# Research Experiences for Undergraduates

# Student Reports

**INDIANA UNIVERSITY**

**Summer 2009**

**Bloomington, Indiana**

# The Summer 2009 REU Program At Indiana University

During the summer of 2009 eleven students participated in the Research Experiences for Undergraduates program in Mathematics at Indiana University. The program ran for eight weeks, from June 22 through August 14. Ten faculty served as research advisers. One faculty member oversaw a pair of related projects; all other faculty advised one student each.

The program opened with an introductory pizza party. On the following morning, students began meeting with their faculty mentors; these meetings continued regularly throughout the first few weeks. During week one, there were short presentations by faculty mentors briefly introducing the problem to be investigated. Students also received orientations to the mathematics library and to our computing facilities. In week three, students gave short, informal presentations to each other on the status of work on the project. Brief training sessions on using LaTeXwere given during week four. During week six, we hosted the Indiana Mathematics Undergraduate Research conference, which featured 22 lectures by 34 students from Rose-Hulman Institute of Technology, Goshen College, Wabash College, and Indiana University. The program concluded with the students giving formal, hourlong presentations to the REU students and faculty, and the turning in of final reports, contained in this volume.

It took the help and support of many different groups and individuals to make the program a success.

We thank the National Science Foundation for major financial support through the REU program. We also thank the College of Arts and Sciences for crucial additional funding. We thank Indiana University for the use of facilities, including library, computers, and recreational facilities. We thank the staff of the Department of Mathematics for support, especially Mandie McCarty for coordinating the complex logistical arrangments (housing, paychecks, information packets, meal plans, etc.) and Cheryl Miller for her assistance in coordinating the application process. We thank Indiana graduate student Kevin Meek for serving as LaTeXconsultant and for compiling this volume.

We thank Professors Eric Bedford, Richard Bradley, Allan Edmonds, David Hoff, Charles Livingston, and Matthias Weber for volunteering to give lectures on their favorite topics during the program. We also thank Professor Jee Koh for his plenary lecture at the Indiana Mathematics Undergraduate Research conference.

This program could not exist without the faculty mentors, whose expertise and generous donation of time and energy enabled our participants to have a truly exceptional experience. A special thanks to the professors who led research projects: Eric Bedford, Chris Connell, Allan Edmonds, Matthew Hahn (biology), Elizabeth Housworth, Nets Katz, Michael Lynch (biology), Peter Ortoleva (chemistry), Kevin Pilgrim, and Matthias Weber.

Kevin M. Pilgrim

# REU Participants Summer 2009

From left to right:

Jacek Skryzalin
Zachary Norwood
Adam Abegg
Zachary Doenges
John Brown
Ari Nachison
Joseph Thurman
Ziva Meyer
Leah Wolberg
Chengcheng Yang
(not shown: Jamil Merali)

**REU Participants**
**Summer 2009**

| | |
|---|---|
| Adam Abegg | St. Louis University |
| John Brown | Indiana University |
| Zachary Doenges | Indiana University |
| Jamil Merali | Northwestern University |
| Ziva Meyer | New College of Florida |
| Ari Nachison | University of California, Santa Barbara |
| Zachary Norwood | University of Nebraska, Lincoln |
| Jacek Skryzalin | Indiana University |
| Joseph Thurman | Vanderbilt University |
| Leah Wolberg | Bowdoin College |
| Chengcheng Yang | Rice University |

**Faculty Advisors**

Eric Bedford
Chris Connell
Allan Edmonds
Matthew Hahn
Elizabeth Housworth
Nets Katz
Michael Lynch
Peter Ortoleva
Kevin Pilgrim
Matthias Weber

# Contents

# The Mutational Rate of Emergence of Complex Adaptations

Adam Abegg
St. Louis University

Indiana University REU Summer 2009
Advisor: Michael Lynch

## A.1 Introduction

Mutations are rare events, so the emergence of complex adaptations is expected to take a long time, especially in small populations. Larger populations provide more individual targets for mutational origin, so intuitively the time required for the establishment of mutants is less; however, should the intermediate steps toward a complex adaption be selectively disadvantageous, the larger population size may inhibit adaptational advance, due to the increased efficiency of selection against deleterious intermediate mutants.

Unfortunately, a number of factors may magnify the rate of emergence of complex adaptations in ways that defy these simple expectations. For example, small populations–owing to the reduced efficiency of selection–are vulnerable to an increase in the mutation rate resulting from the accumulation of mutations with mild effects on the efficiency of DNA replication- and repair- loci.[1] Such an increase in the per-capita mutation rate, which is consistent with the known increase in the per-generation mutation rate in unicellular eukaryotes relative to prokaryotes and in multicellular species relative to unicellular species could offset the decline in the number of individual mutational targets in small populations.[2] On the other hand, in large populations, despite the short persistence time of deleterious intermediate-stage mutations, the steady mutational input of the latter will result in a maintenance of a small, stable reservoir of such alleles by selection-mutation balance. Moreover, this general principle extends to the loci that directly influence the mutation rate, ensuring that there will always be a pool of individuals with mutation rates elevated above the population norm. While seemingly maladaptive at the individual level, the individuals residing in this small segment of the population might be a major source of evolutionary novelties. Finally, in extremely large populations, the possibility exists for double mutants to vault a fitness valley in a single bound, avoiding the price of deleterious intermediates altogether.

Because some have questioned whether conventional mutational mechanisms and current principles of population genetics can adequately explain the emergence of complex adaptations on reasonable evolutionary time scales,[3] there is a need to fully incorporate the above-mentioned complexities into a more comprehensive framework for understanding the population-genetic environments in which complex adaptations are most likely to emerge. In the following pages, the beginnings of such a framework are outlined, showing that when natural levels of heterogeneity in the mutation rate are taken into consideration, adaptations involving multiple mutation steps can emerge in populations of small to intermediate size at rates that can be orders of magnitude more rapid than current theoretical expectations.

## A.2 The Model

The focus throughout will be on diploid, Wright-Fisher structured populations, with segregation of completely linked chromosomes occurring every generation. One locus, harboring two potential alleles, are assumed to govern the mutation rate. A second set of loci are targets of natural selection, and shall be (aptly) referred to as the adaptive loci. **M** and **m** denote alternative alleles at the mutator locus, carrying with them three distinct mutation rates, denoted $u_0$, $u_1$, and $u_2$ for respective genotypes **m**/**m**, **M**/**m** (or, possibly, **m**/**M**–hereafter either type will be denoted strictly as **M**/**m**), and **M**/**M**. (Notice the subscript denotes the number of mutant mutator alleles in the genotype.) Under the model, the fraction of mutant (**M**) alleles produced per generation by **m**/**m** genotypes is $u_0$, whereas those produced by **M**/**m** and **M**/**M** are $(1-u_1)/2$ and 1.0, respectively. Back mutations (**M** $\rightarrow$ **m**) are assumed to occur at a negligible rate. The

[1] Lynch (2008)
[2] Lynch (2007)
[3] e.g., Pigliucci (2007)

mutant allele is also assumed to confer a fitness disadvantage, such that the fitnesses of the **m**/**m**, **M**/**m**, and **M**/**M** genotypes are 1.0, $1.0 - h_M s_M$, and $1.0 - s_M$ respectively, which are consistent with standard diploid selection models.

Starting with the simplest case, an adaptive locus with two independently mutating sites is denoted by any combination of capital and lowercase **A**s and **B**s (e.g., **Ab**/**ab**). Initially, both sites are assumed to be fixed in the ancestral state whose genotypic fitness is 1.0. The fitnesses of alternative genotypes involving the two sites are assumed to be additively determined, such that alleles with single mutants at either site have a reduction in fitness equal to $s_1 \geq 0$, and alleles with mutants at both sites have an increment in fitness equal to $s_2 \geq 0$. Under this scheme, denoting mutants with upper-case letters, the fitnesses of (unordered) genotypes **Ab**/**ab** and **aB**/**ab** are $1.0 - s_1$, of **Ab**/**Ab**, **aB**/**aB**, and **Ab**/**aB** are $1.0 - 2s_1$, of **Ab**/**AB** and **aB**/**AB** are $1.0 - s_1 + s_2$, and of **AB**/**AB** is $1.0 + 2s_2$. The overall fitness of an individual is the product of the fitnesses at the mutator and adaptive loci (e.g., **MAB**/**mAb** has fitness $(1.0 - h_M s_M)(1.0 - s_1 + s_2)$).

Because the adaptive mutations are specific to two individual sites, the rates of production of alternative alleles at this locus are assumed to equal the background per-locus mutation rates (defined by the genotypes at the mutator locus) times a constant $k \leq 1$. Akin to the mutator locus, we assume no back mutations. Throughout we assume mutation rates and effective population sizes ($N$) that are well within the limits of existing biological observations.[4]

The following results are based on stochastic computer simulations of the mean time to fixation of a novel adaptation, generally relying on 200 independent evaluations for any given set of parameters. Prior to the initiation of any specific population trajectory, the frequencies of the alleles at the mutator locus were set to their expectations in the absence of selection on the adaptive locus. For situations in which the heterozygous disadvantage of the mutator allele was smaller than the power of random genetic drift, $1/(2N)$, the mutator-allele frequency was set equal to 1.0, and when the power of drift was smaller than the selective disadvantage, the initial mutator-allele frequency was set equal to the expected value under selection-mutation balance.[5]

## A.3 Two Allele Case: No 'M'

We start with the simplest case, where $u_0 = u_1 = u_2$, so as to get the results for the classical situation, where the common mutation rate is denoted only by $u$. For the special situation in which the intermediate state at the adaptive locus is neutral ($s_1 = 0$), some fairly simple analytical approximations are attainable, with three different domains of behavior, depending upon the effective population size.

First, if the population is sufficiently small in size, the evolutionary dynamics will proceed in a two-step process, with a one step mutant (**Ab**) becoming fixed prior to the arrival of a second step mutation. Starting with the population fixed with frequency 1.0 at **ab**/**ab**, $4Nuk$ first-step mutations arise per generation (the four, instead of two, because either **A** or **B** mutations can arise in this step with equal probability), each with fixation probability $1/(2N)$,[6] so the mean arrival time of the first first-step mutation destined to fix is the reciprocal of the product, $1/2uk$ generations. Because the average time to fixation of a neutral allele is $4N$ generations under the Wright-Fisher model, and the rate of origin of second-step mutations is $\leq 2Nuk$ per generation, it is clear that there is a negligible chance of a second-step mutation arising prior to fixation of a first-step mutation if $(4N)(2Nuk) \ll 1$, or equivalently if $N \ll 1/(2\sqrt{2uk})$. Letting

[4] Lynch (2007)

[5] See solution of equation (6) in Lynch (2008)

[6] For validation, notice $\lim_{s_1 \to 0} p_f(s_1) = 1/(2N)$ for $p_f(s_1)$ defined by equation (3.1)

$$p_f(s_2) = \frac{1 - e^{-2s_2}}{1 - e^{-4Ns_2}} \tag{A.1}$$

denote the probability of fixation of a beneficial (second-step) mutation, then for mutations that fix sequentially, the rate of appearance of the first double mutant destined to fix is the reciprocal of the sum of the average arrival times of the two mutations,

$$r_s = \frac{2uk}{1 + [1/(np_f s_2)]}. \tag{A.2}$$

The mean time to complete establishment of the double mutant is

$$\bar{t}_f = \frac{1}{r_s}, \tag{A.3}$$

ignoring the time for the second mutation to fix, which for small $N$ is negligible compared to the arrival times of mutations.

Second, provided $N > 1/(2\sqrt{2uk})$, there is a significant chance that a beneficial second-step mutation will arise on a descendant of a first-step mutation prior to the fixation of the latter. Nearly all first-step mutations ($1 - 1/(2N)$ of neutral mutants) are destined to be lost by drift, but this process can occasionally rescue such mutants, propelling them to fixation by positive selection. The process in which the beneficial double mutants go to fixation prior to the population ever achieving a pure one-step state has been called stochastic tunneling by Komarova et al. (2003) and Iwasa et al. (2004) in the context of cancer development. The rate of tunneling with a neutral intermediate worked out by these authors via the Moran model is readily extended to the current case. After accounting for diploidy and the two-fold reduction in the rate of drift with the Wright-Fisher model, the rate of appearance of the first double mutant destined to fixation by tunneling becomes

$$r_t \simeq 4Nuk\sqrt{ukp_f(s_2)}. \tag{A.4}$$

Accounting for both paths (which are taken to be mutually exclusive events), the mean number of generations until the establishment of the double mutant is

$$\bar{t}_f \simeq \frac{1}{r_s + r_t}. \tag{A.5}$$

Third, for $N > 1/(4uk)$, the system begins to behave in an effectively deterministic fashion, with the expected frequency of the **AB** allele in generation $t$ being $\sim (ukt)^2$. The probability of fixation of a beneficial mutation is essentially equal to 1.0 once the frequency exceeds $1/(4Ns_2)$. Solving for the time required for the **AB** allele to reach this point yields a mean time to establishment of the double mutant of

$$\bar{t}_f \simeq \frac{1}{2uk\sqrt{Ns_2}}. \tag{A.6}$$

When intermediate (first-step) alleles are disadvantageous, for populations sufficiently small that the fixation proceeds only in a sequential manner, the mean time to establishment is

$$r_s \simeq \frac{2Nuk}{\dfrac{1}{2p_f(s_1)} + \dfrac{1}{p_f(s_2)}}, \tag{A.7}$$

where $p_f(s_1)$, the probability of fixation of a newly arisen first-step mutation, is obtained by substituting $-s_1$ for $s_2$ in equation (3.1). At population sizes large enough to prevent fixation of first-step mutations ($4Ns_1 \gg 1$), the latter are expected to rapidly approach the low frequency maintained under selection-mutation balance, providing a launching pad for beneficial second-step alleles. Under these conditions, the mean arrival time of the first beneficial allele to fix by tunneling is obtainable from Iwasa et al. (2004) yields

$$r_f \simeq \frac{4N(uk)^2(1-s_1)p_f(s_2)}{s_1}. \tag{A.8}$$

Again, the mean time of establishment is given by equation (3.5).

## A.4 Two Allele Case with Evolving Mutation Rates

We now turn to the situation where the mutation rate is allowed to evolve. In this case, at population sizes that are sufficiently small ($4Ns_M \ll 1$), selection is incapable of preventing the fixation of the mutant mutator allele. Thus, up to an approximate threshold of $N = 1/(4s_M)$, the mean time to establishment can be obtained using all of the preceding expressions with $u_2$ used as the mutation rate.

In the model applied here, which ignores $\mathbf{M} \rightarrow \mathbf{m}$ back mutations, the $\mathbf{M}$ allele must ultimately fix. However, for $4Ns_M \gg 1$, the probability of such fixation is negligibly small on reasonable biological time scales (assuming $s_M \gg u_0$), so the mutator allele will generally be maintained at levels defined by selection-mutation balance, with $\widehat{q}$ denoting the frequency of allele $\mathbf{M}$.[7] Because the mutator allele is disadvantageous, and the first-step mutation confers no selective advantage, the path of sequential fixation will almost always begin with a gamete containing a nonmutant mutator allele. Conditional on the fixation of a first-step mutation, the rate of arrival of the second-step mutation is then $2N\overline{u}_m k p_f(s_2)$, where $\overline{u}_m = \widehat{q}u_1 + (1-\widehat{q})u_0$ is the average background mutation rate experienced by the first-step mutation. Conditional on fixation of the first-step mutation, the rate of fixation of the second-step mutation is then $2N\overline{u}_m k p_f(s_2)$, where $\overline{u} = (1-\widehat{q})^2 u_0 + 2\widehat{q}(1-\widehat{q})u_1 + \widehat{q}^2 u_2$ is the average background mutation rate experienced by the second-step mutation. The rate of sequential fixation is then

$$r_s \simeq \frac{2Nk}{[2\overline{u}_m(1-\widehat{q})p_f(s_1)]^{-1} + [\overline{u}p_f(s_2)]^{-1}}. \tag{A.9}$$

The rate of tunneling must allow for the fact that first-step mutations can arise linked to the alternative alleles at the mutator locus, the rates of which are $4N\overline{u}_m k[(1-\widehat{q})$ and $4N\overline{u}_M k\widehat{q}$ for the $\mathbf{m}$ and $\mathbf{M}$ alleles respectively, where $\overline{u}_M = (1-\widehat{q})u_1 + \widehat{q}u_2$. In the first case, assuming a neutral intermediate, tunneling proceeds at rate $\sqrt{\overline{u}_m k p_f(s_2)}$, and in the second case, it proceeds at approximate rate $\sqrt{\overline{u}_M k p_f(s_2 - s_M)}$, with the selective advantage of the adaptive mutation being discounted by the selective disadvantage of the mutator allele. The total rate of tunneling with neutral intermediates is then

$$r_t \simeq 4Nk^{3/2}\left((1-\widehat{q})\overline{u}_m\sqrt{\overline{u}_m p_f(s_2)} + \widehat{q}\overline{u}_M\sqrt{\overline{u}_M p_f(s_2 - s_m)}\right). \tag{A.10}$$

When first-step alleles are deleterious,

$$r_t \simeq 4Nk^2\left([(1-s_1)/s_1](1-\widehat{q})\overline{u}_m^2 p_f(s_2) + [(1-s_1-h_M s_m)/(s_1+h_M s_M)]\widehat{q}\overline{u}_M^2 p_f(s_2 - s_M)\right), \tag{A.11}$$

[7] As stated in footnote (5), this value is the solution of equation (6) in Lynch (2008)

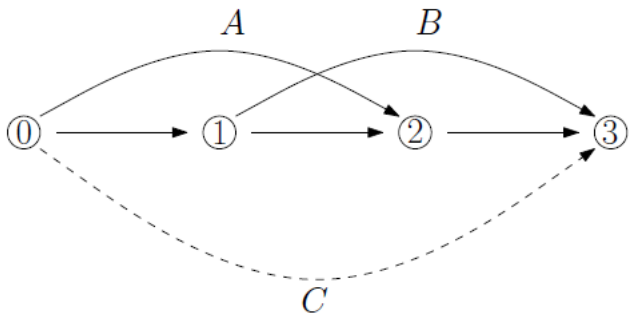
expanding from Iwasa, et al. (2004). The mean time to establishment is again approximated by substituting the previous two expressions into equation (3.5).

## A.5 Expanding to Three or More Adaptive Sites

Throughout this paper, only populations had the alleic form **mab**/**mab**, so expanding the number of sites on the adaptive loci to three, four, or even an arbitrary number of sites is the next step to consider. Equation (4.1) is easily generalized to $n$ adaptive loci in the following manner: recall $s_1 = 0$, and then assume for the purposes of generating a starting point in the analysis that *all* intermediates are neutral. That is, $s_1 = s_2 = \ldots = s_{n-1} = 0$. Then, the rate to go from state[8] 0 to 1 when $n = 2$ is $2u_2k$, from 0 to 1 when $n = 3$ is $3u_2k$, and so on. It is fairly obvious to see that the rate to mutate a mutant destined to fix from state $i$ to $i + 1$ given $n$ sites is $(n - i)u_2k$. From there it is trivial to generalize equation (4.1) to

$$\bar{t}_f \simeq \frac{1}{u_2 k}\left(\sum_{i=2}^{n}\frac{1}{i} + \frac{1}{2Np_f(s_n)}\right). \tag{A.12}$$

To generalize the approach with deleterious intermediates and with $N$ large enough to allow for non-trivial tunneling rates requires deeper analysis. A straightforward method to generalize the nontrivial cases are to try a graphical approach. For example, the following illustration shows an approach for the 3 site case:




The arrows pointing from a state $i$ to a state $i + 1$ is a graphical representation of a sequential fixation, and the arcs are representative of tunneling. The arcs from a state $i$ to a state $i + j$ for $j \geq 1$ are what will be referred to as j-tunneling, tunnels that bypass j fixed states, e.g. they go from state $i$ to state $i + 2$ for $j = 1$ (represented by solid lines in Figure 1). The rate of 1-tunneling has been described at length in this paper. The dashed line is a 2-tunnel, a tunnel that bypasses 2 fixed states. The rate of fixation of 2-tunnels (or any higher tunnels) is unknown. However, they could possibly be so rare as to make little difference in most constructs.

