.5})$. An initial mass $M = 16\pi$ (twice the critical mass) is distributed evenly across all the particles. We choose the time step as $\Delta t = 5 \cdot 10^{-5}$ and run simulations for 2000 time steps until $t = 0.1$. With no flow, our simulation shows that blow-up occurs at $t = 0.0325$, as shown in Figure 2.

In order to see how different flow strengths affect the result, we multiply the flow $\vec{u}$ by a constant C, so that the stochastic particle approximation for the advection

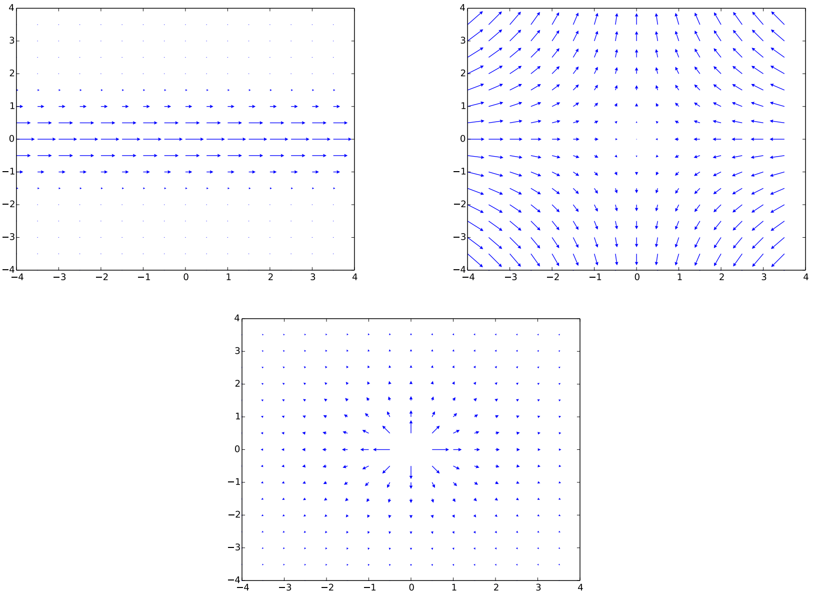


Figure 1. Illustration of the three types of flows tested in our numerical study. Top left: shear flow (3-1). Top right: strain flow (3-2). Bottom: diverging flow (3-3).

term is

$$\mathrm{d}x_n = C\vec{u}(x_n)\,dt.$$

Shear flow is tested with a flow strength, C, of 100 and 1000, strain flow with strengths of 10 and 100, and diverging flow with strengths of 1 and 10. From now on, we will refer to the smaller of the two flow strengths as "weak flow" and the larger of the two as "strong flow" for each flow configuration. Our numerical results

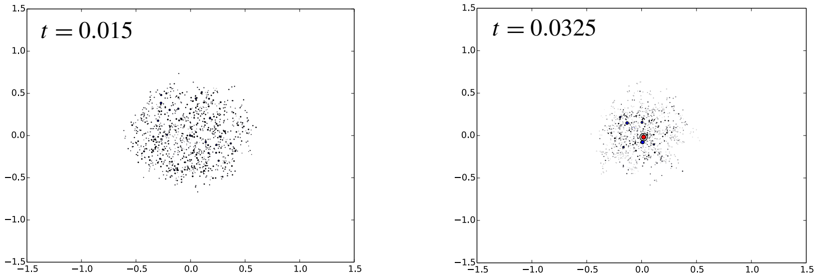


Figure 2. Numerical simulation for the PKS equation with no advection term at different times. Blow-up occurs at $t = 0.0325$. The red dot in the second picture indicates the location of blow-up, where we have a mass concentration of more than 8π.

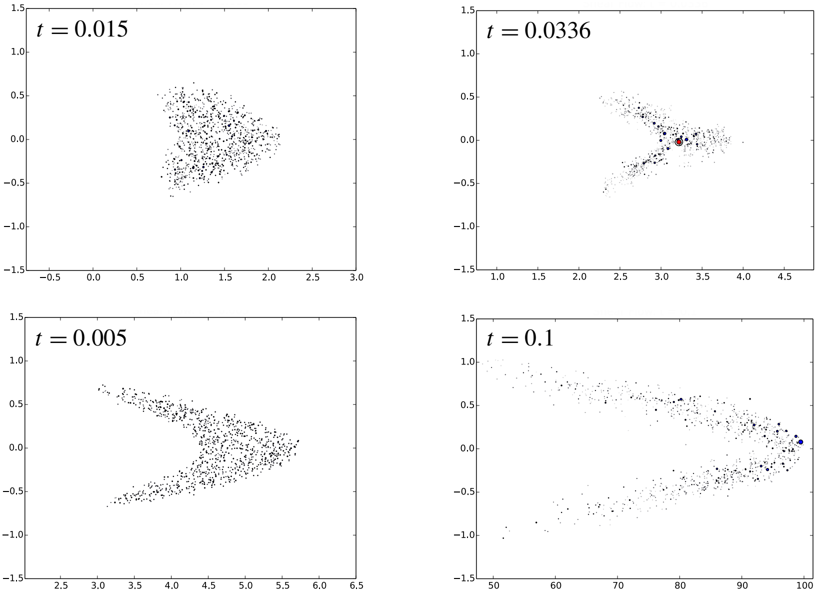


Figure 3. Shear flow (3-1) with strengths $C = 100$ (top row) and 1000 (bottom row). For $C = 100$, blow-up occurs at $t = 0.0336$ (the red dot indicates the location of blow-up). For $C = 1000$, blow-up does not occur before $t = 0.1$.

are shown in Figures 3–5. The results suggest that all of these three flows seem to be able to prevent blow-up when the flow strength is chosen to be sufficiently large.

Remarks on possible future improvements. We now point out some possible improvements that can be made with our current numerical scheme. First, computation time is an issue when many different flows need to be tested. Note that the most time-consuming step is the aggregation step, where for N particles, one has to perform on the order of N^2 calculations to compute their pairwise interactions. Haškovec and Schmeiser [2009] comment on one potential heuristic for speeding up the chemotaxis calculation, where the program, while calculating the step for a given particle, averages the masses of many distant, but close together particles, and uses the essential “center of mass”, instead of the contribution from each individual particle.

Currently, our numerical scheme is not sensitive enough to show a difference between an initial mass that is only slightly above or below $M = 8\pi$. It also is not able to show that for flows $-d/r^{\alpha}$, $\alpha < 1$, with subcritical initial data, the flows will not blow up. These insensitivities may be solved with trying smaller time steps Δt, but it is likely that another numerical method is needed. To help support these claims, it might be better to implement a finite element or finite volume approach, but this is outside the scope of this paper.

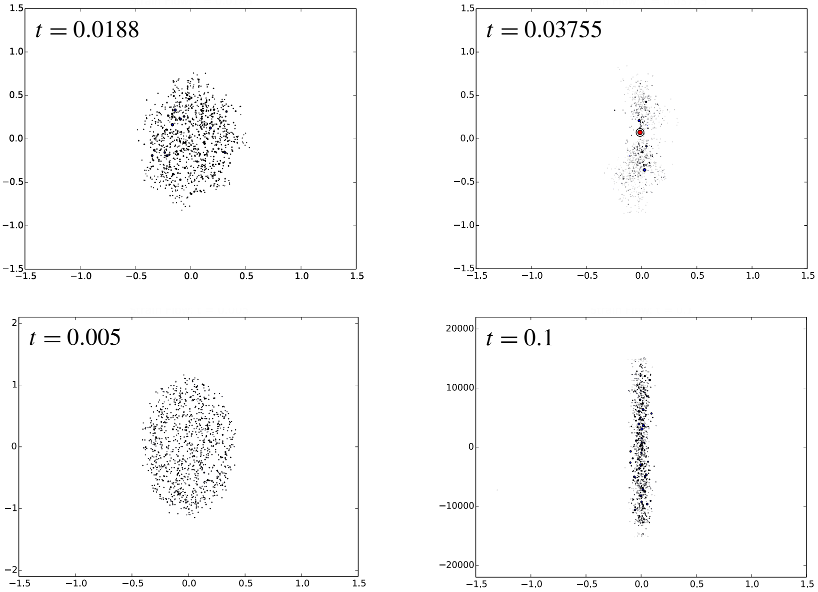


Figure 4. Strain flow (3-2) with strengths $C = 10$ (top row) and 100 (bottom row). For $C = 10$, blow-up occurs at $t = 0.03755$. For $C = 100$, blow-up does not occur before $t = 0.1$.

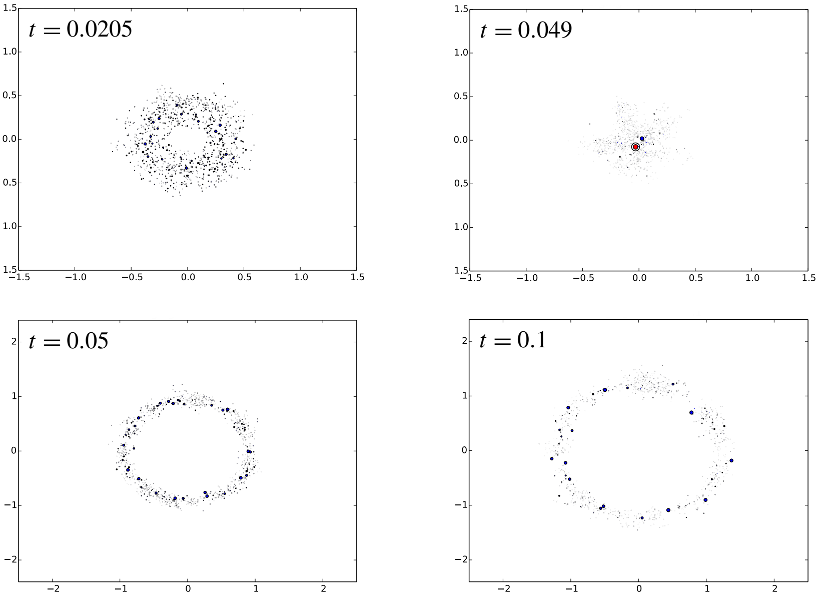


Figure 5. Diverging flow (3-3) with strengths $C = 1$ (top row) and 10 (bottom row). For $C = 1$, blow-up occurs at $t = 0.0490$. For $C = 100$, blow-up does not occur before $t = 0.1$.

Acknowledgements

This research was conducted at the University of Wisconsin–Madison during the summer of 2013 as a part of its Analysis and Differential Equations REU. We would like to thank Professor Alexander Kiselev for helping direct our research and teaching us about the chemotaxis equations. We would also like to thank Tam Do for helping us with the proofs of the theorems and providing guidance.

References

[Barbu 2010] V. Barbu, *Nonlinear differential equations of monotone types in Banach spaces*, Springer, New York, 2010. MR 2011d:34001 Zbl 1197.35002

[Burchard 2009] A. Burchard, "A short course on rearrangement inequalities", Lecture notes, 2009, available at http://www.math.utoronto.ca/almut/rearrange.pdf.

[Haškovec and Schmeiser 2009] J. Haškovec and C. Schmeiser, "Stochastic particle approximation for measure valued solutions of the 2D Keller–Segel system", *J. Stat. Phys.* **135**:1 (2009), 133–151. MR 2010f:92014 Zbl 1173.82021

[Horstmann 2003] D. Horstmann, "From 1970 until present: the Keller–Segel model in chemotaxis and its consequences, I", *Jahresber. Deutsch. Math.-Verein.* **105**:3 (2003), 103–165. MR 2005f:35163 Zbl 1071.35001

[Keller and Segel 1970] E. F. Keller and L. A. Segel, "Initiation of slime mold aggregation viewed as an instability", *J. Theor. Biol.* **26**:3 (1970), 399–415. Zbl 1170.92306

[Keller and Segel 1971] E. F. Keller and L. A. Segel, "Model for chemotaxis", *J. Theor. Biol.* **30**:2 (1971), 225–234. Zbl 1170.92307

[Kim and Yao 2012] I. Kim and Y. Yao, "The Patlak–Keller–Segel model and its variations: properties of solutions via maximum principle", *SIAM J. Math. Anal.* **44**:2 (2012), 568–602. MR 2914242 Zbl 1261.35080

[Kiselev and Ryzhik 2012] A. Kiselev and L. Ryzhik, "Biomixing by chemotaxis and enhancement of biological reactions", *Comm. Partial Differential Equations* **37**:2 (2012), 298–318. MR 2876833 Zbl 1236.35190

[Patlak 1953] C. S. Patlak, "Random walk with persistence and external bias", *Bull. Math. Biophys.* **15** (1953), 311–338. MR 18,424f Zbl 1296.82044

[Perthame 2007] B. Perthame, *Transport equations in biology*, Birkhäuser, Basel, 2007. MR 2007j:35004 Zbl 1185.92006

Received: 2014-08-25 Revised: 2015-01-08 Accepted: 2015-01-09

smk508@nyu.edu
Department of Mathematics, New York University,
New York, NY 10012-1185, United States

johnson.jayrichard@utexas.edu
Department of Mechanical Engineering,
University of Texas at Austin, Austin, TX 78712, United States

evc34@cornell.edu
Department of Mathematics, Cornell University,
Ithaca, NY 14850, United States

yaoyao@math.wisc.edu
Department of Mathematics, University of Wisconsin-Madison,
Madison, WI 53703, United States

INVOLVE 9:1 (2016)
dx.doi.org/10.2140/involve.2016.9.133

On the ribbon graphs of links in real projective space

Iain Moffatt and Johanna Strömberg

(Communicated by Józef H. Przytycki)

Every link diagram can be represented as a signed ribbon graph. However, different link diagrams can be represented by the same ribbon graphs. We determine how checkerboard colourable diagrams of links in real projective space, and virtual link diagrams, that are represented by the same ribbon graphs are related to each other. We also find moves that relate the diagrams of links in real projective space that give rise to (all-A) ribbon graphs with exactly one vertex.

1. Introduction and overview

It is well known that a classical link diagram can be represented by a unique signed plane graph, called its Tait graph (see, for example, the surveys [Bollobás 1998; Ellis-Monaghan and Moffatt 2013; Welsh 1993]). This construction provides a seminal connection between the areas of graph theory and knot theory, and has found impressive applications, such as in proofs of the Tait conjectures [Murasugi 1987; Thistlethwaite 1987]. Tait graphs can also be constructed for checkerboard colourable link diagrams on other surfaces, in which case the resulting graph is embedded on the surface. However, as this construction requires checkerboard colourability, Tait graphs cannot be constructed for arbitrary link diagrams on a surface, or arbitrary virtual link diagrams. Recently, Dasbach, Futer, Kalfagianni, Lin, and Stoltzfus [Dasbach et al. 2008] extended the idea of a Tait graph by associating a set of signed ribbon graphs to a link diagram (see also [Turaev 1987]). Chmutov and Voltz [2008] extended this construction, giving a way to describe an arbitrary virtual link diagram as a signed ribbon graph. These constructions extend to graphs in other surfaces. The ribbon graphs of link diagrams have found numerous applications, and we refer the reader to the surveys [Champanerkar and Kofman 2014; Ellis-Monaghan and Moffatt 2013] for details.

MSC2010: primary 57M15; secondary 05C10.

Keywords: ribbon graph, links in real projective space, Turaev surface, virtual link, partial dual, Tait graph.

Every signed plane graph represents a unique classical link diagram. In contrast, a single signed ribbon graph can represent several different link diagrams or virtual link diagrams. This observation leads to the fundamental problem of determining how link diagrams that are presented by the same signed ribbon graphs are related to each other. It is this problem that interests us here. It was solved for classical link diagrams in [Moffatt 2012]. Here we solve it for checkerboard colourable diagrams of links in $\mathbb{R}\mathrm{P}^3$ (in Theorem 7), and for virtual link diagrams (in Theorem 22).

We also examine the one-vertex ribbon graphs of diagrams of links in $\mathbb{R}\mathrm{P}^3$. Every classical link diagram can be represented as a ribbon graph with exactly one vertex. Abernathy et al. [2014] gave a set of moves that provide a way to move between all of the diagrams of a classical link that have one-vertex all-A ribbon graphs. We extend their work to the setting of links in $\mathbb{R}\mathrm{P}^3$.

This paper is structured as follows. In Section 2 we give an overview of diagrams of links in $\mathbb{R}\mathrm{P}^3$ and of ribbon graphs. In Section 3 we describe how diagrams of links in $\mathbb{R}\mathrm{P}^3$ can be represented by ribbon graphs, and we determine how checkerboard colourable diagrams that give rise to the same ribbon graphs are related to one another. In Section 4 we study the ribbon graphs of diagrams of links in $\mathbb{R}\mathrm{P}^3$ that have exactly one vertex. Finally, in Section 5 we describe how virtual link diagrams that give rise to the same ribbon graphs are related to one another.

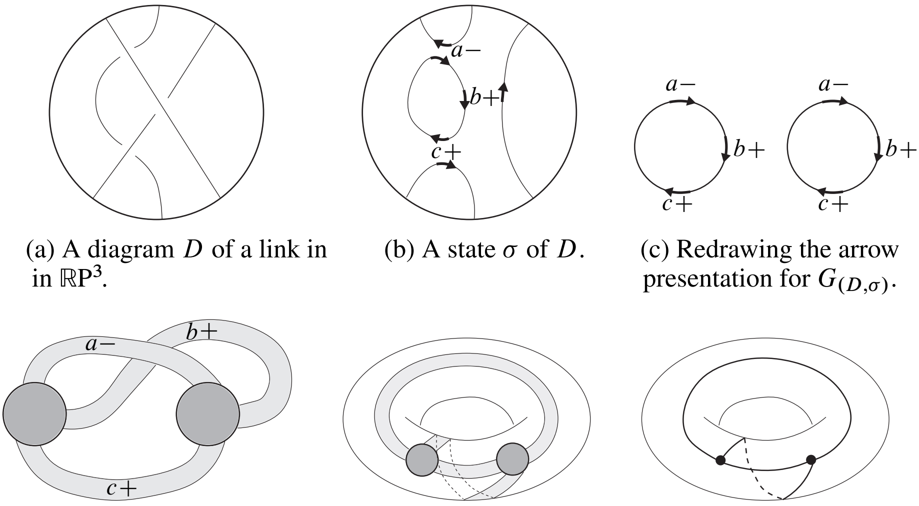
This work arose from Strömberg's undergraduate thesis at Royal Holloway, University of London, which was supervised by Moffatt.

2. Notation and terminology

2.1. ***Links in $\mathbb{R}\mathrm{P}^3$ and their diagrams.*** In this section we provide a brief overview of links in $\mathbb{R}\mathrm{P}^3$ and their diagrams. Further results and details can be found in [Drobotukhina 1994; 1990; Huynh and Le 2008; Mroczkowski 2003; Murasugi 1987; Prasolov and Sossinsky 1997].

A *diagram* of a link in $\mathbb{R}\mathrm{P}^3$ is a disc D^2 in the plane together with a collection of immersed arcs (where an arc is a compact connected 1-manifold possibly with boundary). The end points of arcs with boundary lie on the boundary of the disc ∂D^2, are divided into antipodal pairs, and these are the only points of the arcs that intersect ∂D^2. We further assume that the arcs are generically immersed, in that they have finitely many multiple points and each multiple point is a double point in which the arcs meet transversally. Finally, each double point is assigned an over/under-crossing structure, and is called a *crossing*. Figure 1(a) shows a diagram of a link in $\mathbb{R}\mathrm{P}^3$. Here, D will always refer to a diagram of a link.

A *net* is the real projective plane $\mathbb{R}\mathrm{P}^2$ together with a distinguished projective line, called the *line at infinity*, and a collection of generically immersed closed curves where each double point is assigned an over/under-crossing structure. Let



(a) A diagram D of a link in in $\mathbb{R}\mathrm{P}^3$.

(b) A state σ of D.

(c) Redrawing the arrow presentation for $G_{(D,\sigma)}$.

(d) $G_{(D,\sigma)}$ as a ribbon graph.

(e) Embedding $G_{(D,\sigma)}$ in a surface.

(f) $G_{(D,\sigma)}$ as a cellularly embedded graph.

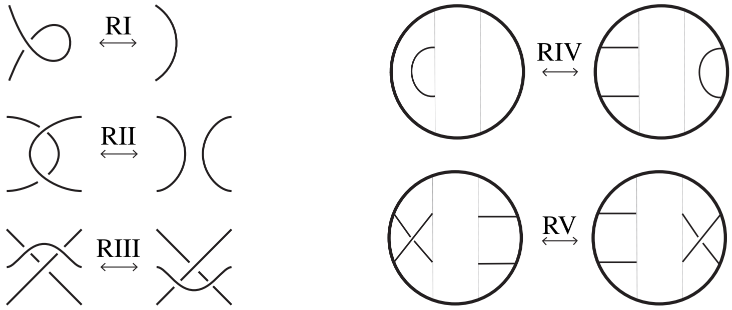
Figure 1. A diagram D of a link in $\mathbb{R}\mathrm{P}^3$ and one of its ribbon graphs.

D be a diagram of a link in $\mathbb{R}\mathrm{P}^3$, then the *net of* D, denoted $\mathcal{N}_D$, is obtained from D by identifying the antipodal points of ∂D^2. The image of ∂D^2 in the net gives the line at infinity.

A *component* of D is a collection of its arcs that give rise to a single closed curve in its net $\mathcal{N}_D$. A component is *null-homologous* if the corresponding curve in $\mathcal{N}_D$ is trivial in $H_1(\mathbb{R}\mathrm{P}^2) = \mathbb{Z}_2$ and is *1-homologous* otherwise. We will say that a diagram is *null-homologous* if each of its components is. The *faces* of D (respectively, $\mathcal{N}_D$) are the components of $D\backslash\alpha$ (respectively, $\mathcal{N}_D\backslash\alpha$), where α is the set of immersed curves. A *region* of D is a collection of its faces that correspond to a single face in its net $\mathcal{N}_D$. A diagram D is *checkerboard colourable* if there is an assignment of the colours black and white to its regions such that no two adjacent regions (those meeting a common arc) are assigned the same colour. A diagram may or may not be checkerboard colourable. For example, the diagram in Figure 1(a) is not, but that in Figure 7(d) is.

The *Reidemeister moves* for diagrams of links in $\mathbb{R}\mathrm{P}^3$ consist of isotopy of the disc that preserves the antipodal pairing (which we call the *R0-move*), together with the five moves in Figure 2 that change the diagram locally as shown (the diagrams are identical outside of the given region). In the figure, the bold lines represent the boundary of the disc. Two diagrams are *equivalent* if they are related by a sequence of Reidemeister moves.

For brevity we work a little informally in this paragraph, referring the reader to [Drobotukhina 1990] for details. Links in $\mathbb{R}\mathrm{P}^3$ give rise to diagrams by representing



(a) The classical moves.

(b) The boundary moves.

Figure 2. The Reidemeister moves for diagrams of links in $\mathbb{R}\mathrm{P}^3$.

$\mathbb{R}\mathrm{P}^3$ as a ball D^3 with antipodal points of its boundary identified, lifting the link from $\mathbb{R}\mathrm{P}^3$ to D^3 and projecting to the equatorial disc D^2. Conversely, given a diagram, regarding D^2 as the equatorial disc of such a representation of $\mathbb{R}\mathrm{P}^3$ and "pulling the over-crossings up a little" gives rise to a link in $\mathbb{R}\mathrm{P}^3$. With this, we have from [Drobotukhina 1990] that two links in $\mathbb{R}\mathrm{P}^3$ are ambient isotopic if and only if their diagrams are equivalent.

2.2. *Ribbon graphs.*

Definition 1. A *ribbon graph* $G = (V(G), E(G))$ is a (possibly nonorientable) surface with boundary represented as the union of two sets of discs, a set $V(G)$ of *vertices*, and a set of *edges* $E(G)$ such that

(1) the vertices and edges intersect in disjoint line segments;

(2) each such line segment lies on the boundary of precisely one vertex and precisely one edge;

(3) every edge contains exactly two such line segments.

An example of a ribbon graph can be found in Figure 1(d), and additional details about them can be found in, for example, [Ellis-Monaghan and Moffatt 2013; Gross and Tucker 2001].

Two ribbon graphs are *equivalent* if there is a homeomorphism taking one to the other that sends vertices to vertices, edges to edges, and preserves the cyclic ordering of the edges at each vertex. The homeomorphism should be orientation-preserving if the ribbon graphs are orientable. Note that any embedding of a ribbon graph is 3-space is irrelevant.

A ribbon graph is topologically a surface with boundary and the *genus* of a ribbon graph is its genus when it is viewed as a surface. It is *orientable* if it is orientable as a surface. A ribbon graph is said to be *plane* if it is homeomorphic

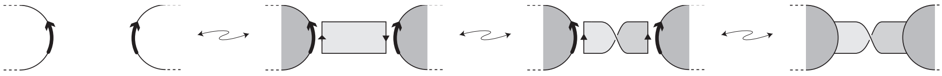


Figure 3. Moving between arrow presentations and ribbon graphs.

to a sphere with holes (or equivalently if it is connected and of genus zero), and is said to be $\mathbb{R}\mathrm{P}^2$ if it is homeomorphic to a real projective plane with holes (or equivalently it is connected, nonorientable and of genus one).

Since a ribbon graph is a surface with boundary, each ribbon graph G admits a unique (up to homeomorphism) cellular embedding into a closed surface Σ. (The cellular condition here means that $\Sigma\backslash G$ is a collection of discs). Using this embedding, it is easy to see that ribbon graphs are equivalent to cellularly embedded graphs (in one direction, contract the ribbon graph to obtain a graph drawn on the surface; in the other direction take a neighbourhood of the graph in a surface) and so are the main object of topological graph theory. See Figure 1(d)–(f).

We will make use of the following combinatorial description of ribbon graphs which is due to Chmutov [2009].

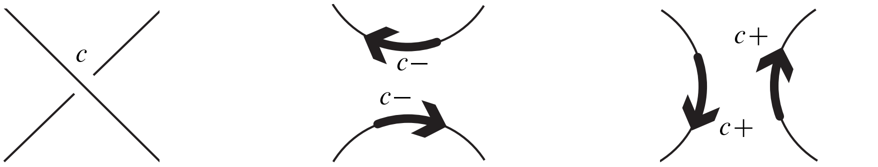
Definition 2. An *arrow presentation* consists of a set of closed curves, each with a collection of disjoint, labelled arrows, called *marking arrows*, lying on them. Each label appears on precisely two arrows.

A ribbon graph can be obtained from an arrow presentation as follows. View each closed curve as the boundary of a disc (the disc becomes a vertex of the ribbon graph). Edges are then added to the vertex discs in the following way: take an oriented disc for each label of the marking arrows; choose two nonintersecting arcs on the boundary of each of the edge discs and direct these according to the orientation; identify these two arcs with two marking arrows, both with the same label, aligning the direction of each arc consistently with the orientation of the marking arrow. This process is illustrated in Figure 3.

Conversely, to describe a ribbon graph G as an arrow presentation, start by arbitrarily labelling and orienting the boundary of each edge disc of G. On each arc where an edge disc intersects a vertex disc, place an arrow on the vertex disc, labelling the arrow with the label of the edge it meets and directing it consistently with the orientation of the edge-disc boundary. The boundaries of the vertex set marked with these labelled arrows give the arrow-marked closed curves of an arrow presentation. See Figure 1(c)–(d) for an example, and [Chmutov 2009; Ellis-Monaghan and Moffatt 2013] for further details.

Arrow presentations are *equivalent* if they describe equivalent ribbon graphs.

We will need to make use of signed ribbon graphs. A *signed ribbon graph* is a ribbon graph G together with a function from $E(G)$ to $\{+,-\}$. Thus it consists of a ribbon graph with a sign associated to each of its edges. Similarly, a *signed*



(a) A Crossing. (b) Its marked A-splicing. (c) Its marked B-splicing.

Figure 4. Marked splicings of a link diagram.

arrow presentation consists of an arrow presentation together with a function from its set of labels to $\{+,-\}$. Signed ribbon graphs and signed arrow presentations are equivalent in the obvious way.

3. The ribbon graphs of links in $\mathbb{R}\mathrm{P}^3$

3.1. ***The ribbon graphs of link diagrams.*** We now describe how a set of ribbon graphs can be associated to a link diagram. Let D be a diagram of a link in $\mathbb{R}\mathrm{P}^3$. Assign a unique label to each crossing of D. A *marked A-splicing* or a *marked B-splicing* of a crossing c is the replacement of the crossing with one of the schemes shown in Figure 4.

Notice that we decorate the two arcs in the splicing with signed labelled arrows that are chosen to be consistent with an arbitrary orientation of the disc. The labels of the arrows are determined by the label of the crossing, and the signs are determined by the choice of splicing.

A *state* σ of D is the result of marked A- or B-splicing each of its crossings. Observe that a state is a signed arrow presentation of a signed ribbon graph. We denote the signed ribbon graph corresponding to the state σ of D by $G_{(D,\sigma)}$. These ribbon graphs are the ribbon graphs of a link diagram:

Definition 3. Let D be a diagram of a link in $\mathbb{R}\mathrm{P}^3$. Then the *set of signed ribbon graphs associated with D*, denoted $\mathbb{G}_D$, is defined by

$$\mathbb{G}_D = \{G_{(D,\sigma)} \mid \sigma \text{ is a marked state of } D\}.$$

If $G \in \mathbb{G}_D$ then we say that G is a *signed ribbon graph of D*. We will also say that *G represents D*.

An example of a ribbon graph $G_{(D,\sigma)}$ for a state σ of a link diagram D is given in Figure 1(a)–(d). The construction of $\mathbb{G}_D$ is a direct extension of the construction for classical links from [Dasbach et al. 2008; Turaev 1987].

If D is checkerboard coloured, then we can construct a signed ribbon graph of D by choosing the splicing that follows the black regions at each crossing. The resulting signed ribbon graph is called a *Tait graph* of D. If D is checkerboard colourable, then it has exactly two Tait graphs, one corresponding to each of the two checkerboard colourings.

Proposition 4. *Let D be a checkerboard colourable diagram of a link in $\mathbb{R}\mathrm{P}^3$. Then its Tait graphs are either plane or $\mathbb{R}\mathrm{P}^2$ ribbon graphs.*

Proof. Checkerboard colour D and let G be its Tait graph. If D is not null-homologous then all of its regions are discs. Since the marked splicings follow the black regions and the black regions are discs, we can embed G in $\mathbb{R}\mathrm{P}^2$ by taking the black regions bounded by the curves of the splicings as vertices, and embedding the edge disc between the pairs of labelled arrows in the obvious way. Since D is checkerboard coloured, all regions of the embedded ribbon graph are discs, and no two face regions or vertex regions share a boundary. Thus G is cellularly embedded in the net and is therefore $\mathbb{R}\mathrm{P}^2$.

If D is null-homologous replace the face of its net that is a Möbius band with a disc to obtain a diagram on the sphere, and repeat the above argument with this embedding. □

We note that it follows from the proof of Proposition 4 that the Tait graphs defined here coincide with the "usual" Tait graphs obtained by placing vertices in black regions and embedding edges through each crossing.

Remark 5. One of the significant applications of the ribbon graphs of links is that they provide a way to connect graph and knot polynomials. A seminal result of Thistlethwaite [1987] expresses the Jones polynomial of an alternating classical link as an evaluation of the Tutte polynomial of either of its Tait graphs. There have been several recent extensions of this result that express the Jones polynomial and Kauffman bracket of virtual and classical links as evaluations of Bollobás and Riordan's extension of the Tutte polynomial to ribbon graphs; see [Bradford et al. 2012; Chmutov 2009; Chmutov and Pak 2007; Chmutov and Voltz 2008; Dasbach et al. 2008; Moffatt 2010; 2011].

Kauffman brackets and Jones polynomials of links in $\mathbb{R}\mathrm{P}^3$ can similarly be expressed in terms of the (multivariate) Bollobás–Riordan polynomials of ribbon graphs that represent their diagrams. In fact, the statement and proofs of the results for links in $\mathbb{R}\mathrm{P}^3$ follow those for the existing results with almost no change. Accordingly we only remark here that they hold. Following the notation of the exposition [Ellis-Monaghan and Moffatt 2013] gives that for a diagram D of a link in $\mathbb{R}\mathrm{P}^3$,

$$\langle D\rangle = d^{k(\mathbb{A})-1}A^{n(\mathbb{A})-r(\mathbb{A})}R(\mathbb{A}; -A^4, A^{-2}d, d^{-1}, 1),$$

and

$$\langle D\rangle = d^{-1}A^{e_-(G_D)-e_+(G_D)}Z(G_D; 1, \boldsymbol{w}, d, 1),$$

where

$$w_e = \begin{cases} A^{-2} & \text{if } e \text{ is negative,} \\ A^2 & \text{if } e \text{ is positive.} \end{cases}$$

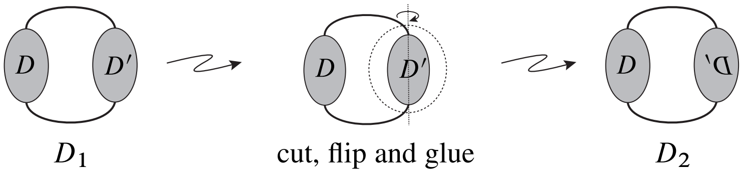


Figure 5. A summand-flip.

In these equations, $\langle D\rangle$ is the Kauffman bracket of [Drobotukhina 1990], $d = -A^2 - A^{-2}$, $\mathbb{A}$ is the ribbon graph of D obtained by choosing the A-splicing at each crossing, R is the Bollobás–Riordan polynomial [2002], and Z is the multivariate Bollobás–Riordan polynomial of [Moffatt 2008]. These identities can be obtained by following Section 5.4.2 of [Ellis-Monaghan and Moffatt 2013].

Furthermore, a connection between the Bollobás–Riordan polynomial and the HOMFLY-PT polynomial of links in $\mathbb{R}\mathrm{P}^3$ from [Mroczkowski 2004] that is analogous to Jaeger's connection [1988] between the Tutte polynomial of a plane graph and the HOMFLY-PT polynomial of a classical link (see also [Jin and Zhang 2012; Moffatt 2008; Traldi 1989]) can also be found:

$$P(\mathcal{L}(G);x,y)=\left(\frac{1}{xy}\right)^{v(G)-1}\left(\frac{y}{x}\right)^{e(G)}(x^2-1)^{k(G)-1}R\left(G;x^2,\frac{x-x^{-1}}{xy^2},\frac{y}{x-x^{-1}}\right).$$

Again the notation here is from [Ellis-Monaghan and Moffatt 2013], and the result can be obtained by following Section 5.5.2 of that text.

3.2. ***Relating link diagrams with the same ribbon graph.*** As mentioned in the introduction, two diagrams can give rise to the same set of signed ribbon graphs. That is, it is possible that $D \neq D'$ but $\mathbb{G}_D = \mathbb{G}_{D'}$. A fundamental question is then if D and D' are diagrams such that $\mathbb{G}_D = \mathbb{G}_{D'}$, how are D and D' related? Here we answer this question in the case when D and D' are both checkerboard colourable. To describe the result, we need to introduce some notation.

Definition 6. Let D and D' be diagrams of links in $\mathbb{R}\mathrm{P}^3$. We say that D and D' are related by a *summand-flip* if D' can be obtained from D by the following process: Orient the disc D^2 and choose a disc $\mathfrak{D}$ in D^2 whose boundary intersects D transversally in exactly two points a and b. Cut out $\mathfrak{D}$ and glue it back in such a way that the orientations of $\mathfrak{D}$ and $D^2\backslash\mathfrak{D}$ disagree, the points a on the boundaries of $\mathfrak{D}$ and $S^2\backslash\mathfrak{D}$ are identified, and the points b on the boundaries of $\mathfrak{D}$ and $S^2\backslash\mathfrak{D}$ are identified. See Figure 5. We say that two link diagrams D and D' are *related by summand-flips* if there is a sequence of summand-flips and R0-moves taking D to D'.

Our first main result is the following.

Theorem 7. *Let D and D' be checkerboard colourable diagrams of links in* $\mathbb{R}\mathrm{P}^3$. *Then $\mathbb{G}_D = \mathbb{G}_{D'}$ if and only if D and D' are related by summand-flips.*

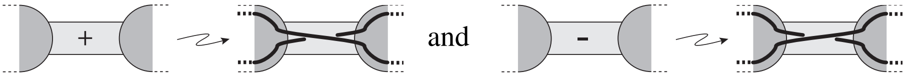


Figure 6. Forming a diagram D_G from a signed ribbon graph G.

Before proving Theorem 7, we note that the requirement that the link diagrams are checkerboard colourable is essential to our approach, and we pose the following.

Open problem. Let D and D' be diagrams of links in $\mathbb{R}\mathrm{P}^3$ (that are not necessarily checkerboard colourable). Determine necessary and sufficient conditions for $\mathbb{G}_D$ and $\mathbb{G}_{D'}$ to be equal.

To prove Theorem 7, we need to be able to recover link diagrams from ribbon graphs. Given a signed $\mathbb{R}\mathrm{P}^2$ or plane ribbon graph, it is straight-forward to recover a link diagram that it represents. Let G be a signed $\mathbb{R}\mathrm{P}^2$ ribbon graph, fill in the holes to obtain a cellular embedding of it in $\mathbb{R}\mathrm{P}^2$, as in Section 2.2. Represent $\mathbb{R}\mathrm{P}^2$ as a disc D^2 with antipodal points identified, and lift the embedding of G to a drawing on D^2. Finally, draw the configuration of Figure 6 on each of its edges, and connect the configurations by following the boundaries of the vertices of G, to obtain the link diagram. See Figure 7 for an example.

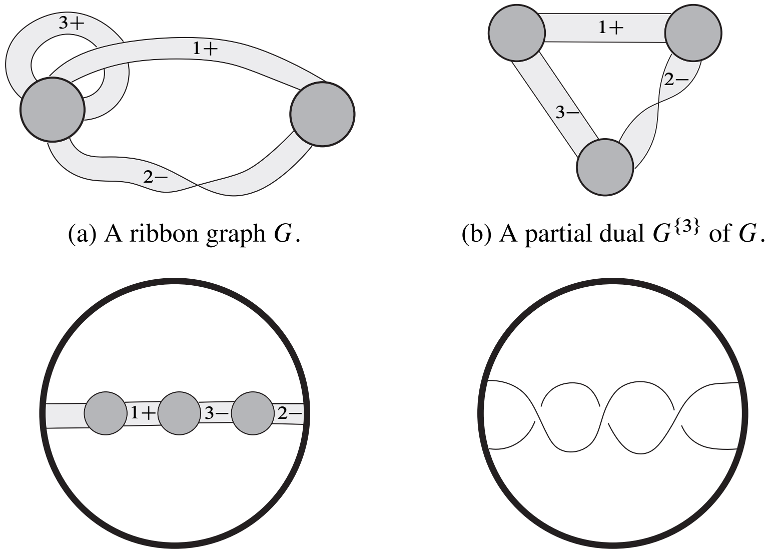
If G is a signed plane ribbon graph, fill in all but one of the holes to obtain a cellular embedding of it in a disc D^2. Drawing the configuration of Figure 6 on each of its edges and connecting the configurations by following the boundaries of the vertices of G gives the required link diagram. In either case, we denote the resulting diagram of a link in $\mathbb{R}\mathrm{P}^3$ by D_G.

Proposition 8. *Let G be a signed $\mathbb{R}\mathrm{P}^2$ or plane ribbon graph. Then D_G is checkerboard colourable.*

Proof. This follows by colouring the regions of D_G that correspond to the vertices of the ribbon graph black. □

To recover a link diagram from a ribbon graph that is not plane or $\mathbb{R}\mathrm{P}^2$ requires more work, and for our application, Chmutov's concept [2009] of a partial dual of a ribbon graph. The idea behind a partial dual is to form the geometric dual of an embedded graph but with respect to only some of its edges. We approach partial duals and geometric duals via arrow presentations as this is particularly convenient for us here. Other descriptions of partial duality can be found in, for example, [Chmutov 2009; Ellis-Monaghan and Moffatt 2013].

Definition 9. Let G be a ribbon graph viewed as an arrow presentation, and let $A \subseteq E(G)$. Then the *partial dual* G^A of G with respect to A is the arrow presentation (or ribbon graph) obtained as follows. For each $e \in A$, suppose α and β are the two arrows labelled e in the arrow presentation of G. Draw a line segment with an



(a) A ribbon graph G.

(b) A partial dual $G^{\{3\}}$ of G.

(c) Drawing $G^{\{3\}}$ in a disc.

(d) Recovering $D_{G^{\{3\}}}$.