Still referring to Figure 1, the rate through a sequence is exactly what one would expect, but there is an interesting note regarding the rates through the 1-tunnels. Tunnel C, as previously stated, is an unknown quantity. Tunnel B has rate exactly as described previously. Tunnel A, however, is triple the rate of Tunnel B, because there are three equally probable ways of traversing Tunnel A: **abc** → **ABc**, **abc** → **AbC**, or **abc** → **aBC**.

[8] From here henceforth, state refers to the number of adaptive mutants

Acknowledging the above, an expected time to fixation of state 3 is

$$\bar{t}_f \simeq \frac{1}{r_s + r_{t_a} + r_{t_b}}, \tag{A.13}$$

where $r_s$ is the rate in going through sequentially, $r_{t_a}$ is the rate in going through Tunnel A and then through the third sequential path, and $r_{t_b}$ is the rate in going through the first sequential path and then through Tunnel B.

In simulations of this type, the expected time was consistently higher than the simulated time, leading one to suspect that j-tunneling plays a significant role in these processes, and obviously it plays larger roles as n increases because each j-tunnel will be multiplied by a constant (or remain the same but is still grouped with a constant-multiplied j-tunnel).

Recall the constant 3 multiplied by the rate of a 1-tunnel in the discussion about Figure 1. This number can be determined for any j-tunnel from the $i^{th}$ state where there are n adaptive sites which becomes apparent if one thinks in the following way: if someone wants to find out the multiplicative coefficient, $c$, on j-tunneling from the $i^{th}$ state with n adaptive site, an equivalent measure is to find out the multiplicative coefficient on j-tunneling from the $0^{th}$ state with n-i adaptive sites. It quickly becomes apparent that the following holds:

$$c = \binom{n-i}{j+1}. \tag{A.14}$$

## A.6 Acknowledgments

The most obvious person to acknowledge is Dr. Michael Lynch, my faculty advisor. His patience with me–I was a biology-novice at the beginning of the summer–is only eclipsed by his talent, and I'm grateful for both. Other key figures who deserve my gratitude are Kevin Pilgrim, Mandie McCarty, Elizabeth Housworth, the members of the Lynch Lab–particularly Tom Doak, the National Science Foundation for funding this REU program, and Indiana University for playing host.

I owe Zach Norwood my life. While editing my LaTeX code mere hours before this paper's deadline, I (of course) fumbled with something and caused a catastrophic error that he had to fix. He's got a wicked serve, too. He's never lost a set to me, but I refuse to acknowledge that fact until there is a stringent steroid testing program for amateur tennis players.

Jacek Skryzalin also deserves his name in print for creating Figure 1 in its present, electronic form, especially since his computer nearly self-destructed moments before he was about to email it to me.

And, of course, no acknowledgments section is complete without mention of Dan Moore, whose musings about the Cardinals, late-nite meal recommendations, and bizarre dead actress obsessions make the world a brighter place. When he's General Manager, I expect a cushy front office job.

## Bibliography

1. Behe, M. J., and D. W. Snoke. 2004. Simulating evolution by gene duplication of protein features that require multiple amino acid residues. Protein Sci. 13: 2651-2664.

2. Christiansen, F. B., S. P. Otto, A. Bergman, and M. W. Feldman. 1998. Waiting with and without recombination: the time to production of a double mutant. Theor. Popul. Biol. 53: 199-215.

3. Durrett, R., and D. Schmidt. 2008. Waiting for two mutations: with applications to regulatory sequence evolution and the limits of Darwinian evolution. Genetics 180: 1501-1509.

4. Gillespie, J. H. 1984. Molecular evolution over the mutational landscape. Evolution 38: 1116-1129.

5. Higgs, P. G. 1998. Compensatory neutral mutations and the evolution of RNA. Genetica 102/103: 91-101.

6. Innan, H., and W. Stephan. 2001. Selection intensity against deleterious mutations in RNA secondary structures and rate of compensatory nucleotide substitutions. Genetics 159: 389-399.

7. Iwasa, Y., F. Michor, and M. A. Nowak. 2004. Stochastic tunnels in evolutionary dynamics. Genetics 166: 1571-1579.

8. Komarova, N. L., A. Sengupta, and M. A. Nowak. 2003. Mutation-selection networks of cancer initiation: tumor suppressor genes and chromosomal instability. J. Theor. Biol. 223: 433-450.

9. Lynch, M. 2007. The Origins of Genome Architecture. Sinauer Assocs., Inc., Sunderland, MA.

10. Lynch, M. 2008. The cellular, developmental and population-genetic determinants of mutation-rate evolution. Genetics 180: 933-943.

11. Pigliucci, M. 2007. Do we need an extended evolutionary synthesis? Evolution 61: 2743-2749.

12. Stephan, W. 1996. The rate of compensatory evolution. Genetics 144: 419-426.

# Random Sampling of Constrained Phylogenies

John R. Brown
Indiana University

Indiana University REU Summer 2009
Advisor: Elizabeth Housworth

## B.1 Introduction

In evolutionary biology, statistical analyses which compare data among present-day species require knowledge of the common evolutionary history of the species. This information consists of a *phylogeny*, a tree which has tips corresponding to present-day species (also called *taxa*) and has branching nodes which correspond to speciation events (when one species becomes two). Phylogenies can also have branch lengths which correspond to the amount of time between nodes. If a phylogeny is known for certain, a standard method exists for incorporating it into the statistical analysis. When the phylogeny is unknown (due to insufficient DNA sequence data for instance), we could use a completely random set of phylogenies to simulate the possible relationships between species. However, often some information about the structure of the tree is available, from morphological data or the fossil record. Thus an intermediate option which generates only random trees which fit the known structure would allow this partial information to be incorporated into the analysis, without assuming that a single phylogeny is correct.

Housworth and Martins showed how to break the set of constrained trees into categories which are defined by a triplet of combinatorial objects $(P, T, C)$. They gave a formula for the number of trees contained in a category. They also showed how to draw a random tree from a given category. Thus to draw a random tree, one would choose a random category and then draw a tree from within that category. The category should be sampled from the probability distribution corresponding to the category weights (i.e. sample larger categories more often and smaller ones less often). The original, naive, implementation of category generation was not computationally feasible for even reasonably large problems. This paper describes an efficient Monte Carlo Markov Chain method for generating $(P, T, C)$ triplets according to the distribution of the category weights.

## B.2 The Problem

The problem consists of taking a uniform random sample from a specific subset of all *bifurcating, rooted, ordered, labeled* trees. *Bifurcating* means that each branch splits into exactly two other branches, and implies that every internal node of the tree has degree three. *Rooted* means that there is one node of degree two from which all branches descend. *Ordered* means that the temporal ordering of the entire tree is relevant, implying for example that the two trees in figure B.1 are distinct. *Labeled* means that the tips of the tree are have names. Thus permutations of the names can create distinct trees.

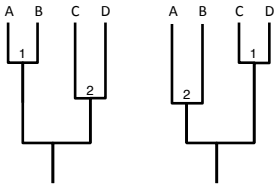



Figure B.1: Two trees which have different ordering

The type of biological information we want to incorporate into the comparative analyses divides the $N$ taxa into *subclades* and *outgroups*. The subclades are subsets of the taxa which are more closely related to each other than to the rest of the taxa. In other words, for each subclade, the most recent common ancestor is a root for a subtree whose leaves are exactly the elements of the subclade. We will label the

subclades $(1, \ldots, r)$ and their corresponding sizes $(k_1, \ldots, k_r)$. Outgroups are taxa which are known to be more distantly related to the given subclades. We will let $n$ be the number of outgroups, which we will also call *external* taxa. An example of a constraint on 12 taxa would be (A,B,C,D),(E,F,G),(H,I),J,K,L. Here subclade 1 consists of (A,B,C,D), subclade two consists of (E,F,G), subclade three is (H,I) and the external taxa are J,K, and L. Thus $k_1 = 4$, $k_2 = 3$, $k_3 = 2$ and $n = 3$. An example of a tree which fits this set of constraints is shown in figure B.2.

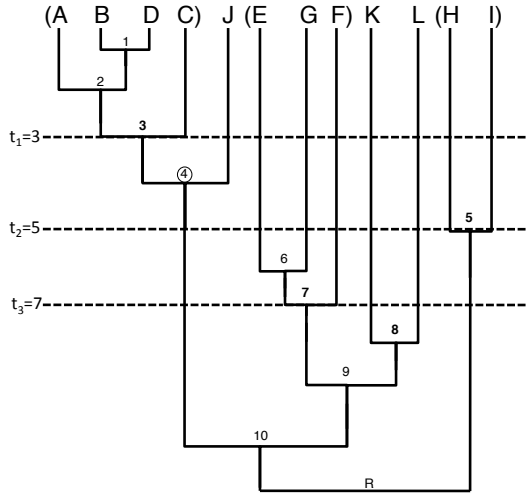



Figure B.2: One phylogeny from the category with $P = (1, 3, 2)$; $T = (3, 5, 7)$; $C = (0, 1, 0)$

To categorize the subset of trees under a constraint of the form above, we will use the triplet $(P, T, C)$. Each of $P, T$, and $C$ is an ordered r-tuple.

1. P is a permuation of 1, ..., r. It specifies the order of the roots of the subclades starting from the tips and going toward the root. $p_i$ is the label of the subclade which is the $i^{th}$ one to be fully joined. In the example above, $P = (1, 3, 2)$ because subclade 1, (A,B,C,D), is joined first, subclade 3, (H,I), is joined second, and subclade 2, (E,F,G) is joined third.

2. T gives the temporal rank of the subclade roots. $t_i$ gives the time (the ordinal rank from the tips to the root) for subclade $p_i$. Thus the interpretation of $T$ depends on $P$. In the example in figure B.2, $T = (3, 5, 7)$ because subclade $p_1 = 1$ is joined at the third event, subclade $p_2 = 3$ is joined at the fifth event, and subclade $p_3 = 2$ is joined at the seventh event.

3. C is a composition. It specifies the number of external joinings which occur between subclade roots. $c_i$ gives the number of external joining events before $t_i$ and after $t_{i-1}$. Once a clade is completely joined at its root, we consider it to be an external node. Thus we start with $n$ external taxa, but after each subclade root, we add one more. In figure B.2, we circle the external joining event between J and the fully joined subclade 1 which occurs at time 4. Since this occurs before $t_2$ and after $t_1$, $c_2 = 1$. Since there are no other external joining events, the other entries of C are zero. Thus $C = (0, 1, 0)$.

## B.3 The Solution

### B.3.1 Counting Constrained Compositions

Let $c$ and $r$ be positive integers. A *composition* of $c$ into $r$ parts is defined as an ordered list $C$ of nonnegative integers $(C_r, \ldots, C_1)$, where $\sum_{i=0}^{r-1} C_{r-i} = c$. (We adopt a somewhat awkward notation now for convenience later). The total number of compositions of $c$ into $r$ is $\binom{c+r-1}{r-1}$. However we want to count only those compositions of c which satisfy constraints of the form $\sum_{i=0}^{j} C_{r-i} \leq \sum_{i=0}^{j} A_{r-i}$ for all $j \in \{0, \ldots, r-1\}$ and some list $A = (A_r, \ldots, A_1)$. In other words, the partial sums of the composition $C$ cannot exceed the partial sums of a fixed constraint list $A$.

We begin by noting that the number we wish to calculate is equal to the sum over every possible value of each entry:

$$\sum_{C_r=0}^{A_r} \sum_{C_{r-1}=0}^{A_r+A_{r-1}-C_r} \cdots \sum_{C_2=0}^{\sum_{i=2}^{r} A_i - \sum_{i=3}^{r} C_i} 1$$

We do not sum over $C_1$ because we assume that the total for the composition $c$ is fixed, so choosing all but one part of the composition leaves only one possibility for the last part. To compute this sum, we will use a recursive algorithm. We will first need the following fact from basic combinatorics:

**Lemma B.3.1.** *For positive integers $n$ and $k$:*

$$\binom{n}{k} = \binom{n+1}{k+1} - \binom{n}{k+1}$$

*Proof.* There are $\binom{n+1}{k+1}$ ways of choosing a $k+1$ sized subset from a set of $n+1$ objects. Let one of the $n+1$ objects be called $x$. When choosing the $k+1$ things, either $x$ is chosen or it is not. In the former case, there are $\binom{n}{k}$ ways of choosing the rest of the $k$ elements of the subset from the other $n$ objects. In the latter, there are $\binom{n}{k+1}$ ways of choosing the $k+1$ subset from the other $n$ objects. Thus $\binom{n+1}{k+1} = \binom{n}{k+1} + \binom{n}{k}$ Q.E.D.

We will use this fact to split terms into a difference of two terms which will telescope in the sum. To illustrate the recursive algorithm, we will show the cases of $r \leq 4$. First note that the case $r = 2$ is trivial, as the sum is simply $\sum_{C_2=0}^{A_2} 1 = A_2 + 1$. For the case $r = 3$, the sum is

$$\begin{aligned}
\sum_{C_3=0}^{A_3} \sum_{C_2=0}^{A_3+A_2-C_3} 1 &= \sum_{C_3=0}^{A_3} \binom{A_3 + A_2 - C_3 + 1}{1} \\
&= \sum_{C_3=0}^{A_3} \binom{A_3 + A_2 - C_3 + 2}{2} - \binom{A_3 + A_2 - C_3 + 1}{2} \\
&= \binom{A_3 + A_2 + 2}{2} - \binom{A_2 + 1}{2}
\end{aligned}$$

Observe that the first equality, though trivial, is equivalent to treating the inner sum as the case $r = 2$, substituting $A_3 + A_2 - C_3$ for $A_2$. We will use an analogous substitution for later steps in the algorithm.

The second equality is by lemma B.3.1 and the third is by telescoping. For the case $r = 4$ the sum is

$$\begin{aligned}
& \sum_{C_4=0}^{A_4} \sum_{C_3=0}^{A_4+A_3-C_4} \sum_{C_2=0}^{(A_4+A_3-C_4)+A_2-C_3} 1 \\
= & \sum_{C_4=0}^{A_4} \binom{A_4+A_3-C_4+A_2+2}{2} - \binom{A_2+1}{2} \\
= & \sum_{C_4=0}^{A_4} \binom{A_4+A_3-C_4+A_2+3}{3} - \binom{A_4+A_3-C_4+A_2+2}{3} - \binom{A_2+1}{2} \\
= & \binom{A_4+A_3+A_2+3}{3} - \binom{A_3+A_2+2}{3} - \binom{A_4+1}{1}\binom{A_2+1}{2}
\end{aligned}$$

The first equality is seen by replacing $A_3$ in the $r = 3$ case with $A_4 + A_3 - C_4$ in the inner two sums. Again, we use lemma B.3.1 and the telescoping of the sum to obtain the next two lines. In the $r = 5$ case, we substitute $A_5 + A_4 - C_5$ for $A_4$ and repeat this process. The first term in the answer to the $r = 4$ case splits into two terms by the lemma, the second term is constant and is multiplied by an additional factor, the third term keeps the constant multiplier $\binom{A_2+1}{2}$ and the $\binom{A_4+1}{1}$ splits into two terms by the lemma. By continuing this process, we can recursively find a formula for the number of constrained compositions for any r. Further sums are tedious by hand and not very illuminating. However, we can program a computer to run this recursion efficiently. The R code to do this is contained in the Appendix, in the subroutine "recursion_count".

### B.3.2 The Jump Process

To implement a Monte Carlo Markov chain, we must have a so-called "jump process," a method of generating candidate points given the current state of the chain. Those candidates will then be accepted or rejected based on some criteria related to the target distribution. A simple type of jump process is one that is also independent of the current state. We attempted to create a process which generates triplets $(P, T, C)$ completely uniformly at random. The chief difficulty lies in the interdependencies of $(P, T, C)$. These constraints mean that early choices in the algorithm force the number of possible later choices to change. To adjust for this non-uniformity we often must calculate the total number of later choices given every possibility of the current choice and then bias our decision based on that information. We succeeded in producing an algorithm which mostly chooses $(P, T, C)$ uniformly at random. One non-uniformity remains, namely that the number of choices for $T$ depends on the permutation, $P$. However, this is easily corrected for later in the Monte Carlo Markov Chain algorithm.

*Algorithm* B.3.2. The process to generate a (P,T,C)

1. Choose P, a permutation of {1,. . . , r }, uniformly at random.
2. Choose c, the total for the composition C, weighted according to number of compositions of each c allowed by the constraints on C
3. Choose C, the composition, uniformly at random, reject if it fails constraints.

4. Choose $\tau$, the total for the composition $\Theta$, according to the number of compositions allowed by the constraints on T

5. Choose composition $\Theta$, uniformly at random, and reject if it fails constraints.

6. Find T, which is completely determined by the previous steps.

For choosing random permutations and combinations in steps (1), (3), and (5), we use standard algorithms found in Nijenhuis and Wilf [2]. For steps (2) and (4) we use the recursion algorithm developed in subsection B.3.1. For step (2), the constraint list $A = (n-1, 1, \ldots, 1, 0, \ldots, 0)$ such that the sum of the list is $c$. This constraint arises because we begin with $n$ external taxa, which can only be joined $n-1$ times. However, after each clade is fully joined, we consider it to be new external node. Thus, the allowed partial sum of C increases by one for each entry of C. In the case where $c \leq (n-1)$, the constraint list is $(c, 0, \ldots, 0)$, and all $\binom{c+r-1}{r-1}$ compositions are allowed. For step (4), the constraint list $A = (k_{p_r} - 2, k_{p_{r-1}} - 2, \ldots, k_{p_2} - 2)$. This constraint reflects the indirect method we use to choose the times using a composition.

### B.3.3 Choosing T by a Constrained Composition

To see how to use a composition to choose T, we must first consider two sets of smallest and largest values for T. This will give us a range for valid times. Throughout this discussion, P and C are fixed and known. First note that already we know the last entry of T. The equation $T_r = \sum_{j=1}^{r} c_j + \sum_{j=1}^{r} k_j - r$ is always true because we know that the $j^{th}$ subclade needs $(k_j - 1)$ times to join and that each external joining uses one time. Let $(t_1, \ldots, t_r)$ denote the list of minimal times, which we call $T_{min}$, and let $(T_1, \ldots, T_r)$ denote the list of maximal times, which we call $T_{max}$. To find the minimal times, we work from the beginning of the list, and find the smallest value for the next entry by leaving enough times for the next subclade to be joined as well as the number of external joinings specified by the composition. For the maximal times, we work from the end of the list, leaving enough time for external joinings. This algorithm (worked out by Housworth and Martins) yields the following equations:

$$\begin{aligned} t_1 &= k_{p_1} - 1 + c_1 \\ t_i &= t_{i-1} + k_{p_i} - 1 + c_i = \sum_{j=1}^{i} k_{p_j} - i + \sum_{j=1}^{i} c_j \qquad \text{(B.1)} \\ T_{r-1} &= T_r - c_r - 1 \\ T_{r-i} &= T_{r-i+1} - c_{r-i+1} - 1 = T_r - \sum_{j=1}^{i} c_{r-j+1} - i \qquad \text{(B.2)} \\ &= \sum_{j=1}^{r} c_j + \sum_{j=1}^{r} k_j - r - \sum_{j=1}^{i} c_{r-j+1} - i \end{aligned}$$

Now we take the entry-wise difference between $T_{max}$ and $T_{min}$ and call this new list $D$, for differences.