Figure 7. Recovering a link diagram from a ribbon graph.

arrow on it directed from the head of α to the tail of β, and a line segment with an arrow on it directed from the head of β to the tail of α. Label both of these arrows e, and delete α and β and the arcs containing them. This process is illustrated locally at a pair of arrows in Figure 8. The ribbon graph $G^{E(G)}$ is the *geometric dual* of G.

If G is a signed ribbon graph then G^A is also a signed ribbon graph with the signs of G^A given by the rule that if an edge e of G has sign $\varepsilon \in \{+, -\}$, then the corresponding edge in G^A has sign $-\varepsilon$ if $e \in A$, and ε if $e \notin A$. (Thus taking the dual of an edge toggles its sign.)

Figure 7(a)–(b) gives an example of a partial dual.

We will need the following properties of partial duals from [Chmutov 2009].

Proposition 10. *Let G be a* (*signed*) *ribbon graph and $A, B \subseteq E(G)$. Then the following hold.*

(1) $G^{\varnothing} = G$.

(2) $G^{E(G)} = G^*$, *where G^* is the geometric dual of G.*

(3) $(G^A)^B = G^{(A \triangle B)}$, *where $A \triangle B = (A \cup B) \backslash (A \cap B)$ is the symmetric difference of A and B.*

(4) *G is orientable if and only if G^A is orientable.*

We emphasise that the construction of the geometric dual G^* of G agrees with the usual graph theoretic construction of the geometric dual of a cellularly embedded

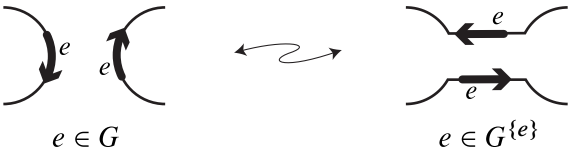


Figure 8. Taking the partial dual of an edge in an arrow presentations.

graph in which a cellularly embedded graph G^* is obtained from a cellularly embedded graph G by placing one vertex in each of its faces, and embedding an edge of G^* between two of these vertices whenever the faces of G they lie in are adjacent, and the edges of G^* are embedded so that they cross the corresponding face boundary (or edge of G) transversally.

Proposition 11. *Let G be a signed $\mathbb{R}\mathrm{P}^2$ or plane ribbon graph. Then $D_G = D_{G^*}$.*

Proof. Upon remembering that taking the dual of a signed ribbon graph changes the sign of each edge, the result is readily seen by comparing Figures 6 and 8. □

Lemma 12. *Let D be a diagram of a link in $\mathbb{R}\mathrm{P}^3$. Then all of the signed ribbon graphs in $\mathbb{G}_D$ are partial duals of each other.*

Proof. Let $G, H \in \mathbb{G}_D$. Then $G = G_{(D,\sigma)}$ and $H = H_{(D,\sigma')}$. It can be seen from Figure 8 that taking partial duals corresponds exactly to choosing another state of D as in Figure 4. □

Lemma 13. *Let D be a checkerboard colourable diagram of a link in $\mathbb{R}\mathrm{P}^3$. Then G represents D if and only if $D = D_{G^A}$, where G^A is a signed plane or $\mathbb{R}\mathrm{P}^2$ ribbon graph.*

Proof. We begin by assuming that $D = D_{G^A}$, where G^A is a signed plane or $\mathbb{R}\mathrm{P}^2$ ribbon graph. Then $G^A = G_{(D,\sigma)}$ for some state σ of D. By Lemma 12, it follows that the partial dual $(G^A)^A = G$ also represents D.

Conversely, assume that G represents D. Since D is checkerboard colourable, it can be represented by a Tait graph T. Clearly $D = D_T$. Then from Lemma 12, it follows that $T = G^A$ for some $A \subseteq E(G)$. Since T is a plane or $\mathbb{R}\mathrm{P}^2$ ribbon graph (by Proposition 4), T is the ribbon graph required by the lemma. □

Lemma 13 provides a way to construct all of the checkerboard colourable link diagrams represented by a given signed ribbon graph: find all of its plane or $\mathbb{R}\mathrm{P}^2$ partial duals and construct the links associated with them. This process is illustrated in Figure 7. The checkerboard colorability requirement here cannot be dropped. For example, if D is the diagram from Figure 1(a), then $\mathbb{G}_D$ contains no plane or $\mathbb{R}\mathrm{P}^2$ ribbon graphs. This leads to the following problem.

Open problem. Let G be a signed ribbon graph. Find an efficient way to construct all of the diagrams of links in $\mathbb{R}\mathrm{P}^3$ that have G as a representative.

We continue with some corollaries of Lemma 13.

Corollary 14. *Let D and D' be checkerboard colourable diagrams of links in $\mathbb{R}\mathrm{P}^3$ such that $\mathbb{G}_D = \mathbb{G}_{D'}$. Then D is null-homologous if and only if D' is.*

Proof. D is null-homologous if and only if it has a plane Tait graph. The result then follows since partial duality preserves orientability. □

Corollary 15. *Let D and D' be checkerboard colourable diagrams of links in $\mathbb{R}\mathrm{P}^3$ such that $\mathbb{G}_D = \mathbb{G}_{D'}$. Then there exists a plane or, respectively, $\mathbb{R}\mathrm{P}^2$ ribbon graph G, and $A \subseteq E(G)$ such that G^A is plane or, respectively, $\mathbb{R}\mathrm{P}^2$ and such that $D = D_G$ and $D' = D_{G^A}$.*

Proof. We have that D and D' give rise to the same set of ribbon graphs. Since D is checkerboard colourable, it gives rise to a plane or $\mathbb{R}\mathrm{P}^2$ ribbon graph G (namely one of its Tait graphs, by Proposition 4). Moreover, since D' is also checkerboard colourable, it also gives rise to a plane or $\mathbb{R}\mathrm{P}^2$ ribbon graph H. We also have that $H \in \mathbb{G}_D$, so $H = G^A$ for some $A \subseteq E(G)$ by Lemma 12. □

Corollary 15 is of key importance here: it tells us that if two checkerboard colourable diagrams of links in $\mathbb{R}\mathrm{P}^3$, D and D', are represented by the same ribbon graphs, then they are both diagrams associated with partially dual plane or $\mathbb{R}\mathrm{P}^2$ ribbon graphs G and G'. Thus if we understand how G and G' are related to each other, we can deduce how D and D' are related to each other. This is our strategy for proving Theorem 7.

In [Moffatt 2012; 2013], rough structure theorems for the partial duals of plane ribbon graphs and $\mathbb{R}\mathrm{P}^2$ ribbon graphs were given. These papers also contained local moves that allow us to move between all partially dual plane or $\mathbb{R}\mathrm{P}^2$ ribbon graphs. To describe this move, we need a little additional terminology.

Let G be a ribbon graph, $v \in V(G)$, and P and Q be nontrivial ribbon subgraphs of G. Then G is said to be the *join* of P and Q, written $P \vee Q$, if $G = P \cup Q$ and $P \cap Q = \{v\}$ and if there exists an arc on v with the property that all edges of P meet it there, and none of the edges of Q do. See the left-hand side of Figure 9, which illustrates a ribbon graph of the form $P \vee Q$. We do not require the ribbon graphs G, P or Q to be connected. Note that since genus is additive under joins, if G is plane then both P and Q are plane, and if G is $\mathbb{R}\mathrm{P}^2$ then exactly one of P or Q is $\mathbb{R}\mathrm{P}^2$ and the other is plane.

Let $G = P \vee Q$ be a ribbon graph. We say that the ribbon graph $G^{E(Q)} = P \vee Q^{E(Q)} = P \vee Q^*$ is obtained from G by a *dual-of-a-join-summand move*. We say that two ribbon graphs are related by *dualling join-summands* if there is a sequence of dual-of-a-join-summand moves taking one to the other, or if they are geometric duals. See Figure 9.

The following result is an amalgamation of Theorem 7.3 of [Moffatt 2012] and Theorem 5.8 of [Moffatt 2013].

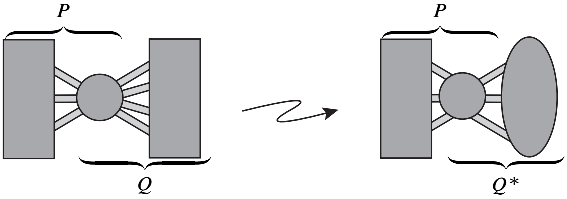


Figure 9. The dual of a join-summand move.

Theorem 16. *Let G and H be connected plane or $\mathbb{R}\mathrm{P}^2$ ribbon graphs. Then G and H are partial duals if and only if they are related by dualling join-summands.*

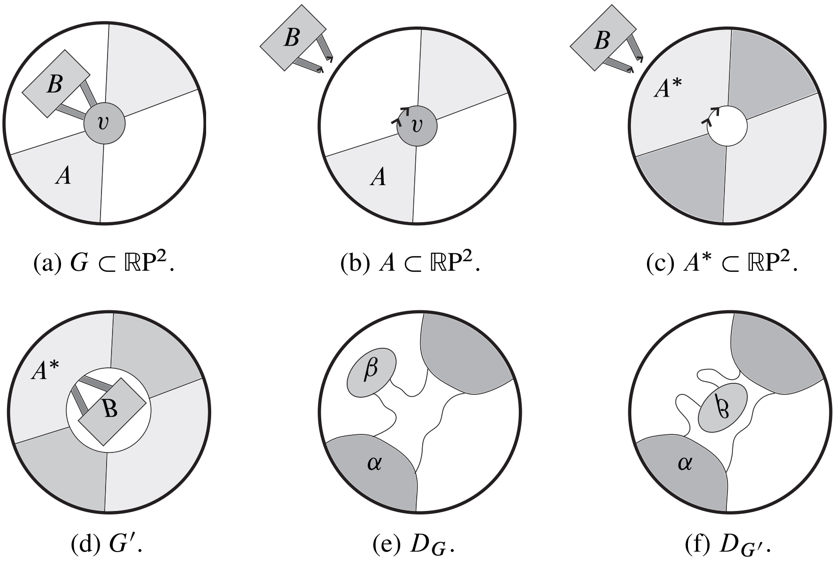
Theorem 16 allows us to prove the following key result.

Lemma 17. *If two $\mathbb{R}\mathrm{P}^2$ ribbon graphs G and G' are related by dualling join-summands, then the link diagrams D_G and $D_{G'}$ they represent are related by summand-flips.*

Proof. It suffices to show that if G and G' are related by a single dual-of-a-join-summand move then D_G and $D_{G'}$ are related by a summand-flip. Suppose that $G = A \vee B$, that $A \cap B = \{v\}$, and that $G' = A^* \vee B$ or $G' = A \vee B^*$. Since we know that genus is additive under joins, we have that one of A or B is $\mathbb{R}\mathrm{P}^2$ and the other is plane. Without loss of generality, suppose that A is the $\mathbb{R}\mathrm{P}^2$ summand.

First suppose that $G' = A^* \vee B$. We start by determining how the cellular embeddings of G and G' are related. From this, we will deduce how the corresponding link diagrams are related. Start by taking the cellular embedding of G in $\mathbb{R}\mathrm{P}^2$. This is illustrated in Figure 10(a). For each edge of B that meets v, place a labelled arrow on the intersection of the edge with v. We can then "detach" B from G, as indicated in Figure 10(b), so that G is recovered from A and B by identifying the corresponding arrows in A and in B with its copy of v removed. After detaching B, we obtain a cellular embedding of A in $\mathbb{R}\mathrm{P}^2$. From this, form the cellular embedding of A^* by interchanging the vertices and faces. (In detail, $A^* \subset \mathbb{R}\mathrm{P}^2$ is obtained from $A \subset \mathbb{R}\mathrm{P}^2$ by reassigning the face (respectively, vertex) discs of $A \subset \mathbb{R}\mathrm{P}^2$ as vertex (respectively, face) discs of $A^* \subset \mathbb{R}\mathrm{P}^2$. Edge discs are unchanged.) This is indicated in Figure 10(c). Finally, obtain an embedding of $G' = A^* \vee B$ by reattaching B according to the labelled arrows, as is indicated in arrows as in Figure 10(d), and notice that B has been "flipped over". Finally consider the diagrams D_G and $D_{G'}$ drawn using these embeddings. Since A and A^* have the same edges and vertex/face boundaries, and by Proposition 11, $D_A = D_{A^*}$, we see that D_G and $D_{G'}$ are related by a summand-flip, as in Figure 10(e)–(f).

Next suppose that $G' = A \vee B^*$. Then, using Proposition 10 and that duality preserves joins, we have $G' = (A \vee B^*) = (A \vee B^*)^{**} = (A^* \vee B^{**})^* = (A^* \vee B)^*$. Then since $D_{(A^* \vee B)^*} = D_{(A^* \vee B)}$, by Proposition 11, this case reduces to the first, completing the proof. □



(a) $G \subset \mathbb{R}P^2$. (b) $A \subset \mathbb{R}P^2$. (c) $A^* \subset \mathbb{R}P^2$.

(d) G'. (e) D_G. (f) $D_{G'}$.

Figure 10. A figure used in the proof of Lemma 17.

Proof of Theorem 7. It is readily seen that if D and D' are related by summand-flips then $\mathbb{G}_D = \mathbb{G}_{D'}$.

For the converse, assume that D and D' are checkerboard colourable link diagrams on $\mathbb{R}P^2$ such that $\mathbb{G}_D = \mathbb{G}_{D'}$. If D and D' are not null-homologous then, by Corollary 15, for some G, we have $D = D_G$ and $D' = D_{G^A}$, where G and G^A are both $\mathbb{R}P^2$. We know by Theorem 16 that G and G^A are related by dualling join-summands. Thus either $G^A = G^*$, in which case the result follows from Proposition 11, or G^A is obtained from G by a sequence of dual-of-a-join-summand moves, in which case the result follows from Lemma 17. □

4. One vertex ribbon graphs

We let $\mathbb{A}_D$ denote the *all-A ribbon graph* of D, which is the ribbon graph obtained from D by choosing the marked A-splicing at each crossing. The all-A ribbon graph is of particular interest since all of the signs are the same, and so a link diagram can be represented by an unsigned ribbon graph (see also Remark 5). It was shown in [Abernathy et al. 2014] that every classical link (i.e., in S^3) can be represented as a ribbon graph with exactly one vertex. Furthermore, the authors of that paper gave a set of moves, analogous to the Reidemeister moves, that provide a way to move between all of the diagrams of a classical link that have one-vertex all-A ribbon graphs. In this section we extend their result to links in $\mathbb{R}P^3$.

Lemma 18. *Every link in $\mathbb{R}P^3$ has a diagram D for which $\mathbb{A}_D$ has exactly one vertex.*

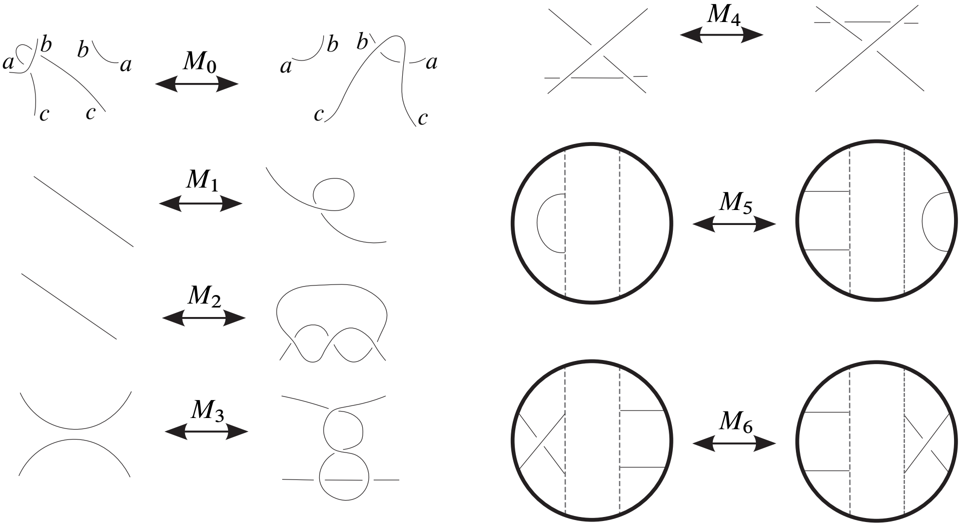


Figure 11. The M-moves.

Proof. Let D be a diagram of a link in $\mathbb{R}P^3$. Let σ_A denote the *all-A state* of D obtained by choosing the marked A-splicing at each crossing. If σ_A has exactly one component then $\mathbb{A}_D$ has exactly one vertex. Otherwise, consider the all-A state $\bar{\sigma}_A$ of the net $\mathcal{N}_D$ of D. There must be two closed curves of $\bar{\sigma}_A$ that can be joined by an embedded arc $\bar{\alpha}$ in $\mathbb{R}P^2 \setminus \bar{\sigma}_A$. Performing an RII-move (possibly with some RIV-moves) along the image of this arc in D gives a new diagram D'. Then $\mathbb{A}_{D'}$ has one less vertex than $\mathbb{A}_D$. Repeat this process until only one curve remains. □

The *M-moves* for diagrams of links in $\mathbb{R}P^3$ consist of isotopy of the disc that preserves the antipodal pairing, together with the moves shown in Figure 11 that change the diagram locally as shown (the diagrams are identical outside of the shown region). For the M_0-move, we require the diagram to be connected in a specific way, as indicated by the labels.

Lemma 19. *Let D be a diagram of a given link in $\mathbb{R}P^3$. Then the M-moves do not change the number of vertices in $\mathbb{A}_D$.*

Proof. For moves M_0–M_4, we refer the reader to [Abernathy et al. 2014]. It is easy to see that the M_5 move does not affect the number of components of the all-A state σ_A of D, since it does not affect the number of, or type of, crossings. It is also easy to see that M_6 does not change the number of vertices of the all-A ribbon graph. □

Let $\mathcal{D}$ denote the set of all diagrams of links in $\mathbb{R}P^3$, $\tilde{\mathcal{D}}$ denote $\mathcal{D}$ modulo the Reidemeister moves, $\mathcal{D}_1 \subset \mathcal{D}$ denote the subset of diagrams such that their all-A ribbon graphs have exactly one vertex, and $\tilde{\mathcal{D}}_1$ denote $\mathcal{D}_1$ modulo the M-moves. Now consider the two natural projections $\phi : \mathcal{D} \to \tilde{\mathcal{D}}$ and $\phi_1 : \mathcal{D}_1 \to \tilde{\mathcal{D}}_1$.

Theorem 20. *Given $D, D' \in \mathcal{D}_1$, we have $\phi(D) = \phi(D')$ if and only if $\phi_1(D) = \phi_1(D')$.*

Proof. First assume that $\phi_1(D) = \phi_1(D')$. Then the link diagrams are related by M-moves. It is easy to see that the link diagrams are then related by Reidemeister moves, so we have that $\phi(D) = \phi(D')$.

Conversely, suppose that $\phi(D) = \phi(D')$. Hence the diagrams are related by Reidemeister moves. We need to show that each Reidemeister move can be described as a sequence of M-moves. For RI–RIII, we refer the reader to [Abernathy et al. 2014]. RIV and RIV are exactly M_5 and M_6 moves, so we have that all the Reidemeister moves can be described as a sequence of M-moves. Hence $\phi_1(D) = \phi_1(D')$, as required. □

5. Virtual link diagrams with same the signed ribbon graphs.

A *virtual link diagram* consists of n closed piecewise-linear plane curves in which there are finitely many multiple points and such that at each multiple point exactly two arcs meet and they meet transversally. Moreover, each double point is assigned either a *classical crossing* structure or is marked as a *virtual crossing*. See the left-hand side of Figure 12, where the virtual crossings are marked by circles. A virtual link is *oriented* if each of its plane curves is. Further details on virtual knots can be found in, for example, the surveys [Kauffman 1999; 2000; 2012; Kaufman and Manturov 2006; Manturov 2004].

Virtual links are considered up to the *generalised Reidemeister moves*. These consist of orientation-preserving homeomorphisms of the plane (which we include in any subset of the moves), the classical Reidemeister moves of Figure 2(a), and the virtual Reidemeister moves of Figure 13. Two virtual link diagrams are *equivalent* if there is a sequence of generalised Reidemeister moves taking one diagram to the other.

Virtual knots are the knotted objects that can be represented by Gauss diagrams. Here a *Gauss diagram* consists of a set of oriented circles together with a set of oriented signed chords whose end points lie on the circles (see the right-hand

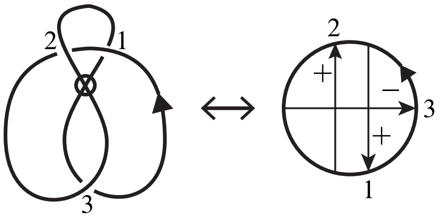


Figure 12. A virtual link (on the left) and its Gauss diagram (on the right). The crossings and chords are numbered for clarity.