For $i \in \{0, \ldots, r-1\}$:

$$\begin{aligned} D_{r-i} = T_{r-i} - t_{r-i} &= \sum_{j=1}^{r} c_j + \sum_{j=1}^{r} k_j - r - \sum_{j=1}^{i} c_{r-j+1} - i \\ &- \left[ \sum_{j=1}^{r-i} c_j + \sum_{j=1}^{r-i} k_{p_j} - (r-i) \right] \\ &= \sum_{j=1}^{i} k_{p_{r-j+1}} - 2i \end{aligned}$$

A consequence of the cancellation of the above sums containing $c_j$ is that the number of possible sets of last joining times $T$ does not depend on the composition $C$. The composition is needed to find $T_{max}$ and $T_{min}$, which give the nominal ranges for the entries $T$, but the size of those ranges is independent of the composition.

We now take another difference, this time between adjacent entries of $D$:

$$\begin{aligned} D_{r-i} - D_{r-i+1} &= \sum_{j=1}^{i} k_{p_{r-j+1}} - 2i - \sum_{j=1}^{i-1} k_{p_{r-j+1}} - 2(i-1) \\ &= k_{p_{r-i+1}} - 2 \end{aligned}$$

Let $T_{ref}$ denote the list of all the double differences: $(D_1 - D_2, D_2 - D_3, \ldots, D_{r-2} - D_{r-1}) = (k_{p_2} - 2, \ldots, k_{p_r} - 2)$. Note that this double differencing procedure leaves a list which depends only on the sizes of the subclades in the permuted list $k_p$.

For example, suppose we have a problem where $(k_1, \ldots, k_6) = (2, 3, 4, 5, 5, 6)$ and $n = 6$ external taxa. Suppose we choose the permutation $P = (3, 2, 5, 4, 1, 6)$ and the composition $C = (0, 0, 2, 4, 3, 1)$. Since the composition total $c$ is 10 and the sum of the sizes of the 6 subclades is 25, we know that $T_r = 10+25-6 = 29$. After finding $T_{max}$ and $T_{min}$ using the equations B.1 and B.2, we subtract twice to obtain $D$ and $T_{ref}$ as in the following table:

| | | | | | | |
|---|---|---|---|---|---|---|
| $T_{max}$ | 14 | 15 | 18 | 23 | 27 | 29 |
| $T_{min}$ | 3 | 5 | 11 | 19 | 23 | 29 |
| $D$ | 11 | 10 | 7 | 4 | 4 | 0 |
| $T_{ref}$ | 1 | 3 | 3 | 0 | 4 | |

To choose our times, we will choose a composition $\Theta$ which will be constrained by $T_{ref}$, and then we will do the reverse of the above differencing procedure to find a corresponding set of times.

*Algorithm* B.3.3. Here we use a constrained composition to choose T.

1. Choose $\tau \in [0, \sum_{i=1}^{r-1} T_{ref_i}]$, weighted according to the number of compositions of $\tau$ into $(r-1)$ parts, constrained as in B.3.1 by $(k_{p_r} - 2, k_{p_{r-1}} - 2, \ldots, k_{p_2} - 2)$ ($T_{ref}$ in reversed order).

2. Generate a random composition, denoted $\Theta$ of $\tau$ into (r-1) parts, and reject if it fails the same constraint as in step (1).

3. Generate a list of length $r$, called $D$. $D_r \leftarrow 0$ ; $D_{r-i} \leftarrow \Theta_i + D_{r-i+1}$ for $i \in \{1, \ldots, r-1\}$
4. Generate $T_{max}$, the set of times where every entry is maximal, using equation B.2
5. $T_i \leftarrow T_{max_i} - D_i$ for $i \in \{1, \ldots, r\}$

Later on we will need the total number of possible choices for the times $T$, which we will denote by $T_{tot}(P)$. This is simply the sum of all the weights calculated in step (1) of algorithm B.3.3. Again note it only depends on the choice of permutation.

Using the same example as above, we would begin algorithm B.3.3 by choosing $\tau$ between 0 and $\sum_{i=1}^{r-1} T_{ref_i} = 11$, since that is the greatest total our composition can take can correspond to a valid set of times. Suppose we choose $\tau = 8$. We now choose $\Theta$ so that it corresponds to a set of valid times. This constraint is exactly the one in **??** with $A = T_{ref}$ (order reversed). For instance in this example $\Theta_a = (1, 2, 4, 0, 2)$ is valid, while $\Theta_b = (3, 1, 4, 1, 0)$ is not. To see that this is true, we can find the set of times $T_a$ and $T_b$ corresponding to $\Theta_a$ and $\Theta_b$ respectively. We can illustrate the process of finding $D_a$ and $D_b$ and $T_a$ and $T_b$ by steps (3) and (5) of algorithm B.3.3 by the following two tables:

| $\Theta_a$ | | 1 | | 2 | | 4 | | 0 | | 2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Reverse ordered $\Theta_a$ | | 2 | | 0 | | 4 | | 2 | | 1 | |
| $D_a$ | 9 | | 7 | | 7 | | 3 | | 1 | | 0 |
| $T_{max}$ | 14 | | 15 | | 18 | | 23 | | 27 | | 29 |
| $T_a = T_{max} - D_a$ | 5 | | 8 | | 11 | | 20 | | 26 | | 29 |
| $T_{min}$ | 3 | | 5 | | 11 | | 19 | | 23 | | 29 |

| $\Theta_b$ | | 3 | | 1 | | 4 | | 1 | | 0 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Reverse ordered $\Theta_b$ | | 0 | | 1 | | 4 | | 1 | | 3 | |
| $D_b$ | 9 | | 9 | | 8 | | 4 | | 3 | | 0 |
| $T_{max}$ | 14 | | 15 | | 18 | | 23 | | 27 | | 29 |
| $T_b = T_{max} - D_b$ | 5 | | 6 | | 10 | | 19 | | 24 | | 29 |
| $T_{min}$ | 3 | | 5 | | 11 | | 19 | | 23 | | 29 |

Note that step (3) can be interpreted as adding each term of a reverse ordered $\Theta$ to $D$, starting from the end of the list. Observe that every entry of $T_a$ falls between the corresponding entries of $T_{max}$ and $T_{min}$, but the third entry of $T_b$ is too small. This is because the sum of the first three entries of $\Theta_b$ is $3+1+4 = 8$ but the sum of the first three entries of $T_{ref}$ is 7.

### B.3.4 The Monte Carlo Markov Chain

We will use the Metropolis-Hastings algorithm to construct a random walk on the space of categories. For a general overview of the Metropolis-Hastings algorithm see [3].

*Algorithm* B.3.4. Given the current state $(P, T, C)$, we go to the next state by this process:

1. Use the Markov Chain (Algorithm B.3.2) to generate a candidate category $(P', T', C')$
2. Calculate the weight of the new category, $W_{P',T',C'}$
3. Generate $\mathcal{U} \in [0, 1]$, a uniform random variable.

4. If $\mathcal{U} < \frac{\mathcal{W}(P',T',C')}{\mathcal{W}(P,T,C)} \frac{\mathcal{J}(P,T,C)}{\mathcal{J}(P',T',C')}$, then $(P,T,C) \leftarrow (P',T',C')$

In step (4), we essentially flip a weighted coin to decide to accept a new category or not. The probability of acceptance is the product of two ratios. The first ratio is given by equation B.3 determined by Housworth and Martins which gives the relative weight of a category. $\mathcal{W}(P,T,C)$ is the number of trees contained in the category divided by a common factor to all categories.Thus the ratio $\frac{\mathcal{W}(P',T',C')}{\mathcal{W}(P,T,C)}$ is equal to the ratio of the total number of trees in the two categories.

$$\mathcal{W}(P,T,C) = \binom{t_1-1}{c_1}\binom{n-c_1+1}{2}\binom{t_1-1-c_1}{k_{p_1}-2} \times \prod_{j=2}^{r}\left[\binom{t_j-t_{j-1}-1}{c_j}\binom{n-\sum_{i=1}^{j}c_i+j}{2}\binom{t_j-\sum_{i=1}^{j}c_i-\sum_{i=1}^{j-1}k_{p_i}+(j-2)}{k_{p_j}-2}\right] \tag{B.3}$$

This acceptance decision based on the values of the target distribution at both points is how Metropolis-Hastings algorithm incorporates the target distribution. By choosing to accept a jump only part of the time in this way, we create a walk which after a large number of samples converges to the target distribution. If the category is large, we are likely to not accept many jumps away from it. Likewise, (in the symmetric Metropolis algorithm) if any candidate category is larger than the current category, we always jump to it.

The second ratio in step (4) of algorithm B.3.4 corrects for having an asymmetrical jump process. By *symmetric*, we mean that the probability that one generates a candidate from the current state must equal the probability of doing the reverse process. We denote by $\mathcal{J}(P,T,C)$ the probability that the triplet $(P,T,C)$ will be generated as a candidate cateogry. A symmetric process would then satisfy the condition $\mathcal{J}((P',T',C')|(P,T,C)) = \mathcal{J}((P,T,C)|(P',T',C'))$. Since our jump process is independent, meaning it does not depend on the current state, we can drop the conditions on these probabilities and simply say $\mathcal{J}(P,T,C) = \mathcal{J}(P',T',C')$. However, this condition is not satisfied because the number of possible times, $T_{tot}$, depends on the permutation, $P$. Thus our process is not completely uniform, and we must incorporate the asymmetry into the Metropolis-Hastings algorithm. The ratio $\frac{\mathcal{J}(P,T,C)}{\mathcal{J}(P',T',C')}$ accomplishes this correction. This ratio of probabilities is easily calculated, since it is simply equal to the reciprocal of the ratio total times.

$$\frac{\mathcal{J}(P,T,C)}{\mathcal{J}(P',T',C')} = \frac{T_{tot}(P')}{T_{tot}(P)} \tag{B.4}$$

We can see this is true by realizing that the permutation and combination are chosen uniformly at random so the probabilities will cancel in the ratio, while the probability that a given set of times is chosen (given a fixed P) is $1/T_{tot}(P)$.

## B.4 Discussion

We programmed the algorithm into the R language and ran the program on a MacBook laptop. The code to run the program is found in the Appendix, along with all necessary subroutines. The acceptance rate (the number of times a jump was accepted divided by the number of trials) when running the Metropolis-Hastings algorithm with our jump process was about 1 percent. To increase this acceptance rate, we instead use every $100^{th}(P,T,C)$ in the chain. The acceptance rate with this adjustment is about 90 percent for 5 subclades and about 25 percent for 10 subclades. The algorithm's runtime is about one minute to generate

100 random categories with 5 subclades and 30 total taxa. It increases to about 7 minutes to generate 100 random categories with 10 subclades and 50 total taxa. This could be improved two ways. First, translating to a lower-level compiled language such as C or C++. Second, we could improve the method of choosing compositions in steps (3) and (5) in algorithm B.3.2. The current way of randomly picking compositions and checking if they satisfy the constraints is inefficient. For larger problems, it often takes many tries to randomly find a composition which fits the constraints (especially with the composition for times). A natural extension of this problem would be support for nested constraints on the trees. Constraints for this problem might look like: ((A,B,C),(D,E),F))(G,H,I),J,K,L. To choose a random tree from these constraints, we could use generate categories for the overall structure: (A,B,C,D,E,F),(G,H,I),J,K,L. Then we would use the process described in [1] to generate a random tree. This would give us a set of times that each subclade uses to join its elements. We would relabel these times and iterate the process for the inside constraints. The only alteration needed is a change in the formula for the weight, equation B.3.

## B.5 Acknowledgements

Thanks goes first to my advisor, Professor Elizabeth Housworth for giving me an appropriate problem and guiding me along the way to solving it. Special thanks goes to Professor Nets Katz for suggesting the recursion method to count constrained compositions. I want to thank Kevin Pilgrim for directing the REU program, Mandie McCarty for all things administrative and edible, and the National Science Foundation for funding the Research Experience of this Undergradute. To my fellow REUers, I'd like to thank you for making this summer an enjoyable and diverse experience. I especially liked listening to you all talk about your projects and math in general. One thing I know I learned this summer is that what I have seen so far is barely the tip of a mathematical iceberg. That

## Bibliography

1. Housworth, E. A. and E. P. Martins, *Random Sampling of Constrained Phylogenies: Conducting Phylogenetic Analyses When the Phylogeny Is Partially Known*, Systematic Biology **50** (2001), 628–639.

2. Nijenhuis, A, and H. Wilf *Combinatorial algorithms for computers and calculators.* Academic Press, New York, 1978.

3. Chib, S. and E. Greenberg. *Understanding the Metropolis-Hastings Algorithm*, The American Statistician **49** (1995), 327–335.

# Appendices

## A Computer Code

The following is the R code which we used to run the algorithms described in the paper. The function Metropolis is called with a vector k of length r, which contains the sizes of the subclades, n, the number of external taxa and trials, the number of desired (P,T,C) triplets. It stores every $100^{th}$ category and calculates the acceptance rate, defined as the number of times the category changed divided by the number of trials. Metropolis calls the subroutines "weight" and "ranPTC". Subroutine "weight" executes the

formula from [1] for the category weight, and the code is written by Housworth. Subroutine "ranPTC" is the jump process (algorithm B.3.2). ranPTC calls subroutines "rancom", "Final_T", and "recursion_count". Subroutine "rancom" is a translation of the algorithm in [2] to find a random composition. Subroutine "Final_T" is a translation of the algorithm in [1] to find the maximal set of times given P and C. Subroutine "recursion_count" executes the algorithm in section B.3.1 and the code is written by Housworth.

```
Metropolis<-function(k,n,trials){
ptc<-ranPTC(k,n)
w1<-weight(k,n,ptc[[1]],ptc[[2]],ptc[[3]])
Perms<-vector(mode = "list", length = trials+1)
Times<-vector(mode = "list", length = trials+1)
Comps<-vector(mode = "list", length = trials+1)
Perms[[1]]<-ptc[[1]]
Times[[1]]<-ptc[[2]]
Comps[[1]]<-ptc[[3]]
accept<-0
for(i in 2:(100*trials+1)){
candidate<-ranPTC(k,n)
w_c<-weight(k,n,candidate[[1]],candidate[[2]],candidate[[3]])
#print(w_c)
u<-runif(1)
if (u<((w_c/w1)*(candidate[[4]]/ptc[[4]]))){
ptc<-candidate
w1<-w_c
changed<-TRUE
}
#Every 100th time, store the result, increment the acceptance if we've moved,
# and reset logical variable changed to FALSE
if (1==i%%100){
Perms[[(i %/% 100)+1]]<-ptc[[1]]
Times[[(i %/% 100)+1]]<-ptc[[2]]
Comps[[(i %/% 100)+1]]<-ptc[[3]]
if(changed==TRUE) accept<-accept+1
changed<-FALSE
}
}
acceptance<-accept/(trials)
print(Perms)
print(Times)
print(Comps)
print(acceptance)
}

#The following code computes the weights. comp is C, Times is T, and
#k_perm is the size of the clade in the permuted list.
```

```
weight<-function(k,n,p,Times,comp){
r<-length(k)
k_perm<-rep(0,r)
for (i in 1:length(k)) {
k_perm[i]<-k[p[i]] }
P_new = 1
P_new = P_new*choose((Times[1] -1), comp[1]) * choose((n - comp[1] + 1), 2)
P_new = P_new*choose((Times[1] - comp[1] - 1), k_perm[1] - 2)
for (j in 2:r){
P_new = P_new*choose((Times[j] - Times[(j-1)] -1), comp[j])*
choose((n-sum(comp[1:j]) + j), 2)*choose( (Times[j] - sum(comp[1:j])
-sum(k_perm[1:(j-1)]) +(j-2) ), k_perm[j] - 2)
}
return(P_new)
}

#this is the code for the jump process
ranPTC<-function(k,n) {
r<-length(k)
#find random permutation
p<-sample(1:r)

#################################################################
#count the allowed compositions for all possible c
allowed<-rep(0,n+r-1)
a<-rep(0,r)
for(i in 0:(n+r-2)){
if ((i-n)<0) a[1]<-i
else {
a[1]<-n-1
for (j in 2:(2+i-n)){
a[j]<-1
}
}
allowed[i+1]<-recursion_count(a,r,i)
}
#choose c according to weighted probability of allowed
c<-sample(0:(n+r-2),1,prob=allowed)
done<-0
#generate random compositions until one fits the constraints
while(!done){
comp<-rancom(c,r)
sum<-0
ok<-1
for(j in 1:(r-1)) {
```

```
sum<-sum+comp[j]
if (sum>max(n-2+j,0)) ok<-0
}
if (ok==1) done<-1
}

#####################################################################
#find t_r, which is set by c already
t<-rep(0,r)
t[r]<-c+sum(k)-r

#find the reference_T corresponding to extreme times
reference_T<-rep(0,r-1)
for (i in 1:(r-1)) reference_T[i]<-k[p[i+1]]-2

#generate a vector allowed_times which contains the number of
#possible constrained compositions for each composition total.
allowed_times<-rep(0,sum(reference_T)-1)
#there's always only one allowed for the sum of 0
allowed_times[1]<-1
#we initialize a_times to the maximum sum, we will decrement it each loop
#to keep its sum equal to i
a_times<-rev(reference_T)
for(i in sum(reference_T):1){
allowed_times[i+1]<-recursion_count(a_times,r-1,i)
entry<-(which(cumsum(rev(reference_T))>(i-1))[1])
a_times[entry]<-a_times[entry]-1
}

#total_times is the total number of times, which depends on p.
#We will need this later for the asymmetry bias correction.
total_times<-sum(allowed_times)

#now we pick a composition total based on allowed_times
t_comptotal<-sample(0:sum(reference_T),1,prob=allowed_times)
#now we pick a random composition and check to see if it fits the constraints
done<-0
while(!done){
t_comp<-rancom(t_comptotal,r-1)
ok<-1
for(j in 1:(r-2)) {
if (cumsum(rev(t_comp))[j]>cumsum(rev(reference_T))[j]) ok<-0
}
if (ok==1) done<-1
}
#now we construct times based on t_comp, Final_T
```

```
differences<-rep(0,r)
for (i in 1:(r-1)) differences[r-i]<-t_comp[r-i]+differences[r-i+1]
t<-Final_T(p,comp,t)-differences

ptc<-list(p,t,comp,total_times)
return(ptc)
}

recursion_count <- function(a, r, c){
count = 0
if (r==2){# it's all trivial
count = a[1] + 1
}
else #r is at least 3
{
recursion = rbind(c(0, 0, -choose(a[r-1] + 1, 2)),
c(a[r-2]+a[r-1]+2, 2, 1))

if(r>3){
for (j in (r-3):1){
recursion = rbind(recursion, c(a[j]+1, 1, recursion[1,3]))
recursion[1,] = c(0, 0, 0)
for (i in 2:(nrow(recursion)-1)){
recursion[1,3]=recursion[1, 3]-recursion[i, 3]*choose(recursion[i,1],recursion[i,2]+1)
recursion[i,1]=recursion[i, 1] + 1 + a[j]
recursion[i,2]=recursion[i, 2] + 1

} # end for i loop
} # end for j loop
}
for (i in 1:length(recursion[,1])){
count = count + recursion[i, 3]*choose(recursion[i, 1], recursion[i, 2])
}
}# end else r is at least 3
count

} # end recursion_count

#This generates a random composition of c into r.
rancom<-function(c,r) {
a<-sample(c+r-1,r-1)
a<-sort(a)
comp<-c(1:r)
comp[1]<-a[1]-1
if(r==2) {comp[2]<-c+1-a[1]}
else{
```

```
for (i in 2:(r-1)) {
comp[i]<-a[i]-a[i-1]-1
}
comp[r]<-c+r-1-a[r-1]
}
return(comp)
}

#This generates the maximal set of times T_max
Final_T<-function(p,comp,t) {
r<-length(p)
final_T<-1:r
final_T[r]<-t[r]
final_T[r-1]<-final_T[r]-comp[r]-1
for (i in 2:(r-1)) {
final_T[r-i]<-final_T[r-i+1]-comp[r-i+1]-1
}
return(final_T)
}
```

# Sorting Signed Permutations With Weighted Reversals

Zachary Doenges
Indiana University

Indiana University REU Summer 2009
Advisor: Matthew Hahn

# C.1 Introduction

**Definition C.1.** The reversal p(i,j) is defined in the usual way.
$(\pi_1...\pi_{i-1}\pi_i\pi_{i+1}...\pi_j\pi_{j+1}...\pi_n) \mapsto (\pi_1...\pi_{i-1}\pi_j\pi_{j-1}...\pi_{i+1}\pi_i\pi_{j+1}...\pi_n)$

e.g. $(1,2,3,4,5)p(2,4) = (1,4,3,2,5)$

**Definition C.2.** A signed reversal p'(i,j) operates the same way as p(i,j) except it flips the signs of every object it moves.

e.g. $(1,2,3,4,5)p(2,3) = (1,-4,-3,-2,5)$

Given an ordering of genes into a chromosome, or genome, it is possible to speak of the distance (the noise) between two adjancent genes.

**Definition C.3.** Let $d_i$ represent the length of noise between gene i-1 and gene i for $1 < i < n$. When i=0,n , then $d_i$ shall be infinite.

Assumption 1: The lengths of a reversal occur exponentially. $f(l) = \lambda e^{-\lambda(l)}$

Assumption 2: The cut of a particular reversal always occurs in the noise and providing the length of the reversal doesn't change one cut is as likely as another.

Assumption 3: Our model will be continous, that is for a reversal of i-j, a $x_1 \in [0, d_1]$ is a valid left hand cut. O shall denote the cut at exactly the gene i and $d_i$ shall denote the cut at the gene i-1

A reversal p(i,j) now requires more information, namely the length of the reversal and where the genome/chorosome was cut. To represent this information the notation p(i,j,k,$x_1$) shall be adopted, with i-j being the interval of genes affected; k being the length of the reversal; and $x_1$ being the left hand cut. Note that $l \geq \sum_{k=i}^{j-1} d_k$ and $0 \leq x_1 \leq d_{i-1}$

Let $S_n^{\pm} = S_n \times \{-1,1\}^n$. Here the first component is the ordering of genes and the second conveys the signs of each of those genes. The set this paper concerns itself with is $S_n^{\pm} \times (\mathbb{R}^+)^{n-1}$. The component $(\mathbb{R}^+)^{n-1} = (d_1, ..., d_{n-1})$ stores the distances.

**Definition C.4.** $L_i^j = \sum_{n=i}^{j-1} d_n$. This is the minimum length of a reversal over markers i-j.

p(i,j,k,$x_1$) sends $(\pi_1...\pi_{i-1}\pi_i\pi_{i+1}...\pi_j\pi_{j+1}...\pi_n) \mapsto (\pi_1...\pi_{i-1} - \pi_j - \pi_{j-1}... - \pi_{i+1} - \pi_i\pi_{j+1}...\pi_n)$ and sends $(d_2, ..., d_n) \mapsto (d_2...d_{i-1}, d_i + (l - L - 2x_1), d_{i+1}...d_{j-1}, d_j + (2x_1 - l + L), d_{j+1}..., d_n)$

It is important to note that the reversal with cuts in the noise moves not only the genes but also the noise. The mapping on the distances above tracks the movement of this noise.

Now that the stage is set, we can address the matter of finding a probability model. This paper will divided into two main parts. The probability of a partiular reversal and an algorithm for uniform sampling of paths.

# C.2 Calculation of The Probability of a Particular Reversal

$P(i,j,l,x_1)$ will be the notation for the probability of the reversal of genes $i-j$ with length l and left cut point $x_1$. Note that the following probabilities were computed given that a observable reversal occurred.

**Definition C.5.** $M_i^j = \sum_{n=i-1}^{j} d_n$. This is the maximum length of a reversal over markers i-j.

**Lemma C.2.1.** *The interval of possible lefthand cuts given l,i, and j is* $(min(0, l - L - d_j), min(l - L, d_i))$ *with* $l \in (L_i^j, M_i^j)$

*Proof.* First consider the right end point of the interval. There are two possibilities. Either l is so short that even with the reversal placed as far right as possible the right hand limit of $d_j$ is never encountered, or the opposite is true. This implies that the right end point of the interval is $min(0, l - L - d_j)$. Similarly there are two possiblities for the left end point. Either l is so short that it encounters the marker j or it is long enough that it encounters $d_i$ first. So the left end point of the interval is $min(l - L, d_i)$ Q.E.D.

**Definition C.6.** $K_i^j(l) = |(min(0, l - L - d_j), min(l - L, d_i))|$.

**Lemma C.2.2.** *For a reversal i-j and in an small range of l, a large portion of $K_i^j(l')$ overlap. The length of the intervals that does not overlap varies at most linearly with respect to l.*

*Proof.* With each $l \in (L_i^j, M_i^j)$ there is a region of possible lefthand cuts in $(0, d_i)$. Consider the $k \in (l+\epsilon, l-\epsilon)$. The k with a associated interval with the smallest right and left end point will be $l-\epsilon$. The k with a associated interval with the largest right and left end point will be $l+\epsilon$ $|(min(0, l + \epsilon - d_j) - min(0, l - \epsilon - d_j)| \leq 2\epsilon$ and $|min(l + \epsilon, d_i) - min(l - \epsilon, d_i)| \leq 2\epsilon$ Q.E.D.

Accordingly we shall take the simplfying assumption that in a small interval aroung l, the interval of left hand cuts is independent of the length of the reversal.

**Lemma C.2.3.** $P(x_1 \,|l, i, j) = \frac{1}{K_i^j(l)}$

*Proof.* Given one $p(i, j, l, x_1)$, we can slide the reversal to the left by a small epsilon. This results in $p(i, j, l, x_1 + \epsilon)$. Since our model veiws $p(i, j, l, x_1)$ and $p(i, j, l, x_1 + \epsilon)$ as equally likely, $x_1$ is uniformly distributed on K. Q.E.D.

**Lemma C.2.4.** $P(l \,|i, j) = \frac{\lambda e^{-\lambda(l)}}{e^{-\lambda(L_i^j)} - e^{-\lambda(M_i^j)}}$

*Proof.* $P(l \,|i, j) = P(l \,\Big|\, l \in (L_i^j, M_i^j)\Big) = \frac{P(l)}{P(l \in (L_i^j, M_i^j))}$ Q.E.D.

**Lemma C.2.5.** $P(i, j \mid$ *a reversal affecting at least one marker occurs*$) = \int_{L_i^j}^{M_i^j} \frac{K_i^j(l)}{\sum_{i \leq j \text{ with } i,j \in \mathbb{Z}_{n+1}^*} K_i^j(l)} \lambda e^{-\lambda(l)} dl$

*Proof.* $P(i, j \mid$ a reversal affecting at least one marker occurs$) = \int_{L_i^j}^{M_i^j} P(i, j \,|)$ a reversal affecting at least one marker occurs and l)P(l). The relative probability of l is $\lambda e^{-\lambda(l)}$. Given the length l, the choice of left hand cut is uniformly distributed over all the possible left hand cuts that result in a observed reversal. Thus $P(i, j \,|)$ a reversal affecting at least one marker occurs and l)$= \frac{K_i^j(l)}{\sum_{i \leq j \text{ with } i,j \in \mathbb{Z}_{n+1}^*} K_i^j(l)}$ Q.E.D.

**Theorem C.2.6.** $\sum_{i \leq j \text{ with } i,j \in \mathbb{Z}_{n+1}^*} P(i, j \mid$ *a reversal affecting at least one marker occurs*$) = \sum_{i \leq j \text{ with } i,j \in \mathbb{Z}_{n+1}^*} \int_{L_i^j}^{M_i^j} \frac{K_i^j}{\sum_{i' \leq j' \text{ with } i',}}$

1

*Proof.* Choose $l_1 \in \mathbb{R}^+$ with $l \notin \Big\{M_i^j\Big\} \cup \Big\{L_i^j\Big\}$. Form the interval $(l^-, l^+)$ where $l^- = max(\Big\{M_i^j < l\Big\} \cup \Big\{L_i^j < l\Big\})$ and $l^+ = min(\Big\{M_i^j > l\Big\} \cup \Big\{L_i^j > l\Big\})$.

Choose a second $l_2 \in \mathbb{R}^+$ with $l \notin \Big\{M_i^j\Big\} \cup \Big\{L_i^j\Big\} \cup (l^-, l^+)$. Form $(l_2^-, l_2^+)$.

Pictorially we consider the positive real line with markers at each $M_i^j$ and $L_i^j$. The $(l_i^-, l_i^+)$ represent the intervals of the disjoint finite partition of the real line by the markers. (The partition is finite as the number of markers is finite.)

**Definition C.7.** Let N equal the number of intervals in the above partition.

We know that $K_i^j(l) \neq 0$ on $(L_i^j, M_i^j)$ and is zero otherwise.

**Definition C.8.** $\overline{K}(l)$ =set of non-zero $K_i^j(l)$

The only way $\overline{K}(l)$ can change as l changes is to lose a $K_i^j$ or to gain a $K_i^j$. This requires that as l changes it encounters $L_i^j$ or $M_i^j$. Thus $\overline{K}(l)$ is constant on $(l^-, l^+)$, by construction.

**Definition C.9.** $\overline{K}_{l_\alpha}$ =set of non-zero $K_i^j(l)$ on $(l_\alpha^-, l_\alpha^+)$. Let $U = \left|\overline{K}_{l_\alpha}\right|$. Order the elements of $\overline{K}_{l_\alpha}$ i.e. the ith element of the set shall be denoted as $K_{l_{\alpha i}}$ with $i \in \mathbb{Z}_{U+1}^*$

This is well defined by the discussion above.
Thus $\sum_{i\in\mathbb{Z}_{N+1}^*} K_i^j = \sum_{i\leq j \text{ with } i,j\in\mathbb{Z}_{n+1}^*} K_i^j(l)$ on $(l_\alpha^-, l_\alpha^+)$.
Thus $\sum_{i\leq j \text{ with } i,j\in\mathbb{Z}_{n+1}^*} \int_{L_i^j}^{M_i^j} \frac{K_i^j(l)}{\sum_{i'\leq j' \text{ with } i',j'\in\mathbb{Z}_{n+1}^*} K_{i'}^{j'}(l)} \lambda e^{-\lambda(l)} dl =$
$\sum_{i\in\mathbb{Z}_{N+1}^*} \int_{l_i^-}^{l_i^+} \frac{\sum_{i'\in\mathbb{Z}_{U+1}^*} K_{l_{\alpha_{i'}}}(l)}{\sum_{i''\leq j'' \text{ with } i'',j''\in\mathbb{Z}_{n+1}^*} K_{i''}^{j''}(l)} \lambda e^{-\lambda(l)} dl = \sum_{i\in\mathbb{Z}_{N+1}^*} \int_{l_i^-}^{l_i^+} \lambda e^{-\lambda(l)} dl =$
$\int_0^\infty \lambda e^{-\lambda(l)} dl = 1$
All of this rearrangement is possible as the number of sums and integrals involved is finite as a direct consquence of the fact that the number of reversals is $n^2$. Furthermore the partition is in fact one of the entire positive real line as $\forall l \in \mathbb{R}^+ \; K_1^1(l) \neq 0$ by $d_0$ being infinite and $d_1$ is non-zero by definition.

Q.E.D.

Thus $P(i,j,l,x_1) = P(l, x_1 \,|i,j)\, P(i,j) = P(x_1 \,|l,i,j)\, P(l \,|i,j)\, P(i,j) = \int_{L_i^j}^{M_i^j} \frac{K_i^j(l)}{\sum_{i'\leq j' \text{ with } i',j'\in\mathbb{Z}_{n+1}^*} K_{i'}^{j'}(l)} \lambda e^{-\lambda(l)} dl \frac{1}{K_i^j(l)} \frac{}{e^{-}}$

*Remark* C.1. If one's purposes do not allow, the assumption of the neglible length of the genes can be thrown out. This paper shall only breifly outline the changes required. Note also that the section of uniform sampling is independent of the section on probability. Provided the user can work out the probability of a given reversal in $\pi$, the rest of the paper can still be used even under a different probability distribution.

**Definition C.10.** Let $l_i$ represent the length of gene $i \in (1, ..., n)$.

**Definition C.11.** $M' = \sum_{n=i-1}^{j} d_n + \sum_{n=i}^{j} l_n$

**Definition C.12.** $L' = \sum_{n=i}^{j-1} d_n + \sum_{n=i}^{j} l_n$

The vaule of $K_i^j(l)$ remains the same. So does the probability distribution above. The proof of theorem 2.8 still holds with slight alteration. The interval (0,min$\left\{L_i^j\right\}$) should not be represented in the partition. Or rather it can be but the values of K in that interval are zero for all i-j reversals. In the proof of theorem 2.8, the markers of L and M are merely moved around. However we must now normalize the probability since the minimum possible length of a reversal is no longer zero but the length of the smallest gene.

## C.3 Overview of previously constructed algorithms

**Definition C.13.** Given $\pi$, a hurdle is a section of the permutation of the form $\pi_j+1 = i, ...., \pi_{j+k-1} = i+k$ with all $\pi_t \in \{\pm(i+1), ..., \pm(k-1)\}$ $\forall t \in \{j+2, ..., j+k-2\}$ and s.t. there is no smaller segment of this section for which the above property holds. e.g. (2,-4,-3,5) is a hurdle of (6,2,-4,-3,5,1)

We can identify a signed permutation to an unsigned one in the following manner: $\pi_i$ goes to $2\pi_i - 1, 2\pi_i$ if by is positive and to $2\pi_i, 2\pi_i - 1$ is it is negative. e.g. $(0, 6, 2, -4, -3, 5, 1, 7) \mapsto (0, 11, 12, 3, 4, 8, 7, 1, 9, 10, 1, 2, 13)$. 0 and 7 are added merely as placeholders and perform that same function in the unsigned permutation, so $0 \mapsto 0$ and $n+1 \mapsto 2(n+1)-1$. It is clear that if we restrict what types of reversals are allowed on the unsigned permutation to those that do not seperate the pairs $2\pi_i, 2\pi_i - 1$, minimal distance on the unsigned permuation is the same as that on the signed permuation. The breakpoint graph will be constructed from the unsigned permutation. Say that $i \approx j$ when $|i-j| = 1$. Call $(\pi_i, \pi_j)$ a break point when $i \approx j$ and but not $\pi_i \approx \pi_j$. Define a breakpoint graph to be the vertices $\pi_0$ thru $\pi_n$ with as black edge connecting $\pi_i$ to $\pi_j$ if it is a breakpoint and a grey edge connecting $\pi_i$ to $\pi_j$ if it is a break point of the inverse permutation. i.e. $\pi_i \approx \pi_j$ but not $i \approx j$. We are interested in alternating cycles, one edge grey the next black the next grey, in the breakpoint path. It was proven by Bafna and Pezner that the maximum cycle decomposition of the breakpoint graph is an important factor in minimal reversal distance.

In [4],it was proved that $b(\pi) - c(\pi) + h(\pi) \leq d(\pi) \leq b(\pi) - c(\pi) + h(\pi) + 1$

Where $b(\pi)$ is the number of breakpoints,$h(\pi)$ is the number of hurdles, and $c(\pi)$ is the maximum cycle decomposition.

1. Bader's algorithm for calculating $d(\pi)$ [5]

This algorithm give the distance and the number of hurdles and superhurdles along with the cycles that they contain. Superhurdles are those whose destruction greates new hurdles.

Using a few facts from [6], Istvn Mikls and Aaron E. Darling managed a very efficent method of enumerating possible sorting reversals with a low probability of false acceptance.

1. The reversals of a permutation with no hurdles are only those reversal that are cycle increasing.
2. The reversals of a permutation with only one hurdle is either cycle increasing or hurdle cutting.
3. The reversals of a permutation with two or more hurdles is either cycle increasing, hurdle cutting, or hurdle merging.

Furthermore

1. Cycles increasing reversals act on a single cycle with reality edges of opposite orientation.
2. Hurdles cutting reversals act on a cycle within a single hurdle.
3. Hurdles merging reversals act on the a cycle with the end points within, even on the edges of, two different hurdles.

Together these facts make an efficent though non-deterministic algorithm possible.

## C.4 MCMC algorithm for uniform sampling of R length paths from $\pi$ to the identity

In this section, the distance of a reversal to the identity will be mentioned often. This is to be understood under the metric of minimum reversal distance. Consider a tree constructed in the following way, start with $\pi$. In the next level, add nodes for each sorting resersal on $\pi$. Each node i in the second level corresponds to a path $p_i$. attach to this node all the sorting reversals for $\pi p_i$. Continue for $d(\pi)+1$. Each path on this tree corresponds with a path from $\pi$ to the identity. An MCMC method for sampling uniformly from this tree of minimum paths was created by Istvn Mikls and Aaron E. Darling. That method is enumerated below. All that the I claim is that which is necessary for the extension of that method to the sampling of non-minimal paths. Now consider the extension of such a tree in the following way. If the path lengths are to be over $d(\pi)$, then each node in the tree can be connected to a number of reversals not all of which will be sorting. The way such a tree would have to be constructed should be clear from the algorithm below.

Given a path from $\pi$ to the identity of length R, denote $C_i$ to be the ith long chain of that path i.e. if $\pi p_1 p_2 p_3 p_4 p_5 p_6$ is a full path, then $C_3$ is $\pi p_1 p_2 p_3$. Let $C_0$ is $\pi$.