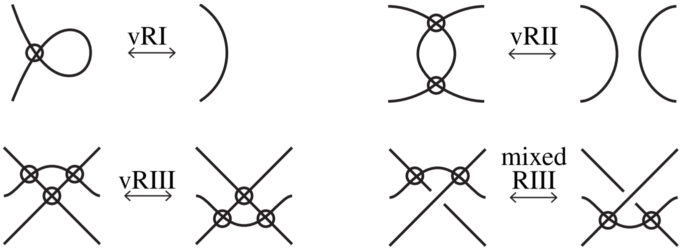


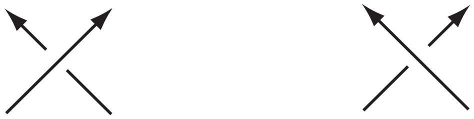
Figure 13. The virtual Reidemeister moves.

side of Figure 12). A Gauss diagram is obtained from an oriented n component virtual link diagram D as follows. Start by numbering each classical crossing. For each component, choose a base point and travel around the component from the base point following the orientation and reading off the numbers of the classical crossings as they are met. Whenever a crossing is met as an over-crossing, label the corresponding number with the letter O. Place each number, in the order met, on an oriented circle corresponding to the component. Connect the points on the circles that have the same number by a chord that is directed away from the O-labelled number. Finally, label each chord with the *oriented sign* of the corresponding crossing, shown in Figure 14, and delete the numbers. The resulting Gauss diagram describes D. See Figure 12 for an example.

Conversely, an oriented virtual link diagram can be obtained from a Gauss diagram by immersing the circles in the plane so that the ends of chords are identified (there is no unique way to do this), and using the direction and signs to obtain a crossing structure. In general, immersing the circles will create double points that do not arise from chords. Mark these as virtual crossings.

The following theorem of Goussarov, Polyak and Viro [Goussarov et al. 2000] provides an important and fundamental relation between Gauss diagrams and virtual links.

Theorem 21. *Let L and L' be two virtual link diagrams that are described by the same Gauss diagram. Then L and L' are equivalent. Moreover, L and L' are related by the virtual Reidemeister moves.*



(a) A positive crossing. (b) A negative crossing.

Figure 14. The oriented signs of a link diagram.

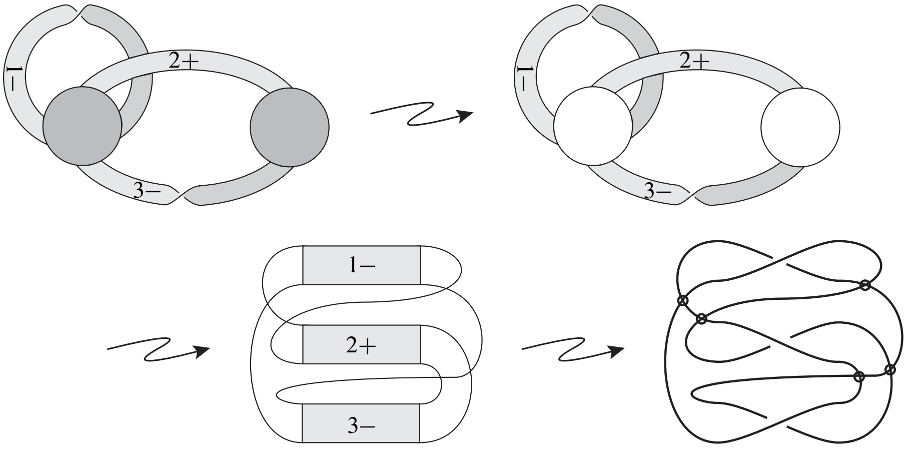


Figure 15. Recovering a virtual link diagram from a signed ribbon graph.

Chmutov and Voltz [2008] observed that the construction of a ribbon graph from a link diagram can be extended to include virtual links. That is, if D is a virtual link diagram and σ is a state of D, then $G_{(D,\sigma)}$ and the set $\mathbb{G}_D$ can be associated with D just as in Section 3.1 (virtual crossings are not smoothed, and the curves of the arrow presentation follow the component of the virtual link through the virtual crossings).

In Theorem 7 we determined how diagrams of links in $\mathbb{R}\mathrm{P}^3$ that are represented by the same set of ribbon graphs are related. We will now consider the corresponding problem for virtual links. We start by determining which ribbon graphs represent virtual link diagrams.

If G is a signed ribbon graph, then we can recover a virtual link diagram D with $G = G_D$ as follows: Delete the interiors of the vertices of G (so that we obtain a set of ribbons that are attached to circles). Immerse the resulting object in the plane in such a way that the ribbons are embedded. (Note that as the circles are immersed, they may cross each other and themselves.) Replace each embedded ribbon with a classical crossing with the crossing structure determined by the sign, as in Figure 6. Make all of the intersection points of the immersed circles into virtual crossings. See Figure 15. The resulting virtual link diagram D has the desired property that $G = G_D$ (as G can be obtained for D by reversing the above construction). Moreover, every virtual link diagram that is represented by G can be obtained in this way. This follows since if $G = G_D$, then we can go through the above process drawing the circles and crossings in such a way that they follow D.

Thus we have that every signed ribbon graph is the signed ribbon graph of some virtual link diagram.

We now determine how virtual link diagrams that are represented by the same ribbon graphs are related. For this we need the concept of virtualisation. The

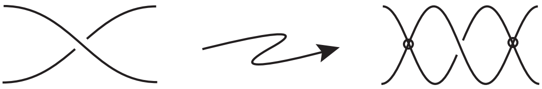


Figure 16. Virtualising a crossing.

virtualisation of a crossing of a virtual link diagram is the flanking of the crossing with virtual crossings as indicated in Figure 16. The crossing in the figure can also be of the opposite type.

Theorem 22. *Let D and D' be two virtual link diagrams. Then D and D' are presented by the same set of signed ribbon graphs if and only if they are related by virtualisation and the virtual Reidemeister moves.*

Proof. Let G be a signed ribbon graph. Label and arbitrarily orient each edge of G. As described above, every virtual link diagram represented by G can be obtained by (1) deleting the interiors of the vertices of G, (2) embedding the edges of G in the plane, (3) immersing the arcs connecting the edges (note that arcs in an immersion may cross each other), and (4) adding the crossing structure as described above.

Suppose D and D' are two virtual link diagrams obtained from G by this procedure. If the edges of G are oriented, in step (2) each embedding of an edge either agrees or disagrees with the orientation of the plane. If in step (2) of the constructions of D and D' the corresponding edges either both agree or both disagree with the orientation of the plane, it is easily seen that for some orientation of their components (in each diagram choose orientations that agree at each pair of crossings that correspond to the same edge of the ribbon graph), D and D' must then be described by the same Gauss diagram. In this case, by Theorem 21, they are related by the purely virtual moves and the semivirtual move.

Now suppose that in step (2) of the construction of D and D', there is an edge e of G such that the orientations of the two plane embeddings disagree with each other, and otherwise the embeddings of the edges and immersions of the arcs in step (3) are identical. Then, by Figure 17, the resulting virtual link diagrams are related by virtualisation.

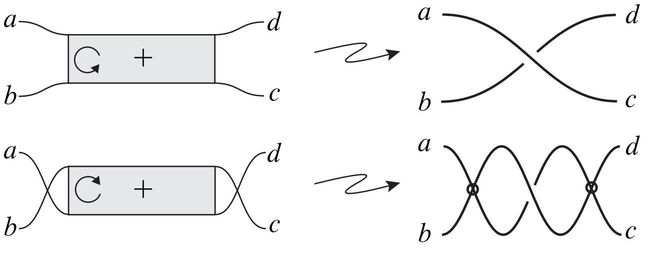


Figure 17. Forming virtual link diagrams from a signed ribbon graph.

It then follows that if G is a signed ribbon graph then the link diagrams it represents are related by virtualisation and the virtual moves. The converse of the theorem is easily seen to hold. □

References

[Abernathy et al. 2014] S. Abernathy, C. Armond, M. Cohen, O. T. Dasbach, H. Manuel, C. Penn, H. M. Russell, and N. W. Stoltzfus, "A reduced set of moves on one-vertex ribbon graphs coming from links", *Proc. Amer. Math. Soc.* **142**:3 (2014), 737–752. MR 3148509 Zbl 1283.05064

[Bollobás 1998] B. Bollobás, *Modern graph theory*, Graduate Texts in Mathematics **184**, Springer, New York, 1998. MR 99h:05001 Zbl 0902.05016

[Bollobás and Riordan 2002] B. Bollobás and O. Riordan, "A polynomial of graphs on surfaces", *Math. Ann.* **323**:1 (2002), 81–96. MR 2003b:05052 Zbl 1004.05021

[Bradford et al. 2012] R. Bradford, C. Butler, and S. Chmutov, "Arrow ribbon graphs", *J. Knot Theory Ramifications* **21**:13 (2012), 1240002. MR 2994589 Zbl 06109753

[Champanerkar and Kofman 2014] A. Champanerkar and I. Kofman, "A Survey on the Turaev genus of knots", *Acta Math. Vietnam.* **39**:4 (2014), 497–514. MR 3292579 Zbl 06388891

[Chmutov 2009] S. Chmutov, "Generalized duality for graphs on surfaces and the signed Bollobás–Riordan polynomial", *J. Combin. Theory Ser. B* **99**:3 (2009), 617–638. MR 2010f:05046 Zbl 1172.05015

[Chmutov and Pak 2007] S. Chmutov and I. Pak, "The Kauffman bracket of virtual links and the Bollobás–Riordan polynomial", *Mosc. Math. J.* **7**:3 (2007), 409–418, 573. MR 2008h:57006 Zbl 1155.57004

[Chmutov and Voltz 2008] S. Chmutov and J. Voltz, "Thistlethwaite's theorem for virtual links", *J. Knot Theory Ramifications* **17**:10 (2008), 1189–1198. MR 2009i:57027 Zbl 1163.57001

[Dasbach et al. 2008] O. T. Dasbach, D. Futer, E. Kalfagianni, X.-S. Lin, and N. W. Stoltzfus, "The Jones polynomial and graphs on surfaces", *J. Combin. Theory Ser. B* **98**:2 (2008), 384–399. MR 2009d:57020 Zbl 1135.05015

[Drobotukhina 1990] Y. V. Drobotukhina, "An analogue of the Jones polynomial for links in $\mathbf{R}P^3$ and a generalization of the Kauffman–Murasugi theorem", *Algebra i Analiz* **2**:3 (1990), 171–191. MR 91i:57001 Zbl 0713.57005

[Drobotukhina 1994] J. Drobotukhina, "Classification of links in $\mathbf{R}P^3$ with at most six crossings", pp. 87–121 in *Topology of manifolds and varieties*, edited by O. Viro, Adv. Soviet Math. **18**, Amer. Math. Soc., Providence, RI, 1994. MR 1296890 Zbl 0866.57007

[Ellis-Monaghan and Moffatt 2013] J. A. Ellis-Monaghan and I. Moffatt, *Graphs on surfaces: Dualities, polynomials, and knots*, Springer, New York, 2013. MR 3086663 Zbl 1283.57001

[Goussarov et al. 2000] M. Goussarov, M. Polyak, and O. Viro, "Finite-type invariants of classical and virtual knots", *Topology* **39**:5 (2000), 1045–1068. MR 2001i:57017 Zbl 1006.57005

[Gross and Tucker 2001] J. L. Gross and T. W. Tucker, *Topological graph theory*, Dover, Mineola, NY, 2001. MR 1855951 Zbl 0991.05001

[Huynh and Le 2008] V. Q. Huynh and T. T. Q. Le, "Twisted Alexander polynomial of links in the projective space", *J. Knot Theory Ramifications* **17**:4 (2008), 411–438. MR 2009h:57006 Zbl 1202.57016

[Jaeger 1988] F. Jaeger, "Tutte polynomials and link polynomials", *Proc. Amer. Math. Soc.* **103**:2 (1988), 647–654. MR 89i:57004 Zbl 0665.57006

[Jin and Zhang 2012] X. Jin and F. Zhang, "The Homfly and dichromatic polynomials", *Proc. Amer. Math. Soc.* **140**:4 (2012), 1459–1472. MR 2012m:57006 Zbl 1241.57004

[Kauffman 1999] L. H. Kauffman, "Virtual knot theory", *European J. Combin.* **20**:7 (1999), 663–690. MR 2000i:57011 Zbl 0938.57006

[Kauffman 2000] L. H. Kauffman, "A survey of virtual knot theory", pp. 143–202 in *Knots in Hellas '98* (Delphi, Greece, 1998), edited by C. M. Gordon et al., Ser. Knots Everything **24**, World Sci. Publ., River Edge, NJ, 2000. MR 2002j:57014 Zbl 1054.57001

[Kauffman 2012] L. H. Kauffman, "Introduction to virtual knot theory", *J. Knot Theory Ramifications* **21**:13 (2012), 1240007. MR 2994594 Zbl 1255.57005

[Kaufman and Manturov 2006] L. K. Kaufman and V. O. Manturov, "Virtual knots and links", *Tr. Mat. Inst. Steklova* **252**:Geom. Topol., Diskret. Geom. i Teor. Mnozh. (2006), 114–133. MR 2008c:57011

[Manturov 2004] V. Manturov, *Knot theory*, Chapman & Hall/CRC, Boca Raton, FL, 2004. MR 2005d:57008 Zbl 1052.57001

[Moffatt 2008] I. Moffatt, "Knot invariants and the Bollobás–Riordan polynomial of embedded graphs", *European J. Combin.* **29**:1 (2008), 95–107. MR 2008j:05116 Zbl 1142.57003

[Moffatt 2010] I. Moffatt, "Partial duality and Bollobás and Riordan's ribbon graph polynomial", *Discrete Math.* **310**:1 (2010), 174–183. MR 2011b:05112 Zbl 1229.05123

[Moffatt 2011] I. Moffatt, "Unsigned state models for the Jones polynomial", *Ann. Comb.* **15**:1 (2011), 127–146. MR 2012b:05087 Zbl 1235.05072

[Moffatt 2012] I. Moffatt, "Partial duals of plane graphs, separability and the graphs of knots", *Algebr. Geom. Topol.* **12**:2 (2012), 1099–1136. MR 2928906 Zbl 1245.05030

[Moffatt 2013] I. Moffatt, "Separability and the genus of a partial dual", *European J. Combin.* **34**:2 (2013), 355–378. MR 2994404 Zbl 1254.05047

[Mroczkowski 2003] M. Mroczkowski, "Diagrammatic unknotting of knots and links in the projective space", *J. Knot Theory Ramifications* **12**:5 (2003), 637–651. MR 2004d:57023 Zbl 1052.57008

[Mroczkowski 2004] M. Mroczkowski, "Polynomial invariants of links in the projective space", *Fund. Math.* **184** (2004), 223–267. MR 2005k:57030 Zbl 1072.57010

[Murasugi 1987] K. Murasugi, "Jones polynomials and classical conjectures in knot theory", *Topology* **26**:2 (1987), 187–194. MR 88m:57010 Zbl 0628.57004

[Prasolov and Sossinsky 1997] V. V. Prasolov and A. B. Sossinsky, *Knots, links, braids and 3-manifolds*, Translations of Mathematical Monographs **154**, Amer. Math. Soc., Providence, RI, 1997. MR 98i:57018 Zbl 0864.57002

[Thistlethwaite 1987] M. B. Thistlethwaite, "A spanning tree expansion of the Jones polynomial", *Topology* **26**:3 (1987), 297–309. MR 88h:57007 Zbl 0622.57003

[Traldi 1989] L. Traldi, "A dichromatic polynomial for weighted graphs and link polynomials", *Proc. Amer. Math. Soc.* **106**:1 (1989), 279–286. MR 90a:57013 Zbl 0713.57003

[Turaev 1987] V. G. Turaev, "A simple proof of the Murasugi and Kauffman theorems on alternating links", *Enseign. Math.* (2) **33**:3-4 (1987), 203–225. MR 89e:57002 Zbl 0668.57009

[Welsh 1993] D. J. A. Welsh, *Complexity: Knots, colourings and counting*, London Mathematical Society Lecture Note Series **186**, Cambridge University Press, 1993. MR 94m:57027 Zbl 0799.68008

Received: 2014-08-28 Revised: 2015-02-06 Accepted: 2015-02-06

iain.moffatt@rhul.ac.uk
Department of Mathematics, Royal Holloway University of London, Egham, Surrey, TW20 0EX, United Kingdom

anna.stromberg.2011@live.rhul.ac.uk
Department of Mathematics, Royal Holloway University of London, Egham, Surrey, TW20 0EX, United Kingdom

mathematical sciences publishers msp

INVOLVE 9:1 (2016)
dx.doi.org/10.2140/involve.2016.9.155

Depths and Stanley depths of path ideals of spines

Daniel Campos, Ryan Gunderson, Susan Morey,
Chelsey Paulsen and Thomas Polstra

(Communicated by Scott T. Chapman)

For a special class of trees, namely trees that are themselves a path, a precise formula is given for the depth of an ideal generated by all (undirected) paths of a fixed length. The dimension of these ideals is also computed, which is used to classify which such ideals are Cohen–Macaulay. The techniques of the proofs are shown to extend to provide a lower bound on the Stanley depth of these ideals. Combining these results gives a new class of ideals for which the Stanley conjecture holds.

1. Introduction

There is a well-known correspondence between square-free monomial ideals generated in degree two and graphs. If G is a graph on n vertices, let $R = k[x_1, \dots, x_n]$ be a polynomial ring over a field k in n variables and define the *edge ideal* $I = I(G)$ to be the ideal generated by all monomials of the form $x_i x_j$, where $\{x_i, x_j\}$ is an edge of G; see [Villarreal 1990]. The use of graphs to study algebraic properties of edge ideals has proven quite fruitful. A natural extension of the edge ideal of a graph is the path ideal of a graph. For each positive integer ℓ, define $P_\ell(G)$ to be the monomial ideal whose generators correspond to paths of length ℓ of G. Since the vertices of a path are distinct, $P_\ell(G)$ is a square-free monomial ideal. Various authors have used combinatorial information from the associated graphs to deduce information about depths of edge ideals [Dao et al. 2013; Dao and Schweig 2013; Fouli and Morey 2014; Herzog and Hibi 2005; Kummini 2009; Morey 2010]. The goal of this article is to examine the depth of a path ideal.

If $(R, \mathfrak{m})$ is a commutative, Noetherian, local ring and I is an ideal of R, the *depth* of R/I is an important algebraic invariant that, loosely speaking, provides one way to measure the size of R/I. More specifically, $\operatorname{depth}(R/I)$ is the maximal length of a sequence in $\mathfrak{m}$ that is regular on R/I. When $\operatorname{depth}(R/I) = \dim(R/I)$, the ring

MSC2010: primary 05E40, 13C14, 13F55; secondary 13A15, 05C25, 05C65, 05C05.
Keywords: Edge ideal, depth, path ideal, Cohen–Macaulay, monomial ideal.
All authors were partially supported by NSF (DMS 1005206). Morey is the corresponding author .

is said to be *Cohen–Macaulay*. There are many ways to detect depth, including the vanishing of Ext modules, local cohomology modules, or Koszul homology, to name a few. See [Herzog and Hibi 2011] for general information about depths and [Miller and Sturmfels 2005, Theorem 13.37] for a sample of the many ways the Cohen–Macaulay property can be detected. In this article, the depths of certain path ideals will be computed. Since the heights of such ideals are easily determined, the depth will be used to classify which such ideals are Cohen–Macaulay.

As a consequence of the method of proof employed, the results regarding the depth of path ideals of a special type of tree extend to a lower bound on the Stanley depth of the ideals. Let I be a monomial ideal. A *Stanley decomposition* of R/I is a direct sum decomposition $R/I = \bigoplus_{i=1}^{s} m_i R_{t_i}$ where m_i is a monomial and $R_{t_i} = k[x_{i_1}, \dots, x_{i_{t_i}}]$ is a polynomial subring of R generated over k by t_i of the variables of R. The depth of this decomposition is the minimum of the t_i, that is, the smallest number of variables used in any summand. The *Stanley depth*, denoted s-depth, of R/I is then the maximum depth of a Stanley decomposition of R/I. Introduced in [Stanley 1982], s-depth is a more geometric invariant attached to a monomial ideal, or more generally to a $\mathbb{Z}^r$-graded module. For a more detailed introduction to Stanley depths, see [Pournaki et al. 2009]. Stanley conjectured that the Stanley depth is always bounded below by the depth. By combining the bound found in Theorem 4.1 with Theorem 3.10, we prove that one class of path ideals is Stanley, that is, the Stanley conjecture holds true for this class of ideals. While other classes of Stanley ideals are known, see for instance [Pournaki et al. 2013] or [Cimpoeaş 2009], the conjecture is still largely open.