**Definition C.14.** Given $\pi$, S=$\{p \mid$ p is a reversal and $d(\pi p) < d(\pi)\}$

**Definition C.15.** Given $\pi$, U=$\{p \mid$ p is a reversal and $d(\pi p) > d(\pi)\}$

**Definition C.16.** Given $\pi$, O=$\{p \mid$ p is a reversal and $d(\pi p) = d(\pi)\}$

**Lemma C.4.1.** *For $d(\pi) = 1$, $O_{(\pi)} = \emptyset$.*

*Proof.* The form of $\pi$ and all permutations with minimal distance one is $\overset{+}{\rightarrow}\overset{-}{\leftarrow}\overset{+}{\rightarrow}$, where the arrow represents a string of consecutive numbers and the sign of the string is recorded above. i.e. (1,2,3,4,-7,-6,-5,8,9) is a possible $\pi$

The forms of all possible $\pi p_1$ are the following:
$\overset{+}{\rightarrow}\overset{-}{\leftarrow}\overset{-}{\leftarrow}\overset{+}{\rightarrow}$
$\overset{+}{\rightarrow}\overset{-}{\leftarrow}\overset{+}{\rightarrow}\overset{-}{\rightarrow}\overset{+}{\leftarrow}$
$\overset{+}{\rightarrow}\overset{+}{\rightarrow}\overset{-}{\leftarrow}\overset{-}{\leftarrow}\overset{+}{\rightarrow}$

Since breaks in arrow of the same sign represent breakspoints in the permutation, we conclude that none of the possible forms of $\pi p_1$ match those of any permutation $\alpha$ with $d(\alpha) = 1$. Q.E.D.

This is not to say or suggest that $O_\pi$ is always empty. Let a permutation with no superhurdles or double hurdels be given. The reversal of the entirity of a small isolated, as in not interweaved with any other cycle, cycle of the breakpoint path with breakpoints at both ends will be a level reversal providing the resultant cycle also has breakpoints at both ends. This is so as the number of breakpoints, internal to the cycle on the edges and outside, is unchanged. As is the maximum cycle decomposition, and the number of hurdles. This is not an irrelevant point as the existance of level moves allows for different types of paths. In effect the level reversal can be a waiting move in the paths progress to the identity. Furthermore since every reversal moves the minimal distance of a permutation out or in by at most one, all paths to the identity without a level are odd when $d(\pi)$ is odd and even otherwise. Thus without a level move a path of length 50 from $\pi \rightarrow$ identity with $d(\pi) = 17$ does not exist.

1. Construct a sample chain $C_i$ for $0 \leq i \leq R$. $C_1 = \pi$ and $C_R$ =identity. Compute $d(C_i)$ for each sub-chain. For each $C_i$ associate a vetor in $\{-1, 0, 1\}^i$. This vector gives the nature of each reversal in the

$C_i$ sub-path. If the jth vector component is 0, then the jth reversal of the path is an element of the O and so forth. Denote said vector as $v^{C_i}$ and the jth component of the vector as $v_j^{C_i}$

2. A swapping operation will be performed on $C_i$ and $C_{i+1}$. This operation will be defined as follows: $C_{i+1} \mapsto C_i'$ by the removal of the last reversal of $C_{i+1}$.

$$d(C_i') = \begin{cases} d(C_{i+1}) + 1 & \text{if } v_{i+1}^{C_{i+1}} = -1 \\ d(C_{i+1}) - 1 & \text{if } v_{i+1}^{C_{i+1}} = 1 \\ d(C_{i+1}) & \text{if } v_{i+1}^{C_{i+1}} = 0 \end{cases} \tag{C.1}$$

At the same time $C_i$ will be extended to $C_{i+1}$ in the following way.
Store R-i=r.
For $r \neq 2$ or r=2 and $d(C_i) \neq 1$

- While $d(C_i) = r$, form S by algorithm and choose uniformly. It should be noted that S contains a small percentage of non-sorting reversals. Once a reversal $(p_{i+1})$ is chosen, Form $C_{i+1}' = C_i \circ p_{i+1}$, where $\circ$ represents the concatenation of $C_i$ with $p_{i+1}$. Now calculate $d(C_{i+1}')$. If $d(C_{i+1}') = d(C_i) - 1$, accept canditate reversal. Reject it otherwise and choose a new candidate.

- While $d(C_i) = r - 1$, use the naive algorithm of enumerating posible reversals affecting them and calculating the new distance to form the sets S,U, and O. This algorithm works in $n^3$. Choose from the union of S and O. Calculate the $d(C_{i+1}')$.

- While $d(C_i) < r - 1$, choose uniformly over all possible reversals and then calculate the new distance.

For $r = 2$ and $d(C_i) = 1$
Backtrack through one accepted reversal, and calculate O again. Continue until.

1. $d(C_j) \neq R - j$. There are two possible cases.In either case we now choose a new p'(j+1) out of the $O \cup U - \{p_{j+1}\}$ , thereby avoiding the dilemma above.

   - d$(C_j) = R - j - 3$ and $p_{j+1}$, the reversal just removed, is an element of U
   - d$(C_j) = R - j - 2$ and $p_{j+1}$ is an element of 0

2. If O is found to be non-empty first, then at that swapping of chains an extension out of O is chosen.

This is a monte carlo markov chain. First any swap can be undone, so reversibility holds. Secondly note that a state of this markov chain is a vector $X = (C_1, ..., C_R)$. Since any chain may be grown by a series of swaps from $C_O$ without altering the chains of greater index, irreducibility holds.

The result of this swapping operation is accepted when u chosen uniformly from (0,1) and $u \leq \frac{P(C_i' \circ p_{i+1} | C_i')}{P(C_i \circ p_i | C_i)}$. That is $\frac{P(C_i' \circ p_{i+1} | C_i')}{P(C_i \circ p_i | C_i)} = \frac{P(X|Y)}{P(Y|X)}$ with $X = (C_1, ., C_i, C_{i+1}, .., C_R)$ and $Y = (C_1, ., C_i', C_{i+1}', .., C_R)$. $C_i' = C_{i+1} / \{p_{i+1}\}$ and $C_{i+1}' = C_i \circ p_{i+1}'$ where $p_{i+1}'$ is an accepted extension. That is X and Y are related by a chain swap. Now given $C_{i+1}$ the new $C_i'$ is determined. The probability of a particular extension is 1/(Q) with Q being the set of possible steps from $C_i$ given the time left and the distance from the identity of the position of the permutation at the end of the chain. This Q varies a good deal, its composition is given in the second step of the algorithm. It is clear that $\frac{P(C_i' \circ p_{i+1} | C_i')}{P(C_i \circ p_i | C_i)} = \frac{|Q_i|}{|Q_{i+1}|}$. In conclusion, given two full chains

each is the last component of the same number of elements. Thus since this algorithm samples X's uniformly it also samples all full paths uniformly.

[4] showed that the mixing time of his algorithm was the following.

Let u a node in the jth level of the tree. Let $L_j(u)$ denote the number of extension of node u,$D_i$ be the number of node is the ith level of the tree.

**Definition C.17.** Let k= $max_i max_{u \in D_i} max_{j>i} \frac{|L_j(u)||D_i|}{|D_j|}$ be called the hiddenness rate.

[4] showed that the mixing time was $O(hR^{2+log_2(9k)})$ with h equal the inverse of the smallest transition probability. I conjecture that my algorithm mixes in less than $O((hR^{2+log_2(9k)})^3+E)$ where E is the expected time spent backtracking through the tree. Furthermore I conjecture that a level reversal is likely to exist at a level greater than or equal to R-j for some $j < 10$. If this could be demonstrated, a probability of the accepting a node outside of the tree could be calculated as well as the time spend on that bad path. This would allow one to estimate E.

## C.5 Enumerating parsimonious orderings of blocks

This section is only a slight extension of Gaul and Blanchette. It is included here merely as a point of interest.

Gaul and Blanchette solved the block ordering problem. Define a block as a small ordered set of gene s.t. the genes are adjancent in the genome. e.g. [c,d,e,f,b] has a block decomposition [c] [d,e] and [f,b]. Usually the blocks [a] and [n] are added as place holders. [a] at the beginning of each possible ordering and [n] at the end. Given two sets of blocks, one for genome A and one set for genome B, what is the simultaneous ordering of blocks that maximizes the cycles in the resultant breakpoint graph? Roughly this means that given partial information regarding genome A and genome B, which possible genomes are the most likely pair. This is the block ordering problem.

In this section, the number of blocks of genome A is assumed to be one. That is the entire genome A is given. Genome B is still fragmented. The problem now is to enumerate all the parsimonious possible genome B.

A glance at Gaul and Blanchette will show that their work readily extends to the case outline above. This paper will not give the program for enumeration explicitly.

First note that the Edge Matching Graph defined in Gaul and Blanchette has only the components a—n and b—f. The exact graphical style of that paper has not been adopted as the two components have only two markers. A quick examination will reveal the following.

**Lemma C.5.1.** *The only possible orders of the edge maching algorithm are [a][b...f][n] and [a'][b'...f'][n'] or [a][-f...-b][n] and [a'][-f'...-b'][n'].*

Since [a] and [n] are place holders, genome B is being considered with respect to genome A and [b...f] is the entire genome A, the two possiblities in the lemma above are equivalent for our purposes.

For all one sided components, list the one sided edges. Label the one sided components 1 thru n. Label [b'...f'] as n+1.

Now Gaul and Blanchette proved that while ordering the blocks of one-sided components which edge is sacrificed is irrelevant. Also the reading, negative or positive, is irrelevant. Thus a program that can list the elements of $S_{n+1} \times \{-1,1\}^n \times \mathbb{Z}_{k_1} \times .... \times \mathbb{Z}_{k_n}$ with $k_n$ denoting the number of edges in the ith one-sided component, will have listed all the parsimonious orders.

$S_{n+1}$ gives the order of the one-sided components and [b'...f']. $\{-1, 1\}^n$ gives the sign of each of the one-sided components. -1 in the ith position means to read the ith one-sided component backwards. 1 means to read it forwards. $\mathbb{Z}_{k_1} \times .... \times \mathbb{Z}_{k_n}$ gives the edge to be left out of each one-sided component.

# Bibliography

1. Eric Gaul and Mathieu Blanchette. Ordering Partially Assembled Genomes Using Gene Arrangements, *McGill centre for Bioinformatics, McGill University*

2. Anne Bergeron. A Very Elementary Presentation of the Hannenhalli-Pevzner Theory, *University of Montreal*

3. Sridhar Hannenhalli and Pavel A. Pevzner. Transforming Cabbage into Turnip: Polynomial Algorithm for Sorting Signed Permutations by Reversals, *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing (Las Vegas, Nev., May 29-June 1* (1995) 178-189

4. Istvn Mikls and Aaron E. Darling. Efficient Sampling of Parsimonious Inversion Histories with Application to Genome Rearrangement in Yersinia http://gbe.oxfordjournals.org/cgi/content/full/2009/0/153 (2009 )

5. Bader DA, Moret BM, Yan M. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J Comput Biol (2001) 8:483491*

6. Siepel AC. An algorithm to enumerate sorting reversals for signed permutations. J Comput Biol (2003) 10:575597

# On the Growth of the Basilica Group

Ziva Kaye Myer
New College of Florida

Indiana University REU Summer 2009
Advisor: Kevin Pilgrim

# D.1 Introduction

It was shown algebraically in 2002 in a Grigorchuk-Zuk paper [2] that the Basilica Group $G =< a, b >$ has exponential growth. The proof showed by contradiction that the semi-group generated by $a$ and $b$ is free by examining cases with two minimal words representing the same element and contradicting their minimality.

The proof presented in this paper is similar in set-up to the Grigorchuk-Zuk proof, but uses Schreier Graphs to prove geometrically that no two words in positive powers of $a$ and $b$ represent the same group element. This intuitive method may be able to be applied to show other groups have exponential growth.

# D.2 Growth of Groups

Let $S = \{s_1, \dots s_k\}$ be a symmetric finite generating set of a group $G =< S >$. Then each $g \in G$ can be written as $g = s_{i1} \dots s_{il}$. The shortest such decomposition is called the length of the element, denoted $l(g) = l_S(g)$. The *growth function* $\gamma(n) = \gamma_G^S(n)$ is the number of $g \in G$ such that $l(g) \leq n$.

Consider two growth functions $\gamma, \gamma' : \mathbb{N} \to \mathbb{N}$ and define $\gamma \preccurlyeq \gamma'$ if $\gamma(n) \leq C\gamma'(\alpha n)$ for all $n > 0$ and some $C, \alpha > 0$. $\gamma$ and $\gamma'$ are equivalent, denoted $\gamma \sim \gamma'$ if $\gamma \preccurlyeq \gamma'$ and $\gamma' \preccurlyeq \gamma$.

**Theorem D.2.1.** *If $S$ and $S'$ are two finite generating sets of a group $G$ then $\gamma_G^S \sim \gamma_G^{S'}$*

*Proof.* Let $S = \{s_1, \dots, s_k\}$ and $S' = \{s'_1, \dots, s'_{k'}\}$. Since both sets generate G, we can write each $s_i \in S$ as a word in elements of $S'$. Let the length each such word be $c_i$, i.e., $l_{S'}(s_i) = c_i$. Define

$$C = max\{c_i\}, 1 \leq i \leq k$$

so that $l_S(g) \leq C l_{S'}(g)\ \forall g \in G$. This implies that $\gamma^{S'}(n) \leq \gamma^S(Cn)$. A similar argument shows that $\gamma^S(n) \leq \gamma^{S'}(C'n)$, proving the growth functions to be equivalent. Q.E.D.

We can now define the different types of growth on a function $f : \mathbb{N} \to \mathbb{R}$. A function f is called *polynomial* if $f(n) \sim n^\alpha$ for some $\alpha > 0$. A function is called *exponential* if $f(n) \sim e^n$.

A group has *superpolynomial growth* if

$$\lim_{n\to\infty} \frac{ln\gamma(n)}{ln(n)} = \infty$$

A group has *subexponential growth* if

$$\lim_{n\to\infty} \frac{ln\gamma(n)}{n} = 0$$

If a group has both superpolynomial and subexponential growth, we say it has *intermediate growth.* [1]

# D.3 Automorphisms of the Infinite Binary Tree

Let $\mathbf{T}$ be an infinite binary tree, as in Figure D.1.

The root of the tree, denoted $\mathbf{r}$ corresponds to the empty word $\varnothing$. V os the set of all vertices $v$, where a vertex of level $n$ denoted by $\mid v \mid = n$ is a word in $\{0, 1\}^n$. $E$ is the set of all edges, and by definition $(v, w) \in E$ if $w = v0$ or $w = v1$. We denote $\mathbf{T}_v$ to be the subtree of $\mathbf{T}$ rooted at the vertex v, and we see $\mathbf{T}_v$ is isomorphic to $\mathbf{T}$.

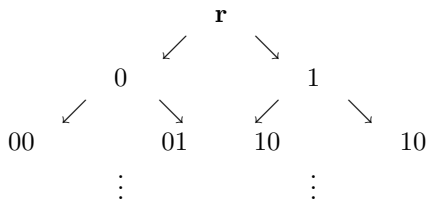



Figure D.1: Infinite Binary Tree

We are concerned with Aut($\mathbf{T}$), the group of bijections of vertices which map edges into edges and preserve the root and level of vertices. $I \in$ Aut($\mathbf{T}$) is the identity automorphism. The simple but extremely important automorphism $\sigma \in$ Aut($\mathbf{T}$) corresponds to a swap at the first level, and maps $\mathbf{T}_0$ into $\mathbf{T}_1$. Precisely, $\sigma$ is defined as $\sigma(0w) = 1w$ and $\sigma(1w) = 0w$ for any word $w$ in 0 and 1.

Since the subtrees $\mathbf{T}_0$ and $\mathbf{T}_1$ are isomorphic to $\mathbf{T}$, we can identify an automorphism of the tree as a swap or identity on the first level along with automorphisms on $\mathbf{T}_0$ and $\mathbf{T}_1$. Formally, this is expressed through the following isomorphism, also called a *wreath product*:

$$\phi : (Aut(\mathbf{T}) \times Aut(\mathbf{T})) \rtimes \mathbb{Z}_2 \longrightarrow Aut(\mathbf{T})$$

$$\varepsilon(g_0, g_1) \longmapsto g, \quad \varepsilon \in \{I, \sigma\}$$

Using $\phi$, we can define finitely generated subgroups of Aut($\mathbf{T}$) by defining the generating elements recursively. Note the property that

$$\sigma(g_0, g_1)\sigma = (g_1, g_0)$$

## D.4 Schreier Graphs

Schreier graphs can assist in understanding the properties, structure, and growth of these finitely generated subgroups of Aut($\mathbf{T}$). In this context, a Schreier graph of a level $n$, denoted $\Gamma_n$ consists of vertices that are all the words in $\{0,1\}^n$ and edges corresponding to each generating element acting on each vertex. In other words, for each vertex $v$ and generator $s$, there exists an edge from $v$ to $s \cdot v$. This edge is labeled with the generator $s$.

Certain subgroups of $Aut(\mathbf{T})$ have the special property that at all levels $n$, we can draw $\Gamma_n$ so that it is *planar*, i.e., so that edges only intersect at their endpoints. In addition, we draw $\Gamma_n$ so that the local picture in the plane is the same at every vertex.

## D.5 The Basilica Group

The remainder of this paper focuses on one subgroup of Aut($\mathbf{T}$) that is generated by two elements, the Basilica Group $G =< a, b >$ with $a$ and $b$ defined recursively in the following way using the above automorphism $\phi$.

$$a = \sigma(b, I)$$

$$b = (a, I)$$

In another notation, for any word $w$ in 0 and 1:

$$a(0w) = 1b(w) \qquad a(1w) = 0w$$

$$b(0w) = 0a(w) \qquad b(1w) = 1w$$

*Example* D.1. Some simple level 5 computations:

$$a(00101) = 1b(0101) = 10a(101) = 10001$$

$$b(10100) = 10100$$

Since we will be concerned with the vertex of $n$ 0's, the following two examples show how $a$ and $b$ act on $v = 00000$:

$$a(00000) = 1b(0000) = 10a(000) = 101b(00) = 1010a(0) = 10101$$

$$b(00000) = 0a(0000) = 01b(000) = 010a(00) = 0101b(0) = 01010$$

The Basilica Group Schreier Graphs, when drawn with the properities outlined in Section D.4, are useful tools in understanding properties of the growth of the group. Each word in $a, b, a^{-1}, b^{-1}$ together with a starting vertex $v$ gives a *path* in $\Gamma_n$, with $a$ and $b$ going forward (i.e., following the directions of the arrows) and $a^{-1}$ and $b^{-1}$ going backwards. A *cycle* in $\Gamma_n$ is defined as a forward path starting and ending at the same vertex.

**Proposition D.5.1.** *The shorest cycles in Basilica Schreier Graphs are those of word in $< a >$ and $< b >$, i.e., of only one generating element. Furthermore, there exists a vertex $v \in \{0,1\}^n$, namely $v = 0 \dots 0$ ($n$ 0's) that disconnects $\Gamma_n$ into two pieces, $L_n$ and $R_n$.*

This is a consequence of results in [3].

## D.6 Exponential Growth of the Basilica Group

**Theorem D.6.1.** *The Basilica Group $G =< a, b >$ as defined above has exponential growth.*

This follows immediately from the following Lemma:

**Lemma D.6.2.** *No two words in positive powers of $a$ and $b$ represent the same group element in $G$.*

*Proof.* Let $w_1$, $w_2$ be different words in positive powers of a and b that represent the same group element $g \in G$. Then without loss of generality we may assume $w_1 = w_1' a$ and $w_2 = w_2' b$. Let

$$L = max\{ \mid w_1' \mid, \mid w_2' \mid \}$$

We want to show that, for a large enough $n$, $w_1$ and $w_2$ acting on the $v$ defined in Proposition D.5.1 will be in different pieces of the graph, a contradiction of their assumed representation the same group element. This can be shown if we pick an $n$ large enough that the cycles in only $a$ and only $b$ starting at $v$ are longer than $2L$, as in:

$$\#\{< a > \cdot v\} \ > 2L \quad and$$

$$\#\{< b > \cdot v\} \ > 2L$$

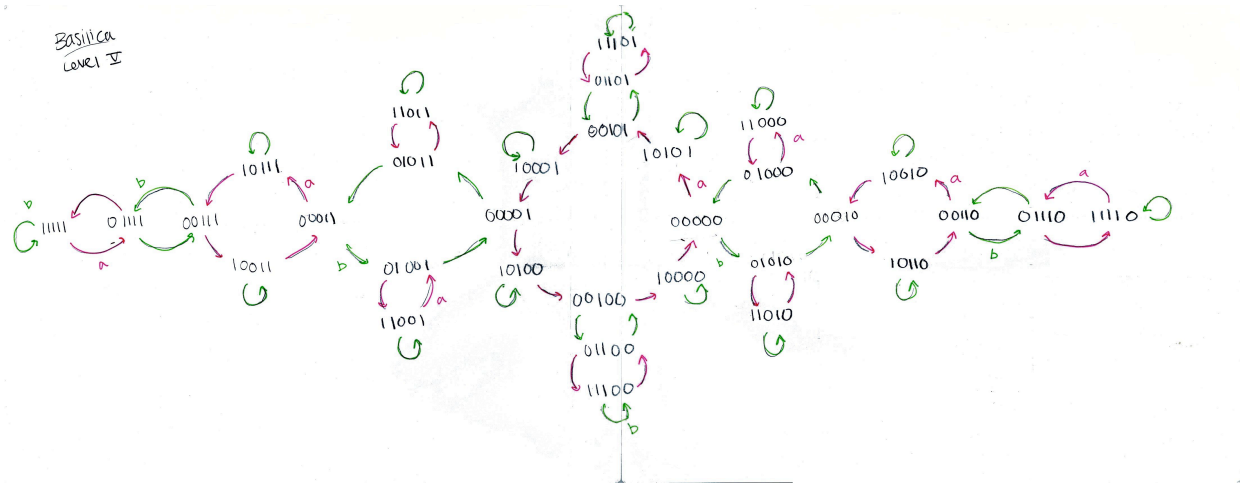



Figure D.2: Basilica Schreier Graph Level 5

For such an $n$, we want to show that, for any word $w$ of length $\leq L$ in positive powers of $a$ and $b$

$$wa \cdot v \neq v \ \ and \ \ wb \cdot v \neq v$$

Since $a \cdot v$ and $b \cdot v$ are in the seperate pieces of $\Gamma_n$, if the above is true, we will not be able to get back to $v$ and thus stay in seperate pieces of the graph. Q.E.D.

**Proposition D.6.3.** *Let $v = 0 \dots 0$ ($n$ 0's, $n$ even). The the orbit of $v$ under $a$ has length $2^{n/2}$.*

*Proof.* We see that it is true for $n = 2$ with orbit $00 \to 10 \to 00$. Assume it is true for $n = k$, $k \geq 2$, $k$ even, i.e., the cycle of $w_k = 0 \dots 0$ (k 0's) under $a$ has lenth $2^{k/2}$. We want to show that the cycle of $00w_k$ under $a$ has length $2^{(k+2)/2}$. This orbit is

$$00w_k \to 10a(w_k) \to 00a(w_k) \to 10a^2(w_k) \to \dots \to 10w_k \to 00w_k$$

twice as long as the orbit for $w_k$ since for each vertex $v_a$ in the cycle of $w_k$ under $a$, there exists a vertex at $00v_a$ and $10v_a$. The result follows from induction. Q.E.D.

A similar argument shows the same result for the cycle of $v$ ($n$ even) under $b$. From these results, we see that can pick a large enough $n$ such that

$$\#\{< a > \cdot v\} \ > 2L \quad and$$

$$\#\{< b > \cdot v\} \ > 2L$$

**Proposition D.6.4.** *If $n$ is large enough (as defined above for instance), then for any word $w$ of length $\leq L$ in positive powers of $a$ and $b$*

$$wa \cdot v \neq v \ \ and \ \ wb \cdot v \neq v$$

*Proof.* Suppose for some word $w$, $wa \cdot v = v$ and $\mid w \mid \leq L$. In the set

$$\{wa \mid wa \cdot v = v, \ \mid w \mid \leq L\}$$

pick the shortest element. $wa \cdot v$ gives a cycle from $v$ to itself and we see from Section D.5 that the shortest cycle is that of $w = a \dots a$. Since we chose a large enough $n$ as defined above, we see that it is impossible to get back to $v$, a contradiction. Q.E.D.

## D.7 Conclusion and Further Work

While exponential growth of the Basilica Group has already been proven [2], the connection between the geometric Schreier Graphs and growth of groups is significant. A similar method could most likely be extended to prove the exponential growth of the Rabbit Group, generated by three elements:

$$a = \sigma(I, c)$$

$$b = (I, a)$$

$$c = (I, b)$$

Furthermore, it is to be seen whether there exists a connection between the structure of Schreier Graphs and intermediate growth. Perhaps there is a geometric method to discovering the growth of groups whose growth is still unknown.

## D.8 Acknowledgements

I would like to thank Dr. Kevin Pilgrim for his expert advising and for directing the REU and Mandie McCarty for all the help and snacks she provided. In addition, I would like to thank the NSF for funding and Indiana University for hosting this REU.

## Bibliography

1. Grigorchuk and Pak Groups of Intermediate Growth, an Introduction *L'Enseignement Mathmatique* **54** (2008) 251-272

2. Grigorchuk and Zuk On a torsion-free weakly branch group defined by a three state automaton *Internat. J. Algebra Comput.* **12** (2002) 1-2, 223–246

3. Nekrashevych *Self-similar groups*, American Mathematical Society, Providence, RI, 2005

# Counting Involutions in Finite Groups

Zach Norwood
University of Nebraska, Lincoln

Indiana University REU Summer 2009
Advisor: Allan Edmonds

# E.1 Introduction

Few make it through an introductory algebra class without seeing the following exercise:

*Exercise.* If $x^2 = 1$ for all $x \in G$, prove that $G$ is abelian; more specifically, $G$ is an elementary abelian 2-group.

The inquisitive algebra student, perhaps having seen other elementary exercises where this is the case, might wonder whether the conclusion of this exercise would follow from a weaker hypothesis. This turns out to be true; Theorem E.5.1 demonstrates that only three fourths of a finite group's elements need be involutions to guarantee it to be elementary abelian.

Before proving this main result, we collect a variety of interesting facts concerning both methods for explicitly determining the number of involutions in a group and also the structure of groups with many involutions. As many of these results were motivated by data collected using GAP, we follow that with a discussion of those data, much of which is available in the Appendix. Some of the accompanying GAP code is also included in the Appendix; the rest is available here[1].

We assume the reader has some familiarity with the basics of finite group theory; The first five chapters of [2], in particular, are a good place to start learning that material. However, for the convenience of the reader, we provide some key points from basic group theory here:

**Definition E.1.**

1. The *center* of a group $G$, denoted $Z(G)$, is $\{ x \in G \mid xy = yx \text{ for every } y \in G \}$.
2. The *centralizer* of an element $g$, denoted $C(g)$ or $C_G(g)$, is the set of all elements that commute with $g$. That is, $C(g) = \{ x \in G \mid xg = gx \}$.
3. A subgroup $H$ of $G$ is *normal* in $G$ (denoted $H \trianglelefteq G$ if it is preserved under conjugation. That is, $H \trianglelefteq G$ if $gHg^{-1} = \{ ghg^{-1} \mid g \in G, h \in H \} = H$.
   When $H$ is normal, the coset space $G/H$ is itself a group (called the *quotient group*) under the operation given by $(aH)(bH) = (ab)H$.
4. If $G$ is acting on a set $S$, then the *stabilizer* of $s \in S$, denoted $\text{stab}(s)$, is $\{ g \in G \mid g \cdot s = s \}$. The *orbit* of $s$ under the action, $\text{orb}(s)$, is $\{ t \in S \mid \text{For some } g \in G,\ g \cdot s = t \}$.
5. If $H$ is a subgroup of $G$, then we call $|G|/|H|$ the *index* of $H$ in $G$ and denote it by $[G : H]$.

$Z(G)$, $C(g)$, and $\text{stab}(s)$ are all subgroups of $G$ (check this!). We call a group "abelian" when $Z(G) = G$. Also, when $G$ is acting on itself by conjugation, we denote the orbit of an element $g$ under this action by $\text{cl}(g)$ for "conjugacy class" of $g$. This is well-defined, as orbits partition a set:

*Exercise.* Let $G$ be a group acting on a set $S$, and let $s$, $t \in S$. Prove that $\sim$, given by $s \sim t$ if $s \in \text{orb}(t)$, is an equivalence relation.

**Theorem E.1.1** (Lagrange's Theorem)**.** *If $H$ is a subgroup of a finite group $G$, then $|H| \big| |G|$.*

Thus the index of a subgroup is always a natural number.

**Theorem E.1.2.** *If $G$ is acting on a set $S$ and $s \in S$, then* $\text{stab}(s) \leqslant G$ *and* $[G : \text{stab}(s)] = |\,\text{orb}(s)|$.

We will be interested in Theorem E.1.2 in a situation wherein $G$ will be acting on itself by conjugation:

[1] idisk.me.com/zbnorwood1/public

**Corollary E.1.3.** $|\operatorname{cl}(g)| = [G : C(x)]$.

**Definition E.2.** It will be especially important for the reader to be familiar with the following groups:

1. The *cyclic* group of order $n$, denoted $C_n$, is a group generated by an element of order $n$: $\{ a^i \mid 0 \leq i \leq n-1 \}$.
2. The *dihedral* group of order $2n$, $D_{2n}$, is the group of symmetries of the $n$-gon. It is usually given by the presentation
$\langle r, s \mid r^n = s^2 = 1,\ srs = r^{-1} \rangle$.
3. The *symmetric* group on $n$-symbols, denoted $S_n$, is the group of permutations of a set of cardinality $n$. The *alternating* group $A_n$ is the subgroup (of index 2) of $S_n$ consisting of exactly the even permutations—products of an even number of transpositions—in $S_n$.

We also refer to cyclic subgroups of groups; for example, the cyclic subgroup generated by an element $a$ of order $n$ is isomorphic to $C_n$ and will be denoted by $\langle a \rangle$.

*Exercise.* Prove that $|S_n| = n!$.

*Exercise.* Let $n$ be a natural number, and let $a$ be a generator of $C_n$. Show that the map $\phi : C_n \to \mathbb{Z}/n\mathbb{Z}$ given by $a^i \mapsto i$ is an isomorphism.

*Exercise.* Describe the conjugacy classes in $D_{2n}$.

**Definition E.3.** The *direct product* of groups $G$ and $H$, denoted $G \oplus H$, is their Cartesian product under componentwise multiplication.
Two groups $G$ and $H$ are *isomorphic*, and we write $G \cong H$, if there exists a bijective homomorphism between the two groups.

For example, $C_{mn} \cong C_m \oplus C_n$ if $m$ and $n$ are relatively prime.

**Theorem E.1.4** (Fundamental Theorem of Cyclic Groups)**.** *Let $C_n$ be the cyclic group of order $n$. Then, for each $d$ dividing $n$, $C_n$ contains exactly one subgroup of order $d$.*

**Theorem E.1.5** (Fundamental Theorem of Finite Abelian Groups)**.** *Every finite abelian group is the direct product of cyclic groups.*

*Exercise.* Every subgroup of an abelian group is normal.

## E.2 Definitions and Conventions

In light of the extensive variation among notational conventions used in algebra, here we state explicitly which conventions we will adopt. First, $D_{2n}$ (not, as is the case in many texts, $D_n$) will denote the dihedral group of order $2n$. (For example, $D_8$ will be the group of symmetries of the square.) The cyclic group of order $n$ will be denoted by $C_n$, groups will always be written multiplicatively, and $\oplus$ will always be used to denote an external direct product of groups. (Note that the tables in the Appendix use $\times$, however.) To denote the group identity we will use 1 or, when more than one group is being addressed, $1_G$. Finally, $|A|$ will serve double duty as the order of a group $A$ if $A$ is a group or the cardinality of a set $A$ if $A$ has no algebraic structure.

The following definition will be of central importance:

**Definition E.4.** For any group $G$ define

$$J(G) := \{\, x \in G \mid x^2 = 1_G \,\} \text{ and } j(G) := |J(G)|.$$

When $G$ is the only group under consideration, $J$ and $j$ might be used in place of $J(G)$ and $j(G)$, respectively.

Although this definition clearly applies to infinite groups, we will, unless otherwise noted, be dealing with only finite groups in this paper.

It is worth noting here that, since we use "involution" to describe an element of order 2, $J(G)$ is actually the set of the involutions (for our purposes, elements of order 2) in $G$ *and* the identity. The convenience adopting this convention provides is evident in Lemmas E.3.4 and E.4.4, for example.

We might, with this new terminology, rephrase Exercise E.1:

*Exercise.* If $G = J(G)$, prove that $G$ is an elementary abelian 2-group.

In our study of groups more than half of whose elements are involutions, it will be convenient to refer to a generalization of the dihedral group. We define that now:

**Definition E.5.** If $G$ is an extension of an abelian group $A$ by $C_2 = \langle x \rangle$, where $x$ acts on each element of $A$ by inversion, then we say $G$ is *of dihedral type* and denote $G$ by $\mathcal{D}_A$.

Finally, we borrow a definition from Edmonds [1]:

**Definition E.6.** If $G$ is a group of order $2^n m$, $m$ odd, and $j(G) = 2^{n-1}(m+1)$, then we call $G$ 2-*maximal*.

This terminology is justified by Edmonds' proof that a 2-maximal group does in fact have at least as many involutions as every other group of its order. For convenience, we will restate that result here, using our conventions:

**Theorem E.2.1** (Edmonds)**.** *If $|G| = 2^n m$, $m$ odd and $n$ at least 1, then $j(G) \leq 2^{n-1}(m+1)$.*

## E.3 Determining $j(G)$ and $J(G)$

It will be useful to record a few almost trivial facts, both to give the reader a feel for the topic and to streamline the proofs of some of the more interesting theorems. Determining $J(C_n)$, for example, is a trivial corollary of the Fundamental Theorem of Cyclic Groups:

**Lemma E.3.1.** *If $G = C_{2n}$, a cyclic group of even order, then $J(G) = \langle x \rangle$, where $x$ is the lone involution in $G$, and $j = 2$.*

The following are worth recording, but too trivial to merit numbering or proof:

1. A group element is contained in $J$ if and only if it is its own inverse.
2. If $|G|$ is odd, then $J(G)$ is trivial.
3. If $|G|$ is even, then $j(G) \geq 2$.[2]

The following forms the foundation for our analysis of actually realized values of $j$.

[2] By Cauchy's Theorem.

**Proposition E.3.2.** *If $|G|$ is even, then $j(G)$ is even.*

The easiest way to see this is to pair every element with its inverse; an element in $J$ will be paired with itself, and it is easy to see that the number of such elements must be even. For variety, we provide a short proof (that expresses the same ideas) using group actions and Burnside's Counting Lemma. It will be useful for the reader to recall the following definition:

If $G$ is acting on $S$, then, for $g \in G$, $\text{fix}(g) = \{ s \in S \mid g \cdot s = s \}$.

*Proof.* Let $G$ be a group of even order, and let $C_2 = \langle x \rangle$ act on $G$ by conjugation, wherein $x$ inverts each element of $G$. That is, $x \cdot g = xgx^{-1} = g^{-1}$ for every $g \in G$. Note that $\text{fix}(x) = \{ g \in G \mid g^{-1} = g \} = J(G)$. Letting $\mathcal{O}$ denote the number of orbits under this action, we have, by Burnside's Counting Lemma,

$$\mathcal{O} = \frac{1}{|C_2|} \sum_{z \in \langle x \rangle} |\,\text{fix}(z)| = \frac{1}{2}\big(|\,\text{fix}(1_{C_2})| + |\,\text{fix}(x)|\big) = \frac{|G|}{2} + \frac{j(G)}{2}.$$

$\mathcal{O}$ and $|G|/2$ are integers, so $j(G)/2$ is an integer. That is, $j(G)$ is even, as required. Q.E.D.

Many of our results concern groups more than half of whose elements are involutions. The perceptive reader might already have identified $D_{2n}$ as such a group; for the rest of you, here is that result with proof:

**Lemma E.3.3.** *If $G = D_{2n}$, then*

$$j(G) = \begin{cases} n+1 & \text{if } n \text{ is odd} \\ n+2 & \text{if } n \text{ is even.} \end{cases}$$

*Proof.* We give $D_{2n}$ its standard presentation:

$$\left\langle r, s \mid r^n = s^2 = 1, srs^{-1} = r^{-1} \right\rangle$$

First note that $(sr^i)(sr^i) = (sr^i s)r^i = r^{-i}r^i = 1$ for all $i$, $1 \leq i \leq n$. As $s \neq (r^i)^{-1}$, $sr^i$ is a nontrivial element of $J(D_{2n})$. So $J(G)$ contains these $n$ distinct elements and the identity, and $r^{n/2}$ if and only if $n$ is even. Q.E.D.

**Lemma E.3.4.** *$j(G \oplus H) = j(G)j(H)$.*

*Proof.* Let $g \in G$ and $h \in H$. Then, since $|(a,b)| = \text{lcm}(|a|, |b|)$,

$$\begin{aligned} (g,h) \in J(G \oplus H) &\iff |(g,h)| \in \{1,2\} \\ &\iff \text{lcm}(|g|,|h|) \in \{1,2\} \\ &\iff |g| \in \{1,2\} \text{ and } |h| \in \{1,2\} \\ &\iff g \in J(G) \text{ and } h \in J(H). \end{aligned}$$

Q.E.D.

**Proposition E.3.5.** *If $G = \mathcal{D}_A$ is of dihedral type, then every element of $G \setminus A$ is an involution, and $j(G) = |G|/2 + j(A)$.*

*Proof.* For $x_i \in C_2 = \langle x \rangle$ and $a_i \in A$, $(a_1, x_1)(a_2, x_2) = (a_1 a_2^{-1}, x_1 x_2)$.
So the elements of $G \smallsetminus A$ correspond to the elements $(a_i, x)$ in $A \rtimes \langle x \rangle$. Clearly each of these is nontrivial, and $(a_i, x)^2 = (a_i a_i^{-1}, x^2) = 1_G$, so each is an involution. Then we have

$$j(G) = j(G \smallsetminus A) + j(A) = |G \smallsetminus A| + j(A) = |G|/2 + j(A),$$

as required. Q.E.D.

Note that, in particular, when $G$ is the dihedral group of order $2|A|$, $A$ is cyclic, so

$$j(G) = \begin{cases} |G|/2 + 1 & \text{if } |A| \text{ is odd} \\ |G|/2 + 2 & \text{if } |A| \text{ is even,} \end{cases}$$

as in Lemma E.3.3.

**Proposition E.3.6.** *If $x \in J(G)$, then $\langle x \rangle$ is normal in $G$ if and only if $x$ is central.*

*Proof.* Every inner automorphism of $G$ must map $x$ to itself, so $gxg^{-1} = x$ for every $g$ in $G$. That is, $gx = xg$ for every $g \in G$. Q.E.D.