The contents of the paper are as follows. In Section 2 we provide the definitions and basic facts used throughout the paper. In Sections 3 and 4 we focus on the particular case where the tree T is a path. An exact formula for the depths of the path ideals is computed in Theorem 3.10. In Lemma 3.13, the dimension of such rings is given. Combining these results, Proposition 3.14 shows that if T is a path on n vertices, $P_\ell(T)$ is Cohen–Macaulay if and only if $n = \ell + 1$ or $n = 2\ell + 2$. In Section 4, using the techniques of Section 3, a bound is given in Theorem 4.1 for the s-depths of path ideals of T and as a result, in Corollary 4.2 these ideals are seen to be Stanley; that is, the Stanley conjecture is satisfied for this class of ideals.

2. Definitions and background

We begin by reviewing some standard notation and terminology regarding graphs and their connections to algebra. By abuse of notation, x_i will be used to denote both the vertex of a graph G and the corresponding variable of the polynomial ring R. For information regarding square-free monomial ideals, see [Villarreal 2001] and for additional background in graph theory, see [Harary 1969].

A *graph* is a vertex set $V = \{x_1, \dots, x_n\}$ together with a set $E = E(G) \subseteq V \times V$ of edges. As previously stated, associated to any graph G is a square-free monomial ideal generated in degree two, $I = I(G)$, called the edge ideal of I. Given a graph G, there is another family of square-free monomial ideals associated to G. For each positive integer ℓ, define $P_\ell(G)$ to be the monomial ideals whose generators correspond to paths of length ℓ of G. Notice that a path of length ℓ contains $\ell + 1$ vertices, so $P_\ell(G)$ is a homogeneous ideal with generators of degree $\ell + 1$. When $\ell = 1$, $P_1(G) = I(G)$ is the edge ideal of G. Notice that since a path is defined to have distinct vertices, $P_\ell(G)$ is a square-free monomial ideal.

The concept of a graph can be easily extended to one of a *clutter*, which is also called a *simple hypergraph*. A clutter $\mathfrak{C}$ is a vertex set V together with a set E of edges, where elements of E are nonempty subsets of V, with no inclusions among elements of E. That is, if $e, f \in E$, then $e \not\subset f$. For a graph, an edge $e \in E$ consists of two vertices, while for a clutter an edge may contain any number of vertices. Since a path ideal can be viewed as a special type of clutter with edges consisting of $\ell + 1$ vertices, tools from combinatorial optimization may be applied to path ideals.

Some basic notions from graph theory will be used throughout the paper and so are presented here for completeness. If $V' \subset V$ is a subset of the vertices of a graph G, the *induced subgraph* on V' is the graph G' given by $V(G') = V'$ and $E(G') = \{e \in E \mid e \subset V'\}$. That is, the edges of G' are precisely the edges of G with both endpoints in V'. If $x \in V(G)$, the *neighbor set* $N(x)$ is the set of all vertices that are adjacent to x, that is, $N(x) = \{y \in V(G) \mid \{x, y\} \in E(G)\}$. The *degree* of a vertex x is the cardinality of $N(x)$. A *leaf* is a vertex of degree one, and a *tree* is a connected graph where every induced subgraph has a leaf. A walk of length s is a collection of vertices and edges $x_0, e_1, x_1, e_2, \dots, e_s, x_s$ where $e_i = x_{i-1}x_i$ for $1 \leq i \leq s$. A walk without repeated vertices is a *path*. If T is a tree, then for any vertices $x, y \in V(G)$, there is a unique path between x and y. The length of this path is the *distance* between x and y, which is denoted by $d(x, y)$. In a general graph, $d(x, y)$ is the minimum of the lengths of all paths connecting x and y. A *forest* is a collection of trees. An *isolated vertex* is a vertex x with $N(x) = \varnothing$. Since $k[x_1, \dots, x_n, y]/(I, y) \cong k[x_1, \dots, x_n]/I$ for any monomial ideal I whose generators lie in $k[x_1, \dots, x_n]$, the graphs throughout this paper are generally assumed to be free of isolated vertices.

There are two common constructions used in combinatorial optimization that take a clutter or graph and produce smaller, related clutters or graphs. One is the *deletion* $\mathfrak{C} \setminus x$, which is formed by removing x from the vertex set of $\mathfrak{C}$ and deleting any edge in $\mathfrak{C}$ that contains x. This has the effect of setting $x = 0$, or of passing to the quotient ring $R/(x)$. The other operation is the *contraction*, $\mathfrak{C}/x$. This is performed by removing x from the vertex set and removing x from any edge that contains x. When $\mathfrak{C} = G$ is a graph, this will result in each vertex in $N(x)$ becoming an isolated vertex since if $y \in N(x)$ then removing x from the edge $\{x, y\}$ isolates y. Any additional edges containing y are removed since the definition of a clutter does not allow containments among edges. This operation has the effect of setting $x = 1$, or of passing to the localization R_x. A *minor* of a graph is formed by performing any combination of deletions and contractions.

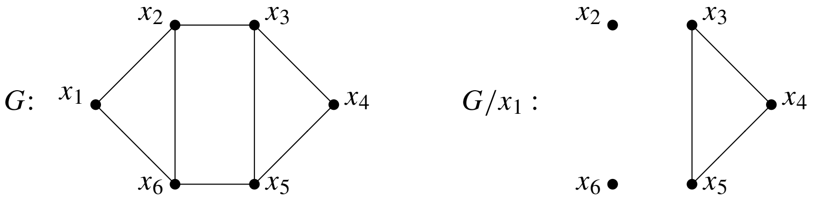
Example 2.1. The graph G shown below corresponds to the ideal

$$I = I(G) = (x_1x_2, x_1x_6, x_2x_3, x_2x_6, x_3x_4, x_3x_5, x_4x_5, x_5x_6).$$

Inverting x_1 yields

$$I_{x_1} = (x_2, x_6, x_2x_3, x_2x_6, x_3x_4, x_3x_5, x_4x_5, x_5x_6) = (x_2, x_6, x_3x_4, x_3x_5, x_4x_5),$$

which corresponds to the graph G/x_1:



If G is a graph, a *minimal vertex cover* of G is a set $C \subset V$ such that for every $e \in E$, $e \cap C \neq \varnothing$ and C is minimal with respect to this property, meaning if C' is any proper subset of C, then there exists an edge $e \in E$ with $e \cap C' = \varnothing$. The minimum cardinality of a minimal vertex cover of G (or $\mathfrak{C}$) is denoted by $\alpha_0 = \alpha_0(G)$. A prime ideal P is a *minimal prime of an ideal* I if $I \subset P$ and if Q is a prime ideal with $I \subset Q \subset P$, then $Q = P$. It is straightforward to check that C is a minimal vertex cover of G if and only if the prime ideal P generated by the variables corresponding to vertices of C is a minimal prime of $I(G)$. Thus $\alpha_0 = \text{height}(I)$.

The definition of depth is usually given for a local ring or with respect to a particular prime ideal. Although $R = k[x_1, \dots, x_n]$ is not a local ring, it has a unique homogeneous maximal ideal $\mathfrak{m} = (x_1, \dots, x_n)$. Since all ideals in this article are homogeneous ideals contained in $\mathfrak{m}$, R may be treated as a local ring. All depths will be taken with respect to $\mathfrak{m}$.

3. Depths of path ideals of spines

In general, it can be quite difficult to determine the precise depth of an ideal. In this section, we give an exact formula for the depth of a path ideal of a tree that does not branch. When combined with the height of the ideal, this formula allows us to determine which path ideals are Cohen–Macaulay. By noting that in this special case the directed path ideal is the same as the path ideal, and by using the Auslander–Buchsbaum formula, the depth formula found can be used to recover the projective dimension result of [He and Van Tuyl 2010, Theorem 4.1] which was also recovered in [Bouchat et al. 2011, Corollary 5.1]. However, the method of proof will allow us in Section 4 to extend the depth result to a bound on the Stanley depths of the ideals, as was done in [Pournaki et al. 2013] for powers of edge ideals. This bound shows that these ideals are Stanley.

The primary tool we will employ for computing depths is to form a family of short exact sequences and then apply the depth lemma (see, for example, [Bruns and Herzog 1993, Proposition 1.2.9], or [Villarreal 2001, Lemma 1.3.9]). In particular, if

$$0 \to A \to B \to C \to 0$$

is a short exact sequence of finitely generated R modules with homogeneous maps and $\operatorname{depth}(C) > \operatorname{depth}(A)$, then $\operatorname{depth}(B) = \operatorname{depth}(A)$. Note that the method used in this section is a variation of the method used in [Hà and Morey 2010; Morey 2010], where instead of using the left term of one sequence to form the subsequent sequence, the right-hand term is used. Starting with the standard short exact sequence

$$0 \to R/(I : z) \xrightarrow{f} R/I \xrightarrow{g} R/(I, z) \to 0$$

and making judicious choices for $z \in R$, we form a family of sequences

$$\begin{array}{ccccccccc} 0 & \to & R/K_1 & \to & R/I & \to & R/C_1 & \to & 0, \\ & & \vdots & & \vdots & & \vdots & & \\ 0 & \to & R/K_i & \to & R/C_{i-1} & \to & R/C_i & \to & 0, \\ & & \vdots & & \vdots & & \vdots & & \\ 0 & \to & R/K_s & \to & R/C_{s-1} & \to & R/C_s & \to & 0, \end{array} \tag{3-1}$$

where $C_0 = I$, $K_i = (C_{i-1} : z_i)$, and $C_i = (C_{i-1}, z_i)$ for $1 \le i \le s$. The goal is to find bounds on the depths of K_i for $1 \le i \le s$ and for C_s. Then applying the depth lemma starting with the last sequence and working back to the first will yield a bound on the depth of R/I. In this section, it will be easier to describe the sequence $\{z_i\}$ using a double index, so the ideals playing the roles of K_i and C_i will be doubly indexed as well.

A tree that does not branch is traditionally referred to as a *path*, however, to avoid the confusion of dealing with path ideals of paths, we will refer to such a graph as a spine. To be precise, we define a *spine* of length $n - 1$ to be a set of n distinct vertices $x_1, \dots, x_n$ together with $n - 1$ edges $x_i x_{i+1}$ for $1 \le i \le n - 1$. We denote such a spine by S_n and we will use $R = k[x_1, \dots, x_n]$ to denote the polynomial ring associated to S_n, or more generally, any graph on n vertices. As subrings of R will be used, define $R_t = k[x_1, \dots, x_t]$ for $t \le n$. While working with these ideals, it will often be convenient to work with subideals generated by selected paths. To facilitate this, define $P_{(\ell,s)}$ to be the ideal generated by the monomials corresponding to all paths of length ℓ of the spine connecting x_1 to x_s. For example, $P_{(2,5)} = (x_1x_2x_3, x_2x_3x_4, x_3x_4x_5)$. Using this notation, $P_\ell(S_n) = P_{(\ell,n)}$.

We first handle a special case.

Lemma 3.1. *Let S_n be a spine on n vertices. If $n \le \ell$, then* $\operatorname{depth}(R/P_{(\ell,n)}) = n$.

Proof. As $\ell \ge n$ we see that S_n does not contain a path of length ℓ. Thus $P_{(\ell,n)} = P_\ell(S_n) = (0)$ and we have $\operatorname{depth}(R/P_{(\ell,n)}) = \operatorname{depth}(R/(0)) = \operatorname{depth}(R) = n$. □

We now fix ℓ and n. In order to define the monomials that will serve the role of z_i above, it is useful to apply the division algorithm to produce unique integers b and c with $0 \le c < \ell + 2$ and $n - \ell - 1 = b(\ell + 2) + c$. It will often be convenient to write $n = (\ell + 1) + b(\ell + 2) + c$ throughout the paper. For $1 \le c \le \ell + 1$, define a sequence $\{a_{(j,k)}\}$ by

$$a_{(j,k)} = \prod_{t=n-\ell-k+1}^{n-j-k+1} x_t$$

for $1 \le j \le \min\{c, \ell\}$ and $1 \le k \le c - j + 1$. Note that for $c = 0$, the sequence is defined to be empty. The order in which the terms appear in this sequence is crucial to the definition of the family of sequences above. Specifically, the terms are ordered $a_{(1,1)}, a_{(1,2)}, a_{(1,3)}, \ldots, a_{(1,c)}, a_{(2,1)}, a_{(2,2)}, \ldots, a_{(2,c-1)}, a_{(3,1)}, \ldots$.

Example 3.2. Suppose $n = 18$ and $\ell = 6$. We then have $b = 1$ and $c = 3$ so our sequence of monomials $\{a_{(j,k)}\}$ is

$$\begin{aligned} a_{(1,1)} &= x_{12}x_{13}x_{14}x_{15}x_{16}x_{17}, & a_{(2,1)} &= x_{12}x_{13}x_{14}x_{15}x_{16},\\ a_{(1,2)} &= x_{11}x_{12}x_{13}x_{14}x_{15}x_{16}, & a_{(2,2)} &= x_{11}x_{12}x_{13}x_{14}x_{15},\\ a_{(1,3)} &= x_{10}x_{11}x_{12}x_{13}x_{14}x_{15}, & a_{(3,1)} &= x_{12}x_{13}x_{14}x_{15}. \end{aligned}$$

Using this sequence, we now define the ideals that will play the roles of C_i and K_i in the sequences above when $I = P_{(\ell,n)}$. Notice that since the sequence used is doubly indexed, the ideals C_i and K_i will require double indices as well, with the same ranges on the indices as above. We first define the ideals $C_{(j,k)} = (I, a_{(1,1)}, a_{(1,2)}, \ldots, a_{(j,k)})$. Note that for $c = 0$, the sequence was defined to be empty, and the only ideal defined is $C_{(0,k)} = P_{(\ell,n)}$ for all k. In general, the sequence of $a_{(j,k)}$ was selected so that many of the terms of $C_{(j,k)} = (I, a_{(1,1)}, a_{(1,2)}, \ldots, a_{(j,k)})$ will be redundant.

Next we define the ideals $K_{(j,k)}$, with the same bounds on j, k as before, by

$$K_{(j,k)} = \begin{cases} (C_{(j-1,c-(j-1)+1)} : a_{(j,1)}) & \text{if } k = 1, \\ (C_{(j,k-1)} : a_{(j,k)}) & \text{if } k > 1. \end{cases} \tag{3-2}$$

Notice that each $K_{(j,k)}$ is formed by taking the quotient ideal of the next term in the sequence with the preceding C ideal. It is straightforward to obtain an explicit formula for $K_{(j,k)}$ (see Proposition 3.4). The selection of the sequence $a_{(j,k)}$ was designed so that these quotient ideals will each have two elements of degree one, and these elements will make all paths of length less than ℓ redundant as generators.

Example 3.3. Assume $n = 18$ and $\ell = 6$. Then $b = 1$, $c = 3$ and the sequence $\{a_{(j,k)}\}$ is given in Example 3.2. Set $I = P_{(6,18)}$. Then

$$I = (x_1x_2x_3x_4x_5x_6x_7, x_2x_3x_4x_5x_6x_7x_8, \dots, \\ x_{11}x_{12}x_{13}x_{14}x_{15}x_{16}x_{17}, x_{12}x_{13}x_{14}x_{15}x_{16}x_{17}x_{18}).$$

By definition,

$$\begin{aligned} C_{(2,1)} &= (I, a_{(1,1)}, a_{(1,2)}, a_{(1,3)}, a_{(2,1)}), \\ C_{(2,2)} &= (I, a_{(1,1)}, a_{(1,2)}, a_{(1,3)}, a_{(2,1)}, a_{(2,2)}), \\ C_{(3,1)} &= (I, a_{(1,1)}, a_{(1,2)}, a_{(1,3)}, a_{(2,1)}, a_{(2,2)}, a_{(3,1)}). \end{aligned}$$

Removing redundant generators yields

$$\begin{aligned} C_{(2,1)} &= (x_1x_2x_3x_4x_5x_6x_7, \dots, x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}, \\ &\qquad x_{10}x_{11}x_{12}x_{13}x_{14}x_{15}, x_{12}x_{13}x_{14}x_{15}x_{16}), \\ C_{(2,2)} &= (x_1x_2x_3x_4x_5x_6x_7, \dots, x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}, \\ &\qquad x_{12}x_{13}x_{14}x_{15}x_{16}, x_{11}x_{12}x_{13}x_{14}x_{15}), \\ C_{(3,1)} &= (x_1x_2x_3x_4x_5x_6x_7, \dots, x_8x_9x_{10}x_{11}x_{12}x_{13}x_{14}, x_{12}x_{13}x_{14}x_{15}). \end{aligned}$$

Now by definition $K_{(2,2)} = (C_{(2,1)} : a_{(2,2)})$ and $K_{(3,1)} = (C_{(2,2)} : a_{(3,1)})$. Notice that $x_{10}a_{(2,2)} = a_{(1,3)} \in C_{(2,1)}$ and $x_{16}a_{(2,2)} = a_{(1,2)} \in C_{(2,1)}$. By the definitions of $a_{(j,k)}$ and of a path, all generators of $C_{(j,k)}$ will be products of consecutive vertices for all allowable j, k. Thus if $ma_{(2,2)} \in C_{(2,1)}$ for some monomial m, either $m \in C_{(2,1)}$, or m is divisible by x_{10} or x_{16} since those are the two vertices adjacent to the consecutive vertices appearing in $a_{(2,2)}$. Thus

$$K_{(2,2)} = (x_1x_2x_3x_4x_5x_6x_7, x_2x_3x_4x_5x_6x_7x_8, x_3x_4x_5x_6x_7x_8x_9, x_{10}, x_{16}).$$

Similarly $x_{11}a_{(3,1)} = a_{(2,2)} \in C_{(2,2)}$ and $x_{16}a_{(3,1)} = a_{(2,1)} \in C_{(2,2)}$, so $x_{11}, x_{16} \in K_{(3,1)}$, and

$$K_{(3,1)} = (x_1x_2x_3x_4x_5x_6x_7, \dots, x_4x_5x_6x_7x_8x_9x_{10}, x_{11}, x_{16}).$$

Proposition 3.4. *The family of ideals $K_{(j,k)}$ has the explicit formulation*

$$K_{(j,k)} = (P_{(\ell,n-\ell-k-1)}, x_{n-\ell-k}, x_{n-j-k+2}). \tag{3-3}$$

Proof. First notice that for $1 \le k \le c$, both $x_{n-\ell-k}a_{(1,k)}$ and $x_{n-k+1}a_{(1,k)}$ are generators of $I = P_{(\ell,n)}$, so $(C_{(1,k-1)}, x_{n-\ell-k}, x_{n-k+1}) \subseteq (C_{(1,k-1)} : a_{(j,k)})$, where $C_{(1,0)} = P_{(\ell,n)}$. The other inclusion is straightforward, so removing redundant elements from the list of generators yields the desired result for $j = 1$. Assume $j \geq 2$. By the definition of the sequence $\{a_{(j,k)}\}$, we have

$$x_{n-\ell-1}a_{(j,1)} = x_{n-\ell-1}\prod_{t=n-\ell}^{n-j} x_t = \prod_{t=n-\ell-1}^{n-(j-1)-2+1} x_t = a_{(j-1,2)},$$

and similarly $x_{n-j+1}a_{(j,1)} = a_{(j-1,1)}$. These equalities show that $x_{n-\ell-1}, x_{n-j+1} \in K_{(j,1)} = (C_{(j-1,c-(j-1)+1)} : a_{(j,1)})$. Since all generators of $C_{(j-1,c-(j-1)+1)}$ are products of consecutive vertices and $x_{n-\ell-1}, x_{n-j+1}$ are the only vertices that extend the consecutive path of vertices of $a_{(j,1)}$, we have

$$K_{(j,1)} = (C_{(j-1,c-(j-1)+1)}, x_{n-\ell-1}, x_{n-j+1}) = (P_{(\ell,n-\ell-1-1)}, x_{n-\ell-1}, x_{n-j+1}),$$

as desired, where the last equality follows from removing redundant generators.

Finally, if $k \geq 2$, we have $(a_{(j-1,k+1)} : a_{(j,k)}) = (x_{n-\ell-k})$ and $(a_{(j-1,k)} : a_{(j,k)}) = (x_{n-j-k+2})$. Thus for $k \geq 2$, $x_{n-\ell-k}, x_{n-j-k+2} \in K_{(j,k)} = (C_{(j,k-1)} : a_{(j,k)})$. Thus as before

$$K_{(j,k)} = (C_{(j,k-1)}, x_{n-\ell-k}, x_{n-j-k+2}) = (P_{(\ell,n-\ell-k-1)}, x_{n-\ell-k}, x_{n-j-k+2}). \qquad \square$$

Given this explicit form for $K_{(j,k)}$, it is easy to see that the depth of $K_{(j,k)}$ can be found inductively from the depth of the path ideal of a shorter spine. Thus the lemma below will allow us to simultaneously control the depth of each of the left-hand terms of the series of sequences. The proof is a direct application of [Morey 2010, Lemma 2.2] and thus is omitted.

Lemma 3.5. *For all j and k,*

$$\operatorname{depth}(R/K_{(j,k)}) = \operatorname{depth}(R_{n-\ell-k-1}/P_{(\ell,n-\ell-k-1)}) + \ell + k - 1.$$

We now need to control the depth of the final term of the final sequence. The nature of this proof will allow us to simultaneously handle the case $c = 0$, which was omitted above. For convenience, we will denote the final $C_{(j,k)}$ by $I_{(1)}$ and the final $a_{(j,k)}$ by $a_{(1)}$ since the final values of j and k depend on the relationship between c and ℓ. Explicitly, define

$$I_{(1)} = \begin{cases} I & \text{if } c = 0, \\ C_{(c,1)} & \text{if } 1 \leq c \leq \ell, \\ C_{(\ell,2)} & \text{if } c = \ell + 1, \end{cases} \qquad a_{(1)} = \begin{cases} a_{(c,1)} & \text{if } 1 \leq c \leq \ell, \\ a_{(\ell,2)} & \text{if } c = \ell + 1, \end{cases}$$

Note that since $I_{(1)}$ is used to denote the final $C_{(j,k)}$, we have

$$I_{(1)} = (I, a_{(1,1)}, a_{(1,2)}, \ldots, a_{(j,k)}),$$

where $I = P_{(\ell,n)}$ and all elements of the sequence $\{a_{(j,k)}\}$ are included in $I_{(1)}$.

The first two cases to consider follow directly from the definition of $C_{(j,k)}$ and an application of [Morey 2010, Lemma 2.2].

Lemma 3.6. *If* $c = \ell$, *then* $\operatorname{depth}(R/I_{(1)}) = \operatorname{depth}(R_{n-\ell-1}/P_{(\ell,n-\ell-1)}) + \ell$.

Proof. Notice that when $c = \ell$, $a_{(c,1)} = x_{n-\ell}$. Also note that $n-\ell-k+1 \leq n-\ell$ and since $k \leq c-j+1$, $n-j-k+1 \geq n-\ell$ when $c = \ell$. Thus $a_{(j,k)} = \prod_{t=n-\ell-k+1}^{n-j-k+1} x_t$ is a multiple of $x_{n-\ell}$ for all j, k when $c = \ell$. Thus $I_{(1)} = C_{(c,1)} = (P_{(\ell,n-\ell-1)}, x_{n-\ell})$ and the result follows from [Morey 2010, Lemma 2.2]. □

Lemma 3.7. *If* $c = \ell+1$, *then* $\operatorname{depth}(R/I_{(1)}) = \operatorname{depth}(R_{n-\ell-2}/P_{(\ell,n-\ell-2)}) + \ell + 1$.

Proof. Notice that when $c = \ell+1$, $a_{(\ell,1)} = x_{n-\ell}$ and $a_{(\ell,2)} = x_{n-\ell-1}$. As before, $a_{(j,k)}$ is a multiple of $x_{n-\ell}$ or of $x_{n-\ell-1}$ for all j, k, and thus the result follows from [Morey 2010, Lemma 2.2]. □

Finding the depth of $I_{(1)}$ for $0 \leq c \leq \ell-1$ will require another family of short exact sequences. Define a sequence of monomials by $b_{(h)} = \prod_{t=n-\ell+h}^{n-c} x_t$ for $1 \leq h \leq \ell - c$.

Example 3.8. As in Example 3.2 assume $n = 18$, $\ell = 6$, $b = 1$, and $c = 3$. Then $\{b_{(h)}\} = \{x_{13}x_{14}x_{15},\ x_{14}x_{15},\ x_{15}\}$.

We again form a family of short exact sequences using the sequence $\{b_{(h)}\}$. For convenience, define $J_{(0)} = I_{(1)}$. Now define $J_{(h)}$ and $L_{(h)}$ by $J_{(h)} = (J_{(h-1)}, b_{(h)})$ and $L_{(h)} = (J_{(h-1)} : b_{(h)})$. Then as in (3-1), we have the following family of short exact sequences:

$$\begin{array}{ccccccccc}
0 & \to & R/L_{(1)} & \to & R/I_{(1)} & \to & R/J_{(1)} & \to & 0, \\
0 & \to & R/L_{(2)} & \to & R/J_{(1)} & \to & R/J_{(2)} & \to & 0, \\
0 & \to & R/L_{(3)} & \to & R/J_{(2)} & \to & R/J_{(3)} & \to & 0, \\
 & & \vdots & & \vdots & & \vdots & & \\
0 & \to & R/L_{(l-c-1)} & \to & R/J_{(l-c-2)} & \to & R/J_{(l-c-1)} & \to & 0, \\
0 & \to & R/L_{(l-c)} & \to & R/J_{(l-c-1)} & \to & R/J_{(l-c)} & \to & 0.
\end{array} \tag{3-4}$$

Note that for each h, $b_{(h)} = x_{n-\ell+h}b_{(h+1)}$ and $a_{(1)} = x_{n-\ell}b_{(1)}$ where $a_{(1)}$ is the final term for the original sequence when $0 < c \leq \ell-1$ and $a_{(1)} = \prod_{t=n-\ell}^{n} x_t$ is the last generator of I when $c = 0$. Now $J_{(\ell-c)} = (I_{(1)}, b_{(1)}, \ldots, b_{(\ell-c)}) = (I_{(1)}, x_{n-c})$ since $b_{(\ell-c)} = x_{n-c}$ and $b_{(h)}$ is a multiple of $b_{(\ell-c)}$ for all other h. Now each $a_{(j,k)}$ is a multiple of x_{n-c}, and $I_{(1)} = (I, a_{(1,1)}, \ldots, a_{(j,k)})$, so removing redundant elements from the generating set yields $J_{(\ell-c)} = (P_{(\ell,n-c-1)}, x_{n-c})$. Similarly $L_{(h)} = (P_{(\ell,n-\ell+h-2)}, x_{n-\ell+h-1})$. Using these explicit forms of $J_{(\ell-c)}$ and $L_{(h)}$, combined with [Morey 2010, Lemma 2.2], we are able to express the depths of all of the left-hand terms and the final right-hand term of the sequences in (3-4) in terms of the depths of path ideals of shorter spines. Note that by the definition of $b_{(h)}$, we will assume $c \leq \ell-1$ whenever we are dealing with $J_{(h)}$ or $L_{(h)}$.

Lemma 3.9. *For all* h, $\operatorname{depth}(R/L_{(h)}) = \operatorname{depth}(R_{n-\ell+h-2}/P_{(\ell,n-\ell+h-2)}) + \ell - h + 1$ *and* $\operatorname{depth}(R/J_{(\ell-c)}) = \operatorname{depth}(R_{n-c-1}/P_{(\ell,n-c-1)}) + c$.

We are now able to prove the main result regarding the depth of a path ideal of a spine.

Theorem 3.10. *Let* S_n *be a spine of* n *vertices. Then*

$$\operatorname{depth}(R/P_\ell(S_n)) = \operatorname{depth}(R/P_{(\ell,n)}) = \begin{cases} \ell(b+1) & \text{if } c = 0, \\ \ell(b+1) + c - 1 & \text{if } c > 0. \end{cases}$$

Proof. We assume ℓ is fixed and induct on n. If $n \le \ell$, we have $b = -1$ and $c = n+1$. By Lemma 3.1, we have $\operatorname{depth}(R/P_{(\ell,n)}) = n$ and $\ell(b+1)+c-1 = \ell(0)+n+1-1 = n$, so the result holds.

Assume $n \ge \ell+1$. When writing $n = (\ell+1)+b(\ell+2)+c$, notice that for $n \ge 0$, $b = -1$ if and only if $n \le \ell$. Thus for $n \ge \ell+1$, $b \ge 0$. In the proof that follows, we will be working with $n-t$ for various values of t. When $b = 0$, this will often result in $n - t \le \ell$. While this situation can easily be handled using separate cases, allowing $b-1 = -1$ creates a more streamlined proof.

Suppose $0 \le c \le \ell - 1$. Then by Lemma 3.9,

$$\operatorname{depth}(R/L_{(h)}) = \operatorname{depth}(R_{n-\ell+h-2}/P_{(\ell,n-\ell+h-2)}) + \ell - h + 1,$$
$$\operatorname{depth}(R/J_{(\ell-c)}) = \operatorname{depth}(R_{n-c-1}/P_{(\ell,n-c-1)}) + c.$$

Recall that $P_{(\ell,n-\ell+h-2)} = P_\ell(S_{n-\ell+h-2})$. Since $h \le \ell - c$, we have $n-\ell+h-2 < n$. As ℓ has remained fixed, our inductive hypothesis on the number of vertices for a fixed path length applies. Thus by induction, if the division algorithm is used to write $n-\ell+h-2 = \ell+1+b'(\ell+2)+c'$ for some integers b' and c' with $0 \le c' < \ell+2$, then $\operatorname{depth}(R_{n-\ell+h-2}/P_{(\ell,n-\ell+h-2)}) = \ell(b'+1)+c'-1$ if $c' > 0$. Since $1 \le h \le \ell - c$ then $0 < c+h \le \ell$. Now $n-\ell+h-2 = \ell+1+(b-1)(\ell+2)+c+h$. Thus by induction,

$$\operatorname{depth}(R_{n-\ell+h-2}/P_{(\ell,n-\ell+h-2)}) = \ell((b-1)+1) + (c+h) - 1,$$

so Lemma 3.9 yields

$$\operatorname{depth}(R/L_{(h)}) = \ell(b) + c + h - 1 + \ell - h + 1 = \ell(b+1) + c.$$

Also by induction, using a similar argument on the number of vertices,

$$\operatorname{depth}(R_{n-c-1}/P_{(\ell,n-c-1)}) = \ell(b-1+1) + (\ell+1) - 1 = \ell(b+1)$$

since $n-c-1 = (\ell+1)+(b-1)(\ell+2)+\ell+1$, so $\operatorname{depth}(R/J_{(\ell-c)}) = \ell(b+1)+c$. Now repeated use of the depth lemma applied to (3-4) yields $\operatorname{depth} R/I_{(1)} = \ell(b+1)+c$.

Suppose $c = \ell$. Then by Lemma 3.6 we have

$$\operatorname{depth}(R/I_{(1)}) = \operatorname{depth}(R_{n-\ell-1}/P_{(\ell,n-\ell-1)}) + \ell.$$

Then $n - \ell - 1 = \ell + 1 + b(\ell + 2) + \ell - \ell - 1 = \ell + 1 + (b - 1)(\ell + 2) + \ell + 1$. Thus applying the inductive hypothesis with $b' = b - 1$ and $c' = \ell + 1$ yields

$$\operatorname{depth}(R_{n-\ell-1}/P_{(\ell,n-\ell-1)}) = \ell(b - 1 + 1) + (\ell + 1) - 1,$$

so $\operatorname{depth}(R/I_{(1)}) = \ell b + \ell + \ell = \ell(b + 1) + c$.

If $c = \ell + 1$, then by Lemma 3.7 we have

$$\operatorname{depth}(R/I_{(1)}) = \operatorname{depth}(R_{n-\ell-2}/P_{(\ell,n-\ell-2)}) + \ell + 1.$$

Then $n - \ell - 2 = \ell + 1 + (b - 1)(\ell + 2) + c$, so by induction,

$$\operatorname{depth}(R_{n-\ell-2}/P_{(\ell,n-\ell-2)}) = \ell(b - 1 + 1) + c - 1,$$

and $\operatorname{depth}(R/I_{(1)}) = \ell(b) + c - 1 + \ell + 1 = \ell(b + 1) + c$.

We now have $\operatorname{depth}(R/I_{(1)}) = \ell(b + 1) + c$ for all possible values of c. Notice that if $c = 0$, we have $P_{(\ell,n)} = I_{(1)}$ and $\operatorname{depth}(R/P_{(\ell,n)}) = \ell(b + 1)$ for any b, and the result holds. Thus we may now assume $c > 0$ for the remainder of the proof.

By Lemma 3.5, for all j, k,

$$\operatorname{depth}(R/K_{(j,k)}) = \operatorname{depth}(R_{n-\ell-k-1}/P_{(\ell,n-\ell-k-1)}) + \ell + k - 1.$$

Now if $n = (\ell+1)+b(\ell+2)+c$, then $n-\ell-k-1 = (\ell+1)+(b-1)(\ell+2)+c-k+1$. Notice that $c - k + 1 > 0$ since $k \leq c - j + 1$. Thus we have

$$\operatorname{depth}(R_{n-\ell-k-1}/P_{(\ell,n-\ell-k-1)}) = \ell(b - 1 + 1) + c - k + 1 - 1 = \ell(b) + c - k$$

by induction. Then $\operatorname{depth}(R/K_{(j,k)}) = \ell(b) + c - k + \ell + k - 1 = \ell(b + 1) + c - 1$. Now repeated application of the depth lemma to the sequences in (3-1) yields $\operatorname{depth}(R/P_{(\ell,n)}) = \ell(b + 1) + c - 1$ when $c > 0$. □

There are some interesting reformulations of the depth found in Theorem 3.10. They are stated here without proof as the proofs are basic computations and summation arguments.

Corollary 3.11. *Theorem 3.10 can be reformulated as*

$$\operatorname{depth}(R/P_{(\ell,n)}) = \begin{cases} m\ell & \text{if } \ell \leq (n - 2m + 2)/m, \\ n - 2m + 2 & \text{if } \ell > (n - 2m + 2)/m, \end{cases}$$

where $m = \lceil n/(\ell + 2)\rceil$, *or as*

$$\operatorname{depth}(R/P_{(\ell,n)}) = \sum_{i=0}^{\ell-1} \left\lceil \frac{n - i}{\ell + 2} \right\rceil.$$

Notice that when ℓ is large relative to n, the depth of $R/P_{(\ell,n)}$ is large. If $\ell > n$, then the depth is n, as was noted in Lemma 3.1. However it is interesting to note that as long as ℓ is roughly half of n or larger, the depth remains quite large.

Corollary 3.12. *If* $\ell \geq (n-2)/2$, *then* $\operatorname{depth}(R/P_{(\ell,n)}) = n-2$ *for* $\ell \neq n-1$ *and for* $\ell = n-1$, $\operatorname{depth}(R/P_{(\ell,n)}) = n-1$.

Proof. Since $\ell \geq (n-2)/2$, we have $b = 0$, where $n = (\ell+1) + b(\ell+2) + c$ and $c \leq \ell+1$. By Theorem 3.10, if $c = 0$, $\operatorname{depth}(R/P_{(\ell,n)}) = \ell(b+1) = \ell = n-1$ and if $c > 0$, then $\operatorname{depth}(R/P_{(\ell,n)}) = \ell(b+1) + c - 1 = \ell + c - 1 = n-2$. □

To determine when the ideal is Cohen–Macaulay, the dimension is first needed. Since that is of independent interest, it is stated separately.

Lemma 3.13. *If* $I = P_{(\ell,n)}$, *then* $\dim(R/I) = n - \lfloor n/(\ell+1) \rfloor$.

Proof. Let $m = \lfloor n/(\ell+1) \rfloor$. The set of vertices $M = \{x_{\ell+1}, x_{2\ell+2}, \dots, x_{m\ell+m}\}$ forms a minimal vertex cover of minimal cardinality of $I = P_{(\ell,n)}$, so $\operatorname{height}(I) = \lfloor n/(\ell+1) \rfloor$. Since R is a polynomial ring of dimension n, $\dim(R/I) = n - \lfloor n/(\ell+1) \rfloor$. □

Proposition 3.14. *Let* $I = P_{(\ell,n)}$. *Then* R/I *is Cohen–Macaulay if and only if* $n = \ell+1$ *or* $n = 2\ell+2$.

Proof. Let $I = P_{(\ell,n)}$. If $n = 2\ell+2$, then by Lemma 3.13,

$$\dim(R/I) = n - \left\lfloor \frac{2\ell+2}{\ell+1} \right\rfloor = n-2,$$

and by Corollary 3.12, $\operatorname{depth}(R/P_{(\ell,n)}) = n-2$. Thus $R/P_{(\ell,n)}$ is Cohen–Macaulay. For $n = \ell+1$, Lemma 3.13 yields

$$\dim(R/P_{(\ell,n)}) = n - \left\lfloor \frac{\ell+1}{\ell+1} \right\rfloor = n-1,$$

and Corollary 3.12 gives $\operatorname{depth}(R/P_{(\ell,n)}) = n-1$, which again shows that $R/P_{(\ell,n)}$ is Cohen–Macaulay.

For the converse, consider $n = \ell+1+b(\ell+2)+c$ with $0 \leq c < \ell+2$. By Lemma 3.13,

$$\dim(R/I) = n - \left\lfloor \frac{n}{\ell+1} \right\rfloor = \left\lceil \frac{n(\ell+1)-n}{\ell+1} \right\rceil = (b+1)\ell + \left\lceil \frac{(b+c)\ell}{\ell+1} \right\rceil.$$

If $c = 0$ then $\operatorname{depth}(R/I) = \ell(b+1)$ by Theorem 3.10. If R/I is Cohen–Macaulay, then $(b+1)\ell + \lceil b\ell/(\ell+1) \rceil = \ell(b+1)$. Thus $\lceil b\ell/(\ell+1) \rceil = 0$, or $b = 0$. Since $b = c = 0$, we have $n = \ell+1$.

If $c > 0$, then $\operatorname{depth}(R/I) = \ell(b+1) + c - 1$ by Theorem 3.10. If R/I is Cohen–Macaulay, then

$$(b+1)\ell + \left\lceil \frac{(b+c)\ell}{\ell+1} \right\rceil = \ell(b+1) + c - 1 \quad \text{or} \quad \left\lceil \frac{(b+c)\ell}{\ell+1} \right\rceil = c-1.$$

Now

$$\left\lceil \frac{(b+c)\ell}{\ell+1} \right\rceil = b+c-\left\lfloor \frac{b+c}{\ell+1} \right\rfloor,$$

so we have $b - \lfloor (b+c)/(\ell+1) \rfloor = -1$. If $b \geq 1$, the left side of this equation is nonnegative, a contradiction. If $b = 0$, then $-\lfloor (b+c)/(\ell+1) \rfloor = -1$ if and only if $c = \ell+1$ since $c < \ell+2$. Thus if $c > 0$, R/I is Cohen–Macaulay if and only if $b = 0$ and $c = \ell+1$. In this case $n = 2\ell+2$.

Thus if R/I is Cohen–Macaulay, $n = \ell+1$ or $n = 2\ell+2$. □

Proposition 3.14 is particularly interesting when compared to [Campos et al. 2014, Theorem 3.8]. In fact, in each of the two instances where the path ideal of a spine is Cohen–Macaulay, the graph can be viewed as a suspension. When $n = 2\ell+2$, $P_{(\ell,n)}$ is the suspension of length ℓ of a graph that consists of a single edge connecting two vertices $(x_{n/2}, x_{n/2+1})$ and when $n = \ell+1$, $P_{(\ell,n)}$ is the suspension of length ℓ of a graph that consists of a single isolated vertex (x_n).

Note that the arguments in Proposition 3.14 can be used to determine the *Cohen–Macaulay defect*, that is $\dim(R/I) - \operatorname{depth}(R/I)$, for a path ideal. For example, if $\ell+1 < n < 2\ell+2$, $\operatorname{depth}(R/P_{(\ell,n)}) = n-2$ and $\dim(R/P_{(\ell,n)}) = n-1$ so the Cohen–Macaulay defect is 1.