**Proposition E.3.7.** *$J(G) \cap Z(G)$ is a subgroup of $G$.*

*Proof.* Let $J$ denote $J(G)$ and $Z$ denote $Z(G)$.
Let $x$ and $y$ be elements of $J \cap Z$. Then $xy \in Z$ and $(xy)^2 = x^2 y^2 = 1_G$, as $x$ and $y$ are in $J$, so $xy \in J$. Furthermore, $(x^{-1})^2 = (x^2)^{-1} = 1_G$, so $x^{-1}$ is also in $J \cap Z$. Q.E.D.

It seems time to add the symmetric and alternating groups to the list of groups for which we have a formula for the number of involutions:

**Theorem E.3.8** (Formulas for $j(S_n)$ and $j(A_n)$)**.**

$$j(S_n) - 1 = \sum_{i=1}^{\lfloor n/2 \rfloor} \frac{n!}{2^i (n-2i)! i!} \text{ and}$$
$$j(A_n) - 1 = \sum_{i=1}^{\lfloor n/4 \rfloor} \frac{n!}{2^{2i} (n-4i)! (2i)!}.$$

*Proof.* A nontrivial element in $S_n$ is an involution if and only if it is the product of disjoint 2-cycles. For $k \geq 0$, let $N_k$ denote the number of distinct products of $k$ 2-cycles in $S_n$. It is not difficult to see that $N_k = 0$ for $k > n/2$; in $S_7$, for example, there are no products of 4, 5, 6, etc. disjoint 2-cycles.

It is now clear that

$$j(S_n) = \sum_{\substack{k \\ k \leq n/2}} N_k = \sum_{k=1}^{\lfloor n/2 \rfloor} N_k, \tag{E.1}$$

so we need only to determine $N_k$. Notice first that, modulo permutation of disjoint cycles[3], we must simply count the number of ways to choose 2 elements from $n$, iterated $k$ times without replacement. This is clearly

$\binom{n}{2}\binom{n-2}{2}\cdots\binom{n-k+2}{2}$, but we have counted each element $k!$ times, so we must divide by $k!$, yielding

$$\begin{aligned} N_k &= \frac{1}{k!}\prod_{j=0}^{k-1}\binom{n-2j}{2} \\ &= \frac{1}{k!}\left(\frac{n!}{2(n-2)!}\right)\left(\frac{(n-2)!}{2(n-4)!}\right)\cdots \\ &= \frac{1}{k!}\left(\frac{n!}{2^k(n-2k)!}\right). \end{aligned}$$

Substituting this into Equation (E.1) yields the formula for $j(S_n)$. The formula for $j(A_n)$ is obtained by only counting $N_k$ for $k$ even. Q.E.D.

Lemmas E.3.1, E.3.3, and E.3.4 and Theorem E.3.8 allow us to determine $j(G)$ for a large class of groups. Our focus for the remainder of the paper will be the reverse: given $j(G)$, what can we say about the structure of $G$?

## E.4 Toward the Main Result

The following special case of the main result is worth proving separately, as the proof reveals some structural properties of 2-groups with many involution, and seems to suggest how *not* to prove the main theorem!

**Theorem E.4.1.** *If $|G| = 2^n$, $n \geq 4$, then $j(G) \neq 2^n - 2$.*

*Proof.* Suppose not. Then $G$ has exactly two elements of order greater than 2; call them $x$ and $x^{-1}$. Clearly $|x| = |x^{-1}|$ must be a power of 2, but if $|x| \geq 8$, we have a cyclic subgroup that contains more than two elements of order greater than 2, a contradiction. So $|x| = |x^{-1}| = 4$. It is clear that $\langle x\rangle \trianglelefteq G$ (in fact, it is characteristic), so the size of the conjugacy class of $x$ is at most 2. Thus $[G : C(x)] \leq 2$, so, since we assumed $|G| \geq 16$, $C(x)$ contains an involution $y \neq x^2$. But then $\langle x\rangle \times \langle y\rangle = \langle x\rangle\langle y\rangle$ is a subgroup isomorphic to $C_2 \oplus C_4$, which contains four distinct elements of order 4, a contradiction. Q.E.D.

**Proposition E.4.2.** *If $G$ is abelian, then $j(G)$ is a power of 2.*

*Proof.* Without loss of generality, we can partition $G$ into a product of cyclic groups of even order and a product of cyclic groups of odd order. That is,

$$G = C_{a_1} \oplus \cdots \oplus C_{a_m} \oplus C_{a_{m+1}} \oplus \cdots \oplus C_{a_n}$$

where $a_i$ is even for $i \leq m$ and $a_i$ is odd for $m+1 \leq i \leq n$.
Then by Lemma E.3.4,

$$\begin{aligned} j(G) &= j(C_{a_1})j(C_{a_2})\cdots j(C_{a_m})j(C_{a_{m+1}})\cdots j(C_{a_n}) \\ &= \left(\prod_{i=1}^{m} j(C_{a_i})\right)\left(\prod_{k=m+1}^{n} j(C_{a_k})\right) \end{aligned}$$

which, by Lemma E.3.1, is $2^m$, as required. Q.E.D.

[3] Disjoint $n$-cycles commute, so we should count $\alpha_1\alpha_2$ and $\alpha_2\alpha_1$ as the same permutation.

**Lemma E.4.3.** $\frac{j(G)}{|H|} \leq j(G/H) \leq \frac{|G|}{|H|}$ *for any normal subgroup* $H$ *of* $G$.

*Proof.* The first inequality is a restatement of Proposition 4.8 of [1]; the second is trivial. Q.E.D.

For convenience, we restate Proposition 4.9 of [1] using our conventions:

**Lemma E.4.4.** *For a central subgroup* $H$, $j(G) \leq j(G/H)j(H)$.

We state the following as a theorem, though we will need it only for the proof of Theorem E.5.1:

**Theorem E.4.5.** *If* $j(G) > |G|/2$, *then* $Z(G)$ *is an (possibly trivial)*[4] *elementary abelian* 2*-group.*

*Proof.* Let $Z = Z(G)$ and let $S$ be a Sylow 2-subgroup of $G$. By Corollary 4.4 of Edmonds, $N_G(S) = S$; so, as $Z$ normalizes every subgroup, $Z \leqslant S$.
So $Z$ is a 2-group; let $|Z| = 2^a$. Since $Z$ is abelian, by Proposition E.4.2, $j(Z) = 2^b$ for some $0 \leq b \leq a$.
Now we have

$$\begin{aligned} \frac{|G|}{2} &< j(G) \\ &\leq j(G/Z)j(Z) \qquad \text{(by Lemma E.4.4)} \\ &= j(G/Z)2^b \\ &\leq \frac{|G|}{|Z|}2^b \qquad \text{(by Lemma E.4.3)} \end{aligned}$$

So $|Z| < 2^{b+1}$, but $Z$ is a 2-group of order at least $j(Z) = 2^b$, so $|Z| = j(Z)$ and $Z = J(Z)$. Q.E.D.

Note that the inequality in the hypothesis of our proposition must be strict:
Consider $G = C_4 \oplus {C_2}^{n-2}$. Then $j(G) = 2^{n-1} = |G|/2$, but $Z(G) = G$, which is not an elementary abelian 2-group.

## E.5 On a Standard Algebra Exercise: Main Result

We have now accumulated enough firepower to prove the main result:

**Theorem E.5.1.** *If* $G$ *is a finite group and* $j(G) > \frac{3}{4}|G|$, *then* $G$ *is an elementary abelian* 2*-group.*

*Proof.* By the main theorem of Edmonds [1], $G$ must be a 2-group: if $|G| = 2^n m$, $m \geq 3$, then

$$j(G) \leq 2^{n-1}(m+1) = 2^{n-2}(2m+2) < 2^{n-2}(3m) = \frac{3}{4}|G|.$$

Let $|G| = 2^n$, and suppose $n > 3$, as the case when $n \leq 3$ follows trivially from Proposition E.3.2. We proceed by induction on $n$. As $G$ is a 2-group, its center is nontrivial, hence contains an involution. Let $a$ be a central involution in $G$. By Lemma E.4.4, we have $j(G) \leq j(G/\langle a \rangle)j(\langle a \rangle)$; so

$$j(G/\langle a \rangle) \geq \frac{j(G)}{2} > \frac{3}{8}|G| = \frac{3}{4}|G/\langle a \rangle|,$$

[4] Consider the dihedral group of order $2n$, $n$ odd.

so by our inductive hypothesis, $G/\langle a\rangle$ is an elementary abelian 2-group of order $2^{n-1}$. We thus have an extension

$$C_2 \rightarrowtail G \twoheadrightarrow (C_2)^{n-1}$$

which we must show to be a direct product.

If $G$ is not an elementary abelian 2-group, there must be an element of order 4; suppose $x \in G$ is such an element. Note that, because $G/\langle a\rangle$ is of exponent 2, $x^2 = a$. Then, under the natural projection onto the quotient group $G/\langle a\rangle$, $\langle x\rangle$ is a cyclic subgroup of order 2. As that cyclic subgroup must be normal in the quotient, $\langle x\rangle \lhd G$. Consequently, conjugation by an element of $G$ must send $x$ to either $x$ or $x^{-1}$: those are the only two candidates inside $\langle x\rangle$. In particular, the length of the orbit of $x$ under conjugation is at most 2. In fact, since by Theorem E.4.5 there are no elements of order 4 in $Z(G)$, $[G : C(x)] \neq 1$. So

$$|\operatorname{cl}(x)| = [G : C(x)] = 2,$$

where $\operatorname{cl}(x)$ denotes the orbit under this action. Now note that

$$j(C(x)) = j(G) - j(G \setminus C(x)) > \frac{3}{4}|G| - \frac{1}{2}|G| = \frac{1}{4}|G|.$$

Combining, we have

$$j(C(x)) > \frac{1}{4}|G| = \frac{1}{2}|C(x)|,$$

so we can apply Theorem E.4.5 to $C(x)$. Therefore $Z(C(x))$ is an elementary abelian 2-group, but $x$, an element of order 4, is necessarily in the center of its own centralizer, a contradiction.

So $G$ contains no element of order 4 and is thus an elementary abelian 2-group, as required. Q.E.D.

The inequality in the hypothesis of our theorem must again be strict. $G = D_8 \oplus (C_2)^{n-3}$, in particular, is not elementary abelian, but $j(G) = 2^{n-3}(6) = \frac{3}{4}|G|$.

## E.6 GAP

GAP is a computer algebra program packaged with extensive group libraries. Our chief strategy was to use GAP to export in a convenient way correlations between the structural properties of a group and its involution count. The following function of $G$ whose purpose is to, using a counter variable, determine $j(G)$, will be called by nearly all of the included code.

```
gap> jinvol := function(g) local a,b;
b := 0;
for a in Elements(g) do
if Order(a)=2 then
b := b+1;
fi;
od;
return b+1;
end;
```

Although this is quite unsophisticated, GAP executes this function for fairly large and complicated groups with surprising efficiency.

A characteristic example, this code verifies the accuracy of the formulas given in Theorem E.3.8 for $n \leq 10$:

```
gap> sninvol := function(n); return
 Sum([1..Int(n/2)], x -> Factorial(n)/(2^x*Factorial(n-2*x)
 *Factorial(x)))+1; end;
function( n ) ... end
gap> for x in [1..10] do  Print(jinvol(SymmetricGroup(x))
 ," = ",sninvol(x),"\n"); od;
1 = 1
2 = 2
4 = 4
10 = 10
26 = 26
76 = 76
232 = 232
764 = 764
2620 = 2620
9496 = 9496

gap> aninvol := function(n); return
 Sum([1..Int(n/4)], x -> Factorial(n)/(2^(2*x)*Factorial(n-4*x)
 *Factorial(2*x)))+1; end;
function( n ) ... end
gap> for x in [1..10] do  Print(jinvol(AlternatingGroup(x))
 ," = ",aninvol(x),"\n"); od;
1 = 1
1 = 1
1 = 1
4 = 4
16 = 16
46 = 46
106 = 106
316 = 316
1324 = 1324
5356 = 5356
```

I also made use of the `orderfrequency` function as defined in Gallian and Rainbolt's GAP User Manual.

A clear theme of the paper is that not all possible even values of $j(G)$ for a given $|G|$ are actually realized by any finite group. The patterns of actually realized $j$s are, of course, interesting, and a table of such values (along with the GAP code that generated the table) is included in the Appendix.

For example, Tables E.3 and E.4 in the Appendix motivate the following conjecture, which suggests a full characterization of the "2-almost-maximal" groups:

**Conjecture E.1.** *Let $G$ be a finite group of order $2^n m$, with $m$ odd, and suppose $n \geq 3$. If $j(G) > 2^{n-2}(2m+1)$, then $G$ is 2-maximal.*
*If $j(G) = 2^{n-2}(2m+1)$, then $G$ is the direct product of ${C_2}^{n-3}$ and a group of dihedral type of order $8m$.*

Table E.2 in the Appendix motivates

**Conjecture E.2.** *Let $G$ be such that $j(G) > |G|/2$, and suppose $4 \big| |G|$. Then the center of $G$ is nontrivial.*

Finally, a cursory glance at Table E.1 would seem to justify the following:

**Conjecture E.3.** *If $j(G) > |G|/2$, then $G$ is isomorphic to a direct product of cyclic groups of order 2 and groups of dihedral type.*

However, a closer look at, for example, the first listed group of order 128 suggests otherwise. If we assume every nontrivial extension by a $C_2$ is such that the generator of $C_2$ acts by conjugation on the group (which GAP's `StructureDescription` command alone does not guarantee, of course), then it seems that no counterexample to E.3 appears until $|G| = 128$. Further research should investigate whether groups with $j > |G|/2$ can be fully characterized using a further generalization of the dihedral group.

# Appendix

It should be noted that the included code often requires allocating GAP more than the default memory; this can be accomplished using the `-o` or `-m` command line option.[5]

Table E.1: Every group $G$ with $j(G) \geq |G|/2$ and $|G| \leq 128$.

| $\|G\|$ | $j(G)$ | Isomorphism Type |
|---|---|---|
| 2 | 2 | $C_2$ |
| 4 | 2 | $C_4$ |
| | 4 | $C_2 \times C_2$ |
| 6 | 4 | $S_3$ |
| 8 | 4 | $C_4 \times C_2$ |
| | 6 | $D_8$ |
| | 8 | $C_2 \times C_2 \times C_2$ |
| 10 | 6 | $D_{10}$ |
| 12 | 8 | $D_{12}$ |
| 14 | 8 | $D_{14}$ |
| 16 | 8 | $(C_4 \times C_2) \rtimes C_2$ |
| | 8 | $C_4 \times C_2 \times C_2$ |
| | 8 | $(C_4 \times C_2) \rtimes C_2$ |
| | 10 | $D_{16}$ |
| | 12 | $C_2 \times D_8$ |
| | 16 | $C_2 \times C_2 \times C_2 \times C_2$ |
| 18 | 10 | $D_{18}$ |
| | 10 | $(C_3 \times C_3) \rtimes C_2$ |
| 20 | 12 | $D_{20}$ |
| 22 | 12 | $D_{22}$ |
| 24 | 14 | $D_{24}$ |
| | 16 | $C_2 \times C_2 \times S_3$ |
| 26 | 14 | $D_{26}$ |
| 28 | 16 | $D_{28}$ |
| 30 | 16 | $D_{30}$ |
| 32 | 16 | $C_2 \times ((C_4 \times C_2) \rtimes C_2)$ |
| | 16 | $(C_4 \times C_2 \times C_2) \rtimes C_2$ |
| | 16 | $(C_2 \times D_8) \rtimes C_2$ |
| | 16 | $C_4 \times C_2 \times C_2 \times C_2$ |
| | 16 | $C_2 \times ((C_4 \times C_2) \rtimes C_2)$ |
| | 18 | $D_{32}$ |
| | 20 | $(C_2 \times C_2 \times C_2 \times C_2) \rtimes C_2$ |
| | 20 | $(C_4 \times C_4) \rtimes C_2$ |
| | 20 | $C_2 \times D_{16}$ |
| | 20 | $(C_2 \times D_8) \rtimes C_2$ |
| | 24 | $C_2 \times C_2 \times D_8$ |

[5]For example, `-o 500m` allocates GAP 500 megabytes of memory.

Table E.1: (continued)

| $\|G\|$ | $j(G)$ | Isomorphism Type |
|---|---|---|
| | 32 | $C_2 \times C_2 \times C_2 \times C_2 \times C_2$ |
| 34 | 18 | $D_{34}$ |
| 36 | 20 | $D_{36}$ |
| | 20 | $C_2 \times ((C_3 \times C_3) \rtimes C_2)$ |
| 38 | 20 | $D_{38}$ |
| 40 | 22 | $D_{40}$ |
| | 24 | $C_2 \times C_2 \times D_{10}$ |
| 42 | 22 | $D_{42}$ |
| 44 | 24 | $D_{44}$ |
| 46 | 24 | $D_{46}$ |
| 48 | 24 | $D_8 \times S_3$ |
| | 26 | $D_{48}$ |
| | 28 | $C_2 \times D_{24}$ |
| | 32 | $C_2 \times C_2 \times C_2 \times S_3$ |
| 50 | 26 | $D_{50}$ |
| | 26 | $(C_5 \times C_5) \rtimes C_2$ |
| 52 | 28 | $D_{52}$ |
| 54 | 28 | $D_{54}$ |
| | 28 | $(C_9 \times C_3) \rtimes C_2$ |
| | 28 | $(C_3 \times C_3 \times C_3) \rtimes C_2$ |
| 56 | 30 | $D_{56}$ |
| | 32 | $C_2 \times C_2 \times D_{14}$ |
| 58 | 30 | $D_{58}$ |
| 60 | 32 | $D_{60}$ |
| 62 | 32 | $D_{62}$ |
| 64 | 32 | $C_2 \times ((C_4 \times C_2 \times C_2) \rtimes C_2)$ |
| | 32 | $(C_2 \times ((C_4 \times C_2) \rtimes C_2)) \rtimes C_2$ |
| | 32 | $(C_2 \times C_2 \times D_8) \rtimes C_2$ |
| | 32 | $(C_2 \times C_2 \times D_8) \rtimes C_2$ |
| | 32 | $C_4 \times C_2 \times C_2 \times C_2 \times C_2$ |
| | 32 | $C_2 \times C_2 \times ((C_4 \times C_2) \rtimes C_2)$ |
| | 32 | $(C_2 \times C_2 \times D_8) \rtimes C_2$ |
| | 32 | $(C_2 \times ((C_4 \times C_2) \rtimes C_2)) \rtimes C_2$ |
| | 32 | $(C_2 \times C_2 \times D_8) \rtimes C_2$ |
| | 32 | $(C_2 \times D_{16}) \rtimes C_2$ |
| | 32 | $C_2 \times ((C_2 \times D_8) \rtimes C_2)$ |
| | 32 | $C_2 \times C_2 \times ((C_4 \times C_2) \rtimes C_2)$ |
| | 34 | $D_{64}$ |
| | 36 | $C_2 \times D_{32}$ |
| | 36 | $(C_8 \times C_4) \rtimes C_2$ |
| | 36 | $D_8 \times D_8$ |
| | 40 | $C_2 \times ((C_4 \times C_4) \rtimes C_2)$ |