4. Stanley depths of path ideals of spines

As remarked before, Theorem 3.10 together with the Auslander–Buchsbaum formula, recovers the projective dimension found in [He and Van Tuyl 2010, Theorem 4.1] and in [Bouchat et al. 2011, Corollary 5.1]. However, the method of proof has the advantage of also yielding information about the Stanley depth. There are three key factors that allow us to extend the depth result to a lower bound on the Stanley depth, or s-depth for brevity. The first two are well-known basic facts. If I is a monomial ideal of a polynomial ring R and y is an indeterminate, then

$$\text{s-depth}(R[y]/IR[y]) = \text{s-depth}(R/I) + 1, \tag{4-1}$$

and $\text{s-depth}(R) = n$ when R is a polynomial ring in n variables. The third result we will need is that s-depth satisfies a partial version of the depth lemma. In particular, it was shown in [Rauf 2010, Lemma 2.2] that if

$$0 \to A \to B \to C \to 0$$

is a short exact sequence of finitely generated R modules then

$$\text{s-depth}(B) \geq \min\{\text{s-depth}(A), \text{s-depth}(C)\}.$$

Now by carefully examining the proof of Theorem 3.10, we are able to extend the result to a lower bound on the s-depth of the path ideal of a spine. Note that the explicit calculations closely follow those of Theorem 3.10 and so details have been condensed in the proof.

Theorem 4.1. *Let S_n be a spine on n vertices. Then*

$$\text{s-depth}(R/P_\ell(S_n)) = \text{s-depth}(R/P_{(\ell,n)}) \geq \begin{cases} \ell(b+1) & \text{if } c=0, \\ \ell(b+1)+c-1 & \text{if } c>0. \end{cases}$$

Proof. We assume ℓ is fixed and induct on n. Write $n=(\ell+1)+b(\ell+2)+c$. If $n \leq \ell$, s-depth$(R/P_{(\ell,n)}) =$ s-depth$(R) = n$ and $\ell(b+1)+c-1 = \ell(0)+n+1-1 = n$ and the result holds. Define the sequences $a_{(j,k)}$ and $b_{(h)}$ and the related ideals $K_{(j,k)}$, $C_{(j,k)}$, $L_{(h)}$, $J_{(h)}$ and $I_{(1)}$ as before. By Proposition 3.4,

$$\text{s-depth}(R/K_{(j,k)}) = \text{s-depth}(R_{n-\ell-k-1}/P_{(\ell,n-\ell-k-1)}) + \ell + k - 1,$$

and by induction

$$\text{s-depth}(R_{n-\ell-k-1}/P_{(\ell,n-\ell-k-1)}) \geq \ell(b-1+1)+c-k+1-1 = \ell(b)+c-k,$$

so s-depth$(R/K_{(j,k)}) \geq \ell(b)+c-k+\ell+k-1 = \ell(b+1)+c-1$.

If $c=\ell$ or $c=\ell+1$, then as in Lemma 3.6 or Lemma 3.7 with [Morey 2010, Lemma 2.2] replaced by (4-1),

$$\text{s-depth}(R/I_{(1)}) = \text{s-depth}(R_{n-\ell-1}/P_{(\ell,n-\ell-1)}) + \ell$$

when $c=\ell$, and

$$\text{s-depth}(R/I_{(1)}) = \text{s-depth}(R_{n-\ell-2}/P_{(\ell,n-\ell-2)}) + \ell + 1$$

when $c=\ell+1$. In either case, applying the inductive hypothesis as in Theorem 3.10 yields

$$\text{s-depth}(R/I_{(1)}) \geq \ell(b+1)+c.$$

Suppose $0 \leq c \leq \ell-1$. Then as in Lemma 3.9 with [loc. cit., Lemma 2.2] replaced by (4-1),

$$\text{s-depth}(R/L_{(h)}) = \text{s-depth}(R_{n-\ell+h-2}/P_{(\ell,n-\ell+h-2)}) + \ell - h + 1,$$
$$\text{s-depth}(R/J_{(\ell-c)}) = \text{s-depth}(R_{n-c-1}/P_{(\ell,n-c-1)}) + c.$$

As in Theorem 3.10, applying the inductive hypothesis yields

$$\text{s-depth}(R_{n-\ell+h-2}/P_{(\ell,n-\ell+h-2)}) \geq \ell((b-1)+1)+(c+h)-1,$$

so s-depth$(R/L_{(h)}) \geq \ell(b+1)+c$. Also by induction

$$\text{s-depth}(R_{n-c-1}/P_{(\ell,n-c-1)}) \geq \ell(b-1+1)+(\ell+1)-1 = \ell(b+1),$$

so s-depth$(R/J_{(\ell-c)}) \geq \ell(b+1)+c$. Now repeated use of [Rauf 2010, Lemma 2.2] applied to (3-4) yields

$$\text{s-depth } R/I_{(1)} \geq \ell(b+1)+c.$$

Notice that if $c = 0$, we have $P_{(\ell,n)} = I_{(1)}$ and s-depth$(R/P_{(\ell,n)}) \geq \ell(b+1)$ for any b, and the result holds. For $c > 0$, repeated application of [loc. cit., Lemma 2.2] to the sequences in (3-1) yields s-depth$(R/P_{(\ell,n)}) \geq \ell(b+1) + c - 1$. □

A monomial ideal I is a Stanley ideal if the Stanley conjecture holds for I. That is, if s-depth$(R/I) \geq$ depth(R/I). Due to the general difficulty of computing the Stanley depth, very few classes of Stanley ideals are known. It is interesting to note that Theorem 4.1 provides a new class of Stanley ideals.

Corollary 4.2. *Let S_n be a spine of n vertices. Then $P_\ell(S_n)$ is a Stanley ideal.*

Proof. This follows directly from Theorems 3.10 and 4.1. □

Acknowledgements

We gratefully acknowledge the computer algebra system Macaulay 2 [Grayson and Stillman 1996] which was invaluable in our work on this paper. The authors of this paper were supported by NSF grant (DMS 1005206) during the initial phase of research. We thank NSF and Texas State University for their support. We wish to thank the other students and faculty mentors working on this grant for helpful suggestions and ideas during various group discussions. In addition, we thank an anonymous referee for suggestions that improved the readability of the paper.

References

[Bouchat et al. 2011] R. R. Bouchat, H. T. Hà, and A. O'Keefe, "Path ideals of rooted trees and their graded Betti numbers", *J. Combin. Theory Ser. A* **118**:8 (2011), 2411–2425. MR 2012g:13032 Zbl 1232.05089

[Bruns and Herzog 1993] W. Bruns and J. Herzog, *Cohen–Macaulay rings*, Cambridge Studies in Advanced Mathematics **39**, Cambridge University Press, 1993. MR 95h:13020 Zbl 0788.13005

[Campos et al. 2014] D. Campos, R. Gunderson, S. Morey, C. Paulsen, and T. Polstra, "Depths and Cohen–Macaulay properties of path ideals", *J. Pure Appl. Algebra* **218**:8 (2014), 1537–1543. MR 3175038 Zbl 1283.05271

[Cimpoeaş 2009] M. Cimpoeaş, "Stanley depth of monomial ideals with small number of generators", *Cent. Eur. J. Math.* **7**:4 (2009), 629–634. MR 2010j:13039 Zbl 1185.13027

[Dao and Schweig 2013] H. Dao and J. Schweig, "Projective dimension, graph domination parameters, and independence complex homology", *J. Combin. Theory Ser. A* **120**:2 (2013), 453–469. MR 2995051 Zbl 1257.05114

[Dao et al. 2013] H. Dao, C. Huneke, and J. Schweig, "Bounds on the regularity and projective dimension of ideals associated to graphs", *J. Algebraic Combin.* **38**:1 (2013), 37–55. MR 3070118 Zbl 06192152

[Fouli and Morey 2014] L. Fouli and S. Morey, "A lower bound for depths of powers of edge ideals", preprint, 2014. arXiv 1409.7020

[Grayson and Stillman 1996] D. R. Grayson and M. E. Stillman, "Macaulay2, a software system for research in algebraic geometry", 1996, available at http://www.math.uiuc.edu/Macaulay2/.

[Hà and Morey 2010] H. T. Hà and S. Morey, "Embedded associated primes of powers of square-free monomial ideals", *J. Pure Appl. Algebra* **214**:4 (2010), 301–308. MR 2011b:13064 Zbl 1185.13024

[Harary 1969] F. Harary, *Graph theory*, Addison-Wesley, Reading, MA, 1969. MR 41 #1566 Zbl 0182.57702

[He and Van Tuyl 2010] J. He and A. Van Tuyl, "Algebraic properties of the path ideal of a tree", *Comm. Algebra* **38**:5 (2010), 1725–1742. MR 2011e:13028 Zbl 1198.13014

[Herzog and Hibi 2005] J. Herzog and T. Hibi, "The depth of powers of an ideal", *J. Algebra* **291**:2 (2005), 534–550. MR 2006h:13023 Zbl 1096.13015

[Herzog and Hibi 2011] J. Herzog and T. Hibi, *Monomial ideals*, Graduate Texts in Mathematics **260**, Springer, London, 2011. MR 2011k:13019 Zbl 1206.13001

[Kummini 2009] M. Kummini, "Regularity, depth and arithmetic rank of bipartite edge ideals", *J. Algebraic Combin.* **30**:4 (2009), 429–445. MR 2010j:13040 Zbl 1203.13018

[Miller and Sturmfels 2005] E. Miller and B. Sturmfels, *Combinatorial commutative algebra*, Graduate Texts in Mathematics **227**, Springer, New York, 2005. MR 2006d:13001 Zbl 1066.13001

[Morey 2010] S. Morey, "Depths of powers of the edge ideal of a tree", *Comm. Algebra* **38**:11 (2010), 4042–4055. MR 2011m:13039 Zbl 1210.13020

[Pournaki et al. 2009] M. R. Pournaki, S. A. Seyed Fakhari, M. Tousi, and S. Yassemi, "What is ... Stanley depth?", *Notices Amer. Math. Soc.* **56**:9 (2009), 1106–1108. MR 2010k:05346 Zbl 1177.13056

[Pournaki et al. 2013] M. R. Pournaki, S. A. Seyed Fakhari, and S. Yassemi, "Stanley depth of powers of the edge ideal of a forest", *Proc. Amer. Math. Soc.* **141**:10 (2013), 3327–3336. MR 3080155 Zbl 1282.13023

[Rauf 2010] A. Rauf, "Depth and Stanley depth of multigraded modules", *Comm. Algebra* **38**:2 (2010), 773–784. MR 2011g:13029 Zbl 1193.13025

[Stanley 1982] R. P. Stanley, "Linear Diophantine equations and local cohomology", *Invent. Math.* **68**:2 (1982), 175–193. MR 83m:10017 Zbl 0516.10009

[Villarreal 1990] R. H. Villarreal, "Cohen–Macaulay graphs", *Manuscripta Math.* **66**:3 (1990), 277–293. MR 91b:13031 Zbl 0737.13003

[Villarreal 2001] R. H. Villarreal, *Monomial algebras*, Monographs and Textbooks in Pure and Applied Mathematics **238**, Marcel Dekker, New York, 2001. MR 2002c:13001 Zbl 1002.13010

Received: 2014-10-02 Revised: 2014-12-22 Accepted: 2015-01-09

campos.daniell@gmail.com *525 West Mulberry Avenue, San Antonio, TX 78212,*
United States

rgunder@math.duke.edu *Department of Mathematics, Duke University,*
Durham, NC 27708, United States

morey@txstate.edu *Department of Mathematics, Texas State University,*
601 University Drive, San Marcos, TX 78666, United States

chelsey.paulsen@gmail.com *East Chapel Hill High School, 500 Weaver Dairy Road,*
Chapel Hill, NC 27514, United States

thomaspolstra@gmail.com *Department of Mathematics, University of Missouri,*
202 Mathematical Sciences Building, Columbia, MO 65211,
United States

INVOLVE 9:1 (2016)
dx.doi.org/10.2140/involve.2016.9.171

Combinatorics of linked systems of quartet trees

Emili Moan and Joseph Rusinko

(Communicated by Kenneth S. Berenhaut)

We apply classical quartet techniques to the problem of phylogenetic decisiveness and find a value k such that all collections of at least k quartets are decisive. Moreover, we prove that this bound is optimal and give a lower bound on the probability that a collection of quartets is decisive.

1. Overview

Evolutionary biologists represent relationships between groups of organisms with phylogenetic trees. Supertree methods were designed to handle the computationally difficult problem of reconstructing such trees for large data sets. Those methods generate a group of accurate, smaller input trees and combine them into a single supertree. Four-taxa trees, known as quartet trees, are commonly used as inputs in supertree methods.

Most quartet amalgamation algorithms use all quartet trees generated from sequencing data or only remove quartet trees that appear to be incorrect. As quartet trees may contain overlapping information, it is possible that a smaller number of trees may provide sufficient information for accurate reconstruction.

Böcker et al. [1999] developed a sufficient condition for a set of quartet trees to be definitive. For any tree on a taxon set $[n] = \{1, 2, \dots, n\}$, we develop a system of quartet trees that meets the criteria of Böcker et al., known as a linked system. Additionally, we develop collections of linked systems known as meshed systems.

Recently, Steel and Sanderson asked for which collections of sets of taxa do the corresponding induced subtrees determine a unique supertree. They called such collections decisive. The notion of decisiveness can be viewed as a generalization of definitiveness where no information is required about the particular subtrees that the subsets of taxa induce. This notion plays an important role in supertree reconstruction since it a priori addresses the question about which subsets of taxa must be analyzed to ensure that a unique supertree can be reconstructed.

MSC2010: 92B10.
Keywords: phylogenetics, quartets, decisiveness.

We use the term quartet to refer to any four-element taxon subset, and the term quartet tree when referencing a resolved four-taxa tree. Using meshed systems, we find a minimal number $k(n)$ such that every collection of at least k quartets is decisive. We use this number to find a lower bound on the probability that an arbitrary collection of quartets is decisive.

Finally, we find that meshed systems may be useful in amalgamation algorithms, such as maxcut [Snir and Rao 2008], that do not always find the correct supertree when given a definitive system of quartet trees.

2. Linked systems

We adopt the terminology in [Dress et al. 2012], except in noted instances when we follow [Semple and Steel 2003] or [Steel and Sanderson 2010]. Phylogenetic trees display relationships among a finite set of taxonomic units.

Definition 2.1. A *binary phylogenetic tree*, $T = (V, E, \varphi)$ on a finite set of taxa X, is a triple consisting of a finite set of vertices V, a set E of edges between vertices, and a labeling map $\varphi : X \to L$, where $L \subset V$ contains all vertices of degree one, or *leaves*, such that the graph (V, E) is an unrooted binary tree and the map φ induces a bijection between X and the set L of leaves of T.

An edge that contains a leaf is an *exterior edge*. The nonleaf vertex of an exterior edge is the *internal vertex of e*, denoted $v_{\text{int}}(e)$. Two exterior edges sharing an internal vertex form a *cherry*. Any edge that is not an exterior edge is an *interior edge*.

While edge length plays an important role in phylogenetics, we do not take it into account, and adopt instead a topological definition of tree isomorphism.

Definition 2.2. Phylogenetic trees, $T_1 = (V_1, E_1, \varphi_1)$ and $T_2 = (V_2, E_2, \varphi_2)$ on a taxon set X, are *isomorphic* if there exists a bijective map $f : V_1 \to V_2$, called an *isomorphism*, such that if $\{u, v\} \in E_1$ then $\{f(u), f(v)\} \in E_2$ and for every $x \in X$, we have $\varphi_2(x) = f(\varphi_1(x))$.

It is impossible to distinguish phylogenetic relationships from unrooted trees with fewer than four taxa; thus, supertree reconstruction algorithms frequently use four-taxa trees or *quartets trees* as inputs [Snir and Rao 2008; Snir et al. 2008; Strimmer and von Haeseler 1996]. *Quartet trees* are binary phylogenetic trees on four leaves. Such trees are in one-to-one correspondence with two-element subsets of X, such as $\{\{a, b\}, \{c, d\}\}$, according to the separation of the four leaves by the interior edge. The union of all four taxa is the *support of q*, denoted $\text{supp}(q)$.

Quartet trees contain an interior edge which separates the taxa into two pairs. Similarly, removing an interior edge of a tree separates the graph into two connected components. An edge *e separates* taxa a and b from c and d if $\{a, b\}$ and $\{c, d\}$ are subsets of the vertex sets of different connected components of $T - \{e\}$. This separation points to a relationship between edges of a tree and quartet trees. A quartet tree $ab|cd$ is *displayed by a binary phylogenetic tree T* if there exists an edge $e \in E$ that separates a and b from c and d.

Denote the set of all quartet trees on a taxon set X by $Q(X)$. Any subset Q of $Q(X)$ is called a *system of quartet trees* on X with the support defined by $\operatorname{supp}(Q) = \bigcup_{q \in Q} \operatorname{supp}(q)$. Additionally, we denote the set of all quartet trees displayed by a tree T by Q_T. A system of quartet trees Q is *compatible* if there exists a tree T such that $Q \subseteq Q_T$.

Definition 2.3 [Semple and Steel 2003]. Let $T = (V, E, \varphi)$ be a binary phylogenetic tree and let $ab|cd \in Q_T$. An interior edge e of T is *distinguished* by $ab|cd$ if e is the only edge that separates a and b from c and d.

Quartet trees which distinguish edges are a powerful input to quartet amalgamation algorithms. These algorithms must handle noncompatible systems of quartet trees. However, even compatible systems may be difficult to resolve as multiple trees may display a particular collection of quartet trees.

Definition 2.4 [Steel and Sanderson 2010]. A system of quartet trees Q is *definitive* if up to isomorphism, there is a unique binary phylogenetic tree T for which $Q \subseteq Q_T$.

Böcker et al. [1999] described various criteria for a system of quartet trees of size $n - 3$ to be definitive. We construct systems of quartet trees that meet this criteria and make note of some useful applications of these systems.

Proposition 2.5 [Böcker et al. 1999, Example 3.7]. *If T is a binary tree such that the interior edges of T are labeled $E = \{e_1, \dots, e_{n-3}\}$, and Q is a system of quartet trees such that each $q_i \in Q$ distinguishes a unique edge e_i in T with*

$$\left| \operatorname{supp}(q_i) \setminus \bigcup_{j<i} \operatorname{supp}(q_j) \right| = 1$$

for $i = 2, \dots, n - 3$, then Q is definitive.

We create a system of quartet trees that satisfies the hypotheses of Proposition 2.5, known as a linked system, by imposing an ordering on the interior edges of a tree and the quartet trees which distinguish those edges. We define linked systems in terms of the associated graph.

Definition 2.6. For a compatible system of $n - 3$ quartet trees Q on a taxon set X, define the associated graph $G_T(Q)$ with vertex set V and edge set E as follows:

- The vertex set V is the set of all quartet trees $q \in Q$ which distinguish a unique edge in T.
- Vertex pairs $\{q_i, q_j\}$ are connected by an edge $e \in E$ if the edge e_i that q_i distinguishes is adjacent to the edge e_j that q_j distinguishes and $|\operatorname{supp}(\{q_i, q_j\})| = 5$.

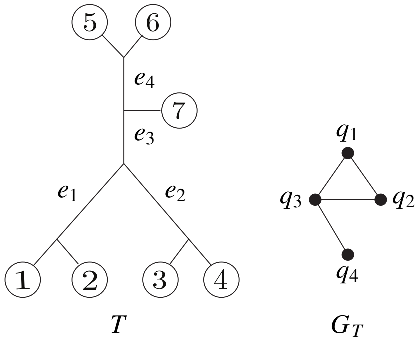


Figure 1. A binary phylogenetic tree T and the associated graph $G_T(Q)$ for the quartet trees $q_1 = 12|35$, $q_2 = 34|15$, $q_3 = 57|13$ and $q_4 = 56|71$.

Definition 2.7. Two quartet trees are *linked* if their vertices are connected in $G_T(Q)$. The system of quartet trees Q is *a linked system* if $G_T(Q)$ is connected. See Figure 1 for an example.

In Section 3 we prove that linked systems are definitive. Linked systems also help illuminate the broader concept of phylogenetic decisiveness, which we review here.

For a binary phylogenetic tree T and a subset Y of X, let $T|Y$ denote the induced binary phylogenetic tree on leaf set Y (the tree obtained from the minimal subtree connecting Y by suppressing any vertices of degree 2). Let $\mathcal{S}$ be the collection of subsets of a set X of size four; we refer to all such subsets as *quartets* [Steel and Sanderson 2010].

Definition 2.8 [Steel and Sanderson 2010]. We say that $\mathcal{S}$ is *phylogenetically decisive* if it satisfies the following property: if T and T' are binary phylogenetic trees, with $T|Y = T'|Y$ for all $Y \in \mathcal{S}$, then $T = T'$.

We will use collections of linked systems to find the minimal number $k(n)$ such that a collection of at least k quartets is phylogenetically decisive.

3. Applications of linked systems

We first show that linked systems meet the criteria of [Böcker et al. 1999] for defining a unique tree.

Theorem 3.1. *Every linked system of quartet trees is definitive.*

Proof. Let Q be a linked system of quartet trees on a tree T on a taxon set X. Linked systems are of size $n-3$ and each quartet tree distinguishes a unique edge in T. Let T be a binary phylogenetic tree on a taxon set X and let e_1 be an interior edge adjacent to a cherry. The tree is connected, which implies we can label the remaining interior edges $\{e_2, \dots, e_{n-3}\}$ such that e_j is adjacent to some e_i with $i < j$. Moreover, because the support of each pair of quartet trees $\{q_i, q_j\}$ that distinguishes adjacent edges $\{e_i, e_j\}$ is five, each pair of quartet trees shares three taxa and each additional quartet tree in Q adds one new taxon to the support of Q. Thus, linked systems meet the criteria in Proposition 2.5 and are definitive. □

Though all linked systems are definitive, we find that not all definitive systems of size $n-3$ are linked.

Example. The system of quartet trees $Q = \{12|36, 23|45, 24|56\}$ meets the criteria established in Proposition 2.5, and is thus definitive. The graph $G_T(Q)$ contains the three vertices q_1, q_2, and q_3, where q_2 and q_3 are connected by an edge and q_1 is an isolated vertex. Thus, Q is not a linked system.

Since a system satisfying Proposition 2.5 and linked systems both contain $n-3$ quartet trees, one might surmise that all compatible systems of quartet trees of a modest size would be definitive. However, there are large systems of compatible quartet trees which are not definitive and large collections of quartets which are not decisive. A collection of quartets and the induced quartet trees on a caterpillar tree provides one such example.

Definition 3.2 [Semple and Steel 2003]. A *caterpillar* on n leaves is a binary phylogenetic tree for which there exists an induced subtree on a sequence of distinct interior vertices $v_1, v_2, \dots, v_k$ such that, for all $i \in \{1, 2, \dots, k-1\}$, v_i and v_{i+1} are adjacent.

The ordering of vertices in a caterpillar tree induces an ordering of interior edges e_i, where e_i connects v_i with v_{i+1}. We use this ordering to construct large families of quartet trees shared by several caterpillar trees.

Theorem 3.3. *The minimal number $k(n)$ such that every collection of quartets S with $|S| \geq k$ is decisive is greater than $\binom{n}{4} - (n-3)$.*

Proof. Let $\{a, b, c\}$ be a subset of X. We define T_1, T_2, and T_3 to be three distinct caterpillar trees of size $n \geq 4$ that differ only in the placement of three taxa a, b, and c such that for each tree, $v_{\text{int}}(a)$, $v_{\text{int}}(b)$ and $v_{\text{int}}(c)$ are incident to e_1. Denote by $\mathcal{S}$ the set of $n-3$ sets

$$\bigcup_{i=1}^{n-3} Y_i = \{a, b, c, y \mid y \in X - \{a, b, c\}\}$$

and let $\mathcal{S}'$ be the complement of $\mathcal{S}$. We observe that for all $Y \in \mathcal{S}'$, we have $T_1|Y = T_2|Y = T_3|Y$, but $T_1 \neq T_2 \neq T_3$. Therefore, S' is not decisive. Since $|\mathcal{S}'| = \binom{n}{4} - (n-3)$, the minimal number $k(n)$ such that every collection $\mathcal{S}$ of quartets with $|\mathcal{S}| \geq k$ is decisive is greater than $\binom{n}{4} - (n-3)$. □

To show that sets of quartets of size $\binom{n}{4}-(n-3)$ are decisive, we prove that Q_T contains at least $n-3$ disjoint linked systems, ensuring the removal of any $n-4$ quartet trees from a compatible system would leave at least one linked system. We introduce a process for building such systems by using a seed quartet tree which distinguishes an edge of a tree, and systematically constructing additional quartet trees which distinguish the same edge.

In a phylogenetic tree, each interior edge $e=(v_l, v_r)$ is adjacent to four edges e_i, e_j, e_h, and e_k, which divide the tree into four components and partition the set of taxa X into four distinct sets A_i, A_j, A_k, and A_h, with $x \in A_n$ if the unique path from x to v_l contains the edge e_n.

Definition 3.4. Let $q=ij\,|\,kh$ be a quartet tree that distinguishes an edge e and let $i \in A_i$, $j \in A_j$, $k \in A_k$, and $h \in A_h$, where A_i, A_j, A_k, and A_h are partitions of X induced by e. For $x \in X-\operatorname{supp}(q)$, define the *quartet-tree substitution* $q(x)$ to be the unique quartet tree in which the taxon $x \in A_n$ replaces the taxon in q that is in $\operatorname{supp}(q)\cap A_n$.

Notice $q(x)$ and q must distinguish the same edge of the tree.

Definition 3.5. Let the quartet tree q distinguish an edge e of a tree. Define the *vine of* q by $v(q)=\{q\}\cup\bigcup_{x\in X-\operatorname{supp}(q)} q(x)$. We refer to q as the *seed* of the vine.

The following shows that if two quartet trees are linked, then so are their vines.

Definition 3.6. Two vines $v(q_i)$ and $v(q_j)$ are *linked* if for each $q_i \in v(q_i)$ there exists a unique $q_j \in v(q_j)$ such that q_i and q_j are linked.

Theorem 3.7. *If q_i and q_j are linked quartet trees, then the associated vines $v(q_i)$ and $v(q_j)$ are linked.*

Proof. Assume that $\{q_i, q_j\}$ are the seeds of the adjacent edges e_i and e_j and are linked in T. Let $v(q_i)$ and $v(q_j)$ be the associated vines.

Since $\operatorname{supp}(q_i, q_j)=5$, each quartet tree contains one taxon that the other does not. Let z be the taxon in $\operatorname{supp}(q_j)-\operatorname{supp}(q_i)$ and y be the taxon in $\operatorname{supp}(q_i)-\operatorname{supp}(q_j)$. Use quartet-tree substitution to construct the quartet trees $q_i(z)$ and $q_j(y)$. By construction, $\operatorname{supp}(q_i, q_j)=\operatorname{supp}(q_i(z), q_j(y))$ and $q_i(z)$ and $q_j(y)$ are linked.

Use quartet-tree substitution with each remaining taxon $x \in X-\operatorname{supp}(q_i, q_j)$ on q_i and q_j to construct the remaining quartet trees in $v(q_i)$ and $v(q_j)$. In the construction of each $\{q_i(x), q_j(x)\}$, one taxon (Case 1) or two taxa (Case 2) are removed and one taxon x is introduced. Thus, for $x \in X-(\operatorname{supp}(q_i, q_j))$, we have $4 \le |\operatorname{supp}(q_i(x), q_j(x))| \le 6$.

In Case 1, x replaces the same taxon in q_i and q_j and $\operatorname{supp}(q_i, q_j)$ remains the same.

In Case 2, x replaces one taxon in q_i and a different taxon in q_j.

Assume that x replaces two different taxa in $\operatorname{supp}(q_i, q_j)-\operatorname{supp}(q_i \cap q_j)$. Then, $|\operatorname{supp}(q_i(x), q_j(x))|=4$ and $q_i(x)=q_j(x)$. This is not possible since

q_i and q_j distinguish different edges. Thus, x does not replace two different taxa in $\text{supp}(q_i, q_j) - \text{supp}(q_i \cap q_j)$ and $|\,\text{supp}(q_i(x), q_j(x))| \neq 4$.

Now assume that x replaces two different taxa in $\text{supp}(q_i \cap q_j)$. Then, we have that $|\,\text{supp}(q_i(x), q_j(x))| = 6$. Recall that q_i and q_j distinguish the edges e_i and e_j. Thus, in order for x to replace two different taxa in $\text{supp}(q_i \cap q_j)$, x would have to be in two different sets of the partition that e_i induces on X. This is not possible because x cannot be in two different sets of a partition. Thus, x does not replace two different taxa in $\text{supp}(q_i \cap q_j)$ and $|\,\text{supp}(q_i(x), q_j(x))| \neq 6$.

Thus, in this case, x replaces one taxon in $\text{supp}(q_i \cap q_j)$ and one taxon in $\text{supp}(q_i, q_j) - \text{supp}(q_i \cap q_j)$ and $|\,\text{supp}(\{q_i(x), q_j(x)\})| = 5$. Therefore the vines $v(q_i)$ and $v(q_j)$ are linked. □

A linking between vines allows us to construct multiple disjoint linked systems of quartet trees. We refer to these systems as *meshed systems* and use them to show that any set of quartets of sufficient size is decisive.

Definition 3.8. A *meshed system* on a tree T with taxon set X is an $(n-3)$ by $(n-3)$ array of quartet trees, where each row is a linked system and each column is a vine.

Note that the existence of a meshed system ensures that the removal of up to $n-4$ quartet trees from Q_T must leave at least one definitive set.

Theorem 3.9. *For any binary phylogenetic tree T on a taxon set X, the system Q_T of all quartet trees displayed by T contains a meshed system.*

Proof. Let T be a binary phylogenetic tree on a taxon set X and let e_1 be an interior edge adjacent to a cherry. The tree is connected, which implies we can label the remaining interior edges $\{e_2, \dots, e_{n-3}\}$ such that e_j is adjacent to some e_i with $i < j$.

Let e_j be adjacent to e_i with $i < j$. We know that e_i separates T into two connected components T_i^a and T_i^b. Moreover, e_j separates T into two connected components T_j^a and T_j^b. Since e_i and e_j are adjacent, $\text{supp}(T_i^b) \cap \text{supp}(T_j^a) \neq \varnothing$. Let $q_i = ab|cd$ and $q_j = ac|de$ such that $a \in \text{supp}(T_i^a)$, $c \in \text{supp}(T_j^b)$, and $d \in \text{supp}(T_i^b) \cap \text{supp}(T_j^a)$. Thus, q_i and q_j are linked for all e_i and e_j in T, and we have a linked system that makes up the first row of our matrix.

Using quartet-tree substitution, construct vines $v(q_i)$ and $v(q_j)$. By Theorem 3.7 the vines $v(q_j)$ and $v(q_i)$ are linked. Thus, we have $n-3$ disjoint columns of quartet trees in our matrix. Additionally, for each pair of linked quartet trees $\{q_i, q_j\}$ in row one, there exists a pair $\{q_i(x), q_j(x)\}$ in the remaining rows of the matrix that are linked. Thus, we have $n-3$ rows of linked systems.

Therefore, Q_T contains a meshed system. □

The existence of a meshed system allows us to find the minimal number, $k(n)$, such that every collection $\mathcal{S}$ of quartets with $|\mathcal{S}| \geq k$ is decisive.

Theorem 3.10. *The number $k(n) = \binom{n}{4} - (n-4)$ is the smallest number such that every collection of quartets $\mathcal{S}$ on a taxon set $X = [n]$ with $\mathcal{S} \geq k$ is decisive.*

Proof. Let $\mathcal{S}$ be a collection of quartets on a taxon set $X = [n]$ with $|\mathcal{S}| \geq \binom{n}{4} - (n-4)$. Let T and T' be two phylogenetic trees such that $T\,|\,Y = T'\,|\,Y$ for all $Y \in \mathcal{S}$. We define $Q \subset Q_T$ to be the collection of $T\,|\,Y$ for all $Y \in \mathcal{S}$. By Theorem 3.9, Q_T contains a meshed system M. By the pigeon hole principle, if $|Q| \geq \binom{n}{4} - (n-4)$ then Q must contain one of the linked systems in M, and by Theorem 3.1, Q is definitive. Thus, T is the unique tree which displays Q. However, since Q is also $T'\,|\,Y$ for all $Y \in \mathcal{S}$, we must have $T = T'$. Therefore, $\mathcal{S}$ is decisive. Moreover, Theorem 3.3 shows that $k \geq \binom{n}{4} - (n-4)$. Therefore $k(n) = \binom{n}{4} - (n-4)$ is the minimal number such that every collection of quartets with $|\mathcal{S}| \geq k$ is decisive. □

In addition to establishing requirements for collections of subsets of $[n]$ to be decisive, [Steel and Sanderson 2010] provides a formula for the probability that a particular collections of subsets of $[n]$ will be decisive for an arbitrarily sampled phylogenetic tree. In this section, we prove a similar result by finding a lower bound for the probability that a collection of subsets of $[n]$ of a particular size will be phylogenetically decisive. This bound is independent of the underlying tree topology.

Theorem 3.11. *The probability $p(X, k)$ that an arbitrary collection of k quartets is decisive has the property*

$$p(X, k) \geq \frac{\sum_{i=1}^{n-3} (-1)^{i+1} \binom{n-3}{i} \binom{|Q_T| - i(n-3)}{|Q_T| - k}}{\binom{|Q_T|}{k}}.$$

Proof. Let $\mathcal{S}$ be a collection of k quartets. Let T and T' be two phylogenetic trees such that $T\,|\,Y = T'\,|\,Y$ for all $Y \in \mathcal{S}$. We define $Q \subset Q_T$ to be the collection of $T\,|\,Y$ for all $Y \in \mathcal{S}$. Following Theorem 3.10, if Q contains a definitive set of quartet trees, then S is decisive. By Theorem 3.1, if a collection of compatible quartet trees contains a linked system of quartets, then it is definitive. Thus, the probability that a collection $\mathcal{S}$ is decisive is at least the probability that Q contains one of the $n-3$ disjoint linked systems constructed in Theorem 3.9. The formula follows from applying the inclusion-exclusion principle to count the number of subsets of size k which contain one of the disjoint systems of linked quartets. □

To illustrate the utility of the formula, we express the lower bound probability versus the number of quartets selected in Figure 2. In Figure 3, we plot the number of quartets required to ensure a fixed accuracy as a power of n. Notice the number of quartets needed to ensure the sample is decisive with accuracy of 25% is on the order of n^c with $c \sim 3.3$ and is almost indistinguishable from the number required to ensure 99% accuracy.

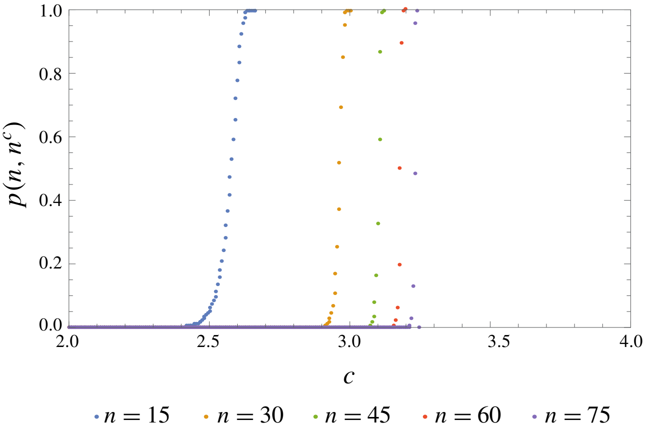


Figure 2. A lower bound on the probability that a set of n^c quartets is decisive.

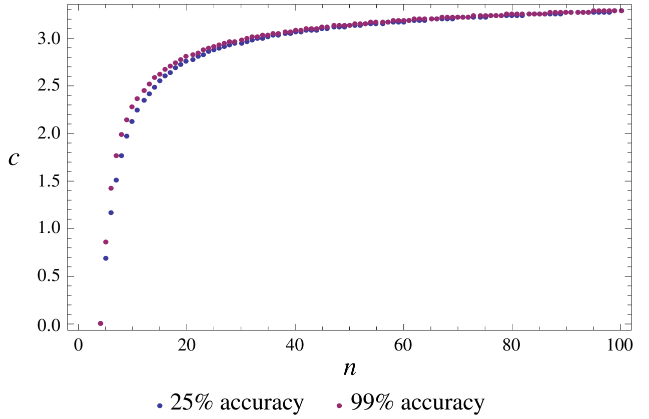


Figure 3. For $|\mathcal{S}| = n^c$, this shows the size of compatible quartets as a power of n required to ensure a decisive subset with fixed probability.

4. Conclusion

Using the criteria established in [Böcker et al. 1999], we have developed a new type of definitive system of quartet trees, linked systems. We have also developed groups of linked systems, known as meshed systems. We have used meshed systems to show that the number of quartets required to ensure decisiveness is on the order of $O(n^4)$. Moreover, we have used meshed systems to show the probability that an arbitrary collection of quartets contains a decisive system. These results lend credence to sampling quartets on the order of $n^{3.3}$.

It has been suggested that smaller sets of representative quartet trees will play a crucial role in developing efficient scalable supertree methods, as the use of all quartet tree samples may be computationally inefficient [Swenson et al. 2011]. Thus, linked systems may be useful inputs in such algorithms. However, some supertree methods, such as quartets maxcut, do not always return a fully resolved tree even when the input sets contain small definitive systems. For example, maxcut does not return a fully resolved tree for the linked system $Q_1 = \{12|35, 13|45, 14|56\}$, but returns the correct tree for the meshed system $M = \{Q_1, Q_2, Q_3\}$, where $Q_2 = \{12|34, 23|45, 24|56\}$ and $Q_3 = \{12|36, 23|46, 34|56\}$. Therefore, we anticipate that both linked and meshed systems will serve as efficient inputs for future supertree algorithms, as these algorithms could be reformulated to emphasize small definitive units.

Acknowledgements

Both authors were supported by grants from the National Center for Research Resources (5 P20 RR016461) and the National Institute of General Medical Sciences (8 P20 GM103499) from the National Institutes of Health. We would also like to thank Dr. Mike Steel for introducing us to the concept of decisiveness and giving input throughout the writing process.

References

[Böcker et al. 1999] S. Böcker, A. W. M. Dress, and M. A. Steel, "Patching up X-trees", *Ann. Comb.* **3**:1 (1999), 1–12. MR 2001d:05038 Zbl 0933.05039

[Dress et al. 2012] A. Dress, K. T. Huber, J. Koolen, V. Moulton, and A. Spillner, *Basic phylogenetic combinatorics*, Cambridge University Press, 2012. MR 2893879 Zbl 1298.92008

[Semple and Steel 2003] C. Semple and M. Steel, *Phylogenetics*, Oxford Lecture Series in Mathematics and its Applications **24**, Oxford University Press, 2003. MR 2005g:92024 Zbl 1043.92026

[Snir and Rao 2008] S. Snir and S. Rao, "Quartets maxcut: a divide and conquer quartets algorithm", *IEEE/ACM Trans. Comput. Biol. Bioinform.* **7**:4 (2008), 701–718.

[Snir et al. 2008] S. Snir, T. Warnow, and S. Rao, "Short quartet puzzling: a new quartet-based phylogeny reconstruction algorithm", *J. Comput. Biol.* **15**:1 (2008), 91–103. MR 2009b:92035

[Steel and Sanderson 2010] M. Steel and M. J. Sanderson, "Characterizing phylogenetically decisive taxon coverage", *Appl. Math. Lett.* **23**:1 (2010), 82–86. MR 2011c:05351 Zbl 1181.92068

[Strimmer and von Haeseler 1996] K. Strimmer and A. von Haeseler, "Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies", *Mol. Biol. Evol.* **13**:7 (1996), 964–969.

[Swenson et al. 2011] M. S. Swenson, R. Suri, C. R. Linder, and T. Warnow, "An experimental study of quartets maxcut and other supertree methods", *Algorithms for Molecular Biology* **6**:7 (2011).

Received: 2014-11-17 Revised: 2015-01-14 Accepted: 2015-02-02

pricee4@winthrop.edu
Department of Mathematics, Winthrop University,
Rock Hill, SC 29733, United States

rusinkoj@winthrop.edu
Department of Mathematics, Winthrop University,
Rock Hill, SC 29733, United States

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the Submission page at the Involve website.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles in *Involve* are usually in English, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not make any reference to the bibliography. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and subject classifications for the article, and, for each author, postal address, affiliation (if appropriate), and email address.

Format. Authors are encouraged to use $\LaTeX$ but submissions in other varieties of $\TeX$, and exceptionally in other formats, are acceptable. Initial uploads should be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of Bib$\TeX$ is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages (Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc.) allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with details about how your graphics were generated.

White space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

involve

2016 vol. 9 no. 1

1944-4176(2016)9:1;1-4