Table E.1: (continued)

| $\|G\|$ | $j(G)$ | Isomorphism Type |
|---|---|---|
| | 40 | $C_2 \times C_2 \times D_{16}$ |
| | 40 | $C_2 \times ((C_2 \times C_2 \times C_2 \times C_2) \rtimes C_2)$ |
| | 40 | $C_2 \times ((C_2 \times D_8) \rtimes C_2)$ |
| | 48 | $C_2 \times C_2 \times C_2 \times D_8$ |
| | 64 | $C_2 \times C_2 \times C_2 \times C_2 \times C_2 \times C_2$ |
| 66 | 34 | $D_{66}$ |
| 68 | 36 | $D_{68}$ |
| 70 | 36 | $D_{70}$ |
| 72 | 38 | $D_{72}$ |
| | 38 | $(C_{12} \times C_3) \rtimes C_2$ |
| | 40 | $C_2 \times C_2 \times D_{18}$ |
| | 40 | $C_2 \times C_2 \times ((C_3 \times C_3) \rtimes C_2)$ |
| 74 | 38 | $D_{74}$ |
| 76 | 40 | $D_{76}$ |
| 78 | 40 | $D_{78}$ |
| 80 | 42 | $D_{80}$ |
| | 44 | $C_2 \times D_{40}$ |
| | 48 | $C_2 \times C_2 \times C_2 \times D_{10}$ |
| 82 | 42 | $D_{82}$ |
| 84 | 44 | $D_{84}$ |
| 86 | 44 | $D_{86}$ |
| 88 | 46 | $D_{88}$ |
| | 48 | $C_2 \times C_2 \times D_{22}$ |
| 90 | 46 | $D_{90}$ |
| | 46 | $(C_{15} \times C_3) \rtimes C_2$ |
| 92 | 48 | $D_{92}$ |
| 94 | 48 | $D_{94}$ |
| 96 | 48 | $C_2 \times D_8 \times S_3$ |
| | 50 | $D_{96}$ |
| | 52 | $(C_{12} \times C_4) \rtimes C_2$ |
| | 52 | $C_2 \times D_{48}$ |
| | 56 | $C_2 \times C_2 \times D_{24}$ |
| | 64 | $C_2 \times C_2 \times C_2 \times C_2 \times S_3$ |
| 98 | 50 | $D_{98}$ |
| | 50 | $(C_7 \times C_7) \rtimes C_2$ |
| 100 | 52 | $D_{100}$ |
| | 52 | $C_2 \times ((C_5 \times C_5) \rtimes C_2)$ |
| 102 | 52 | $D_{102}$ |
| 104 | 54 | $D_{104}$ |
| | 56 | $C_2 \times C_2 \times D_{26}$ |
| 106 | 54 | $D_{106}$ |
| 108 | 56 | $D_{108}$ |

Table E.1: (continued)

| $\|G\|$ | $j(G)$ | Isomorphism Type |
|---|---|---|
| | 56 | $C_2 \times ((C_9 \times C_3) \rtimes C_2)$ |
| | 56 | $C_2 \times ((C_3 \times C_3 \times C_3) \rtimes C_2)$ |
| 110 | 56 | $D_{110}$ |
| 112 | 58 | $D_{112}$ |
| | 60 | $C_2 \times D_{56}$ |
| | 64 | $C_2 \times C_2 \times C_2 \times D_{14}$ |
| 114 | 58 | $D_{114}$ |
| 116 | 60 | $D_{116}$ |
| 118 | 60 | $D_{118}$ |
| 120 | 62 | $D_{120}$ |
| | 64 | $C_2 \times C_2 \times D_{30}$ |
| 122 | 62 | $D_{122}$ |
| 124 | 64 | $D_{124}$ |
| 126 | 64 | $D_{126}$ |
| | 64 | $(C_{21} \times C_3) \rtimes C_2$ |
| 128 | 64 | $C_2 \times ((C_2 \times ((C_4 \times C_2) \rtimes C_2)) \rtimes C_2)$ |
| | 64 | $C_2 \times ((C_2 \times C_2 \times D_8) \rtimes C_2)$ |
| | 64 | $(C_2 \times C_2 \times C_2 \times D_8) \rtimes C_2$ |
| | 64 | $(C_2 \times C_2 \times C_2 \times D_8) \rtimes C_2$ |
| | 64 | $(C_2 \times C_2 \times C_2 \times D_8) \rtimes C_2$ |
| | 64 | $C_2 \times C_2 \times C_2 \times ((C_4 \times C_2) \rtimes C_2)$ |
| | 64 | $C_2 \times ((C_2 \times C_2 \times D_8) \rtimes C_2)$ |
| | 64 | $C_2 \times C_2 \times ((C_4 \times C_2 \times C_2) \rtimes C_2)$ |
| | 64 | $C_2 \times ((C_2 \times D_{16}) \rtimes C_2)$ |
| | 64 | $C_2 \times ((C_2 \times C_2 \times D_8) \rtimes C_2)$ |
| | 64 | $C_2 \times ((C_2 \times C_2 \times D_8) \rtimes C_2)$ |
| | 64 | $C_2 \times C_2 \times ((C_2 \times D_8) \rtimes C_2)$ |
| | 64 | $(C_2 \times ((C_4 \times C_4) \rtimes C_2)) \rtimes C_2$ |
| | 64 | $C_2 \times C_2 \times C_2 \times ((C_4 \times C_2) \rtimes C_2)$ |
| | 64 | $C_4 \times C_2 \times C_2 \times C_2 \times C_2 \times C_2$ |
| | 64 | $C_2 \times ((C_2 \times ((C_4 \times C_2) \rtimes C_2)) \rtimes C_2)$ |
| | 66 | $D_{128}$ |
| | 68 | $(C_{16} \times C_4) \rtimes C_2$ |
| | 68 | $C_2 \times D_{64}$ |
| | 68 | $(C_8 \times C_8) \rtimes C_2$ |
| | 72 | $(C_2 \times C_2 \times C_2 \times C_2 \times C_2 \times C_2) \rtimes C_2$ |
| | 72 | $(C_2 \times ((C_2 \times D_8) \rtimes C_2)) \rtimes C_2$ |
| | 72 | $C_2 \times C_2 \times D_{32}$ |
| | 72 | $C_2 \times ((C_8 \times C_4) \rtimes C_2)$ |
| | 72 | $C_2 \times D_8 \times D_8$ |
| | 72 | $(C_4 \times C_4 \times C_4) \rtimes C_2$ |
| | 80 | $C_2 \times C_2 \times ((C_2 \times C_2 \times C_2 \times C_2) \rtimes C_2)$ |

Table E.1: (continued)

| $\|G\|$ | $j(G)$ | Isomorphism Type |
|---|---|---|
| | 80 | $C_2 \times C_2 \times ((C_4 \times C_4) \rtimes C_2)$ |
| | 80 | $C_2 \times C_2 \times C_2 \times D_{16}$ |
| | 80 | $C_2 \times C_2 \times ((C_2 \times D_8) \rtimes C_2)$ |
| | 96 | $C_2 \times C_2 \times C_2 \times C_2 \times D_8$ |
| | 128 | $C_2 \times C_2 \times C_2 \times C_2 \times C_2 \times C_2 \times C_2$ |

Table E.1 was created using the following GAP code:

```
PrintTo("/GAP_TeX/Table of js",
"\\begin{longtable}{|c|c|p{2.5in}|}\n",
"\\caption{Every group $G$ with $j(G) \\geq |G|/2$ and
 $|G| \\leq 128$.\\label{realizedjs}}\\\\\n",
"\\hline\n",
"$|G|$ & $j(G)$ & Isomorphism Type\\\\\\hline\\endfirsthead\n",
"\\caption[]{(continued)}\\\\\n",
"\\hline\n",
"$|G|$ & $j(G)$ & Isomorphism Type\\\\\\hline\\endhead\n",
"\\hline\\endfoot\n");
for x in List([1..10],y->2*y) do
 dummy := [];
 for g in AllSmallGroups(x) do
  if jinvol(g) >= x/2 then
   Append(dummy,g);
  fi;
  AppendTo("/GAP_TeX/morethanhalf",
   x);
  if Length(dummy) > 0 then
   for y in dummy do
    AppendTo("/GAP_TeX/morethanhalf",
    " & ",jinvol(y)," & ",StructureDescription(y),"\\\\\n");
   od;
  else
   AppendTo("/GAP_TeX/morethanhalf",
   " & \\\\\n");
  fi;
 od;
 AppendTo("/GAP_TeX/morethanhalf",
 "\\hline\n")
od;
AppendTo("/GAP_TeX/morethanhalf",
 "\\end{longtable}");
```

Obviously this code asks GAP to create the file

"/GAP_TeX/morethanhalf.txt", and the following `emacs` commands will format the text file to produce the LaTeX code for Table E.1:

```
M-x replace-regexp RET \(C\|D\|S\) RET \1_ RET
M-x replace-regexp RET _\([0-9]*\) RET _{\1} RET
M-x replace-regexp RET x RET \\times RET
M-x replace-regexp RET : RET \\rtimes RET
```

Be sure to `M-<` between each line to return the mark to the beginning of the file.

Table E.2: The center of $G$, for $|G|$ a multiple of 4 and $j(G) > |G|/2$.

| $\|G\|$ | $j(G)$ | Center |
|---|---|---|
| 4 | 4 | $C_2 \times C_2$ |
| 8 | 6 | $C_2$ |
| | 8 | $C_2 \times C_2 \times C_2$ |
| 12 | 8 | $C_2$ |
| 16 | 10 | $C_2$ |
| | 12 | $C_2 \times C_2$ |
| | 16 | $C_2 \times C_2 \times C_2 \times C_2$ |
| 20 | 12 | $C_2$ |
| 24 | 14 | $C_2$ |
| | 16 | $C_2 \times C_2$ |
| 28 | 16 | $C_2$ |
| 32 | 18 | $C_2$ |
| | 20 | $C_2 \times C_2$ |
| | 20 | $C_2 \times C_2$ |
| | 20 | $C_2 \times C_2$ |
| | 20 | $C_2$ |
| | 24 | $C_2 \times C_2 \times C_2$ |
| | 32 | $C_2 \times C_2 \times C_2 \times C_2 \times C_2$ |
| 36 | 20 | $C_2$ |
| | 20 | $C_2$ |
| 40 | 22 | $C_2$ |
| | 24 | $C_2 \times C_2$ |
| 44 | 24 | $C_2$ |
| 48 | 26 | $C_2$ |
| | 28 | $C_2 \times C_2$ |
| | 32 | $C_2 \times C_2 \times C_2$ |
| 52 | 28 | $C_2$ |
| 56 | 30 | $C_2$ |
| | 32 | $C_2 \times C_2$ |
| 60 | 32 | $C_2$ |
| 64 | 34 | $C_2$ |
| | 36 | $C_2 \times C_2$ |

Table E.2: (continued)

| $\|G\|$ | $j(G)$ | Center |
|---|---|---|
| | 36 | $C_2 \times C_2$ |
| | 36 | $C_2 \times C_2$ |
| | 40 | $C_2 \times C_2 \times C_2$ |
| | 40 | $C_2 \times C_2 \times C_2$ |
| | 40 | $C_2 \times C_2 \times C_2$ |
| | 40 | $C_2 \times C_2$ |
| | 48 | $C_2 \times C_2 \times C_2 \times C_2$ |
| | 64 | $C_2 \times C_2 \times C_2 \times C_2 \times C_2 \times C_2$ |
| 68 | 36 | $C_2$ |
| 72 | 38 | $C_2$ |
| | 38 | $C_2$ |
| | 40 | $C_2 \times C_2$ |
| | 40 | $C_2 \times C_2$ |
| 76 | 40 | $C_2$ |
| 80 | 42 | $C_2$ |
| | 44 | $C_2 \times C_2$ |
| | 48 | $C_2 \times C_2 \times C_2$ |
| 84 | 44 | $C_2$ |
| 88 | 46 | $C_2$ |
| | 48 | $C_2 \times C_2$ |
| 92 | 48 | $C_2$ |
| 96 | 50 | $C_2$ |
| | 52 | $C_2 \times C_2$ |
| | 52 | $C_2 \times C_2$ |
| | 56 | $C_2 \times C_2 \times C_2$ |
| | 64 | $C_2 \times C_2 \times C_2 \times C_2$ |
| 100 | 52 | $C_2$ |
| | 52 | $C_2$ |
| 104 | 54 | $C_2$ |
| | 56 | $C_2 \times C_2$ |
| 108 | 56 | $C_2$ |
| | 56 | $C_2$ |
| | 56 | $C_2$ |
| 112 | 58 | $C_2$ |
| | 60 | $C_2 \times C_2$ |
| | 64 | $C_2 \times C_2 \times C_2$ |
| 116 | 60 | $C_2$ |
| 120 | 62 | $C_2$ |
| | 64 | $C_2 \times C_2$ |
| 124 | 64 | $C_2$ |
| 128 | 66 | $C_2$ |
| | 68 | $C_2 \times C_2$ |

Table E.2: (continued)

| $\|G\|$ | $j(G)$ | Center |
|---|---|---|
| | 68 | $C_2 \times C_2$ |
| | 68 | $C_2 \times C_2$ |
| | 72 | $C_2 \times C_2 \times C_2$ |
| | 72 | $C_2 \times C_2 \times C_2$ |
| | 72 | $C_2 \times C_2 \times C_2$ |
| | 72 | $C_2 \times C_2 \times C_2$ |
| | 72 | $C_2 \times C_2 \times C_2$ |
| | 72 | $C_2$ |
| | 80 | $C_2 \times C_2 \times C_2 \times C_2$ |
| | 80 | $C_2 \times C_2 \times C_2 \times C_2$ |
| | 80 | $C_2 \times C_2 \times C_2 \times C_2$ |
| | 80 | $C_2 \times C_2 \times C_2$ |
| | 96 | $C_2 \times C_2 \times C_2 \times C_2 \times C_2$ |
| | 128 | $C_2 \times C_2 \times C_2 \times C_2 \times C_2 \times C_2 \times C_2$ |

The GAP code used to create Table E.2 is nearly identical to that used for Table E.1, so we will omit it here. The `emacs` commands will also render the resulting text file suitable for a LaTeX `longtable`.

Table E.3: Every realized $j(G)$ for every even $|G| \leq 256$.

| $\|G\|$ | Realized $j(G)$s |
|---|---|
| 2 | 2 |
| 4 | 2, 4 |
| 6 | 2, 4 |
| 8 | 2, 4, 6, 8 |
| 10 | 2, 6 |
| 12 | 2, 4, 8 |
| 14 | 2, 8 |
| 16 | 2, 4, 6, 8, 10, 12, 16 |
| 18 | 2, 4, 10 |
| 20 | 2, 4, 6, 12 |
| 22 | 2, 12 |
| 24 | 2, 4, 6, 8, 10, 14, 16 |
| 26 | 2, 14 |
| 28 | 2, 4, 16 |
| 30 | 2, 4, 6, 16 |
| 32 | 2, 4, 8, 10, 12, 16, 18, 20, 24, 32 |
| 34 | 2, 18 |
| 36 | 2, 4, 8, 10, 16, 20 |
| 38 | 2, 20 |
| 40 | 2, 4, 6, 8, 12, 14, 22, 24 |
| 42 | 2, 4, 8, 22 |

Table E.3: (continued)

| $\|G\|$ | Realized $j(G)$s |
|---|---|
| 44 | 2, 4, 24 |
| 46 | 2, 24 |
| 48 | 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 24, 26, 28, 32 |
| 50 | 2, 6, 26 |
| 52 | 2, 4, 14, 28 |
| 54 | 2, 4, 10, 28 |
| 56 | 2, 4, 6, 8, 16, 18, 30, 32 |
| 58 | 2, 30 |
| 60 | 2, 4, 6, 8, 12, 16, 24, 32 |
| 62 | 2, 32 |
| 64 | 2, 4, 8, 12, 16, 18, 20, 24, 28, 32, 34, 36, 40, 48, 64 |
| 66 | 2, 4, 12, 34 |
| 68 | 2, 4, 18, 36 |
| 70 | 2, 6, 8, 36 |
| 72 | 2, 4, 6, 8, 10, 14, 16, 20, 22, 26, 32, 38, 40 |
| 74 | 2, 38 |
| 76 | 2, 4, 40 |
| 78 | 2, 4, 14, 40 |
| 80 | 2, 4, 6, 8, 10, 12, 16, 22, 24, 26, 28, 32, 36, 42, 44, 48 |
| 82 | 2, 42 |
| 84 | 2, 4, 8, 16, 32, 44 |
| 86 | 2, 44 |
| 88 | 2, 4, 6, 8, 24, 26, 46, 48 |
| 90 | 2, 4, 6, 10, 16, 46 |
| 92 | 2, 4, 48 |
| 94 | 2, 48 |
| 96 | 2, 4, 8, 10, 12, 16, 18, 20, 24, 26, 28, 32, 34, 36, 40, 44, 48, 50, 52, 56, 64 |
| 98 | 2, 8, 50 |
| 100 | 2, 4, 6, 12, 26, 36, 52 |
| 102 | 2, 4, 18, 52 |
| 104 | 2, 4, 6, 8, 28, 30, 54, 56 |
| 106 | 2, 54 |
| 108 | 2, 4, 8, 10, 16, 20, 28, 40, 56 |
| 110 | 2, 6, 12, 56 |
| 112 | 2, 4, 6, 8, 10, 12, 16, 20, 30, 32, 34, 36, 44, 48, 58, 60, 64 |
| 114 | 2, 4, 20, 58 |
| 116 | 2, 4, 30, 60 |
| 118 | 2, 60 |
| 120 | 2, 4, 6, 8, 10, 12, 14, 16, 18, 22, 24, 26, 32, 34, 38, 42, 48, 62, 64 |
| 122 | 2, 62 |

Table E.3: (continued)

| $\|G\|$ | Realized $j(G)$s |
|---|---|
| 124 | 2, 4, 64 |
| 126 | 2, 4, 8, 10, 22, 64 |
| 128 | 2, 4, 8, 12, 16, 20, 24, 28, 32, 34, 36, 40, 44, 48, 52, 56, 60, 64, 66, 68, 72, 80, 96, 128 |
| 130 | 2, 6, 14, 66 |
| 132 | 2, 4, 8, 24, 48, 68 |
| 134 | 2, 68 |
| 136 | 2, 4, 6, 8, 18, 36, 38, 70, 72 |
| 138 | 2, 4, 24, 70 |
| 140 | 2, 4, 6, 12, 16, 48, 72 |
| 142 | 2, 72 |
| 144 | 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 32, 38, 40, 42, 44, 50, 52, 56, 60, 64, 74, 76, 80 |
| 146 | 2, 74 |
| 148 | 2, 4, 38, 76 |
| 150 | 2, 4, 6, 16, 26, 76 |
| 152 | 2, 4, 6, 8, 40, 42, 78, 80 |
| 154 | 2, 8, 12, 78 |
| 156 | 2, 4, 8, 14, 28, 56, 80 |
| 158 | 2, 80 |
| 160 | 2, 4, 8, 10, 12, 16, 18, 20, 24, 28, 32, 36, 40, 42, 44, 48, 50, 52, 56, 60, 64, 68, 72, 82, 84, 88, 96 |
| 162 | 2, 4, 10, 28, 82 |
| 164 | 2, 4, 42, 84 |
| 166 | 2, 84 |
| 168 | 2, 4, 6, 8, 10, 14, 16, 18, 22, 30, 32, 44, 46, 50, 58, 64, 86, 88 |
| 170 | 2, 6, 18, 86 |
| 172 | 2, 4, 88 |
| 174 | 2, 4, 30, 88 |
| 176 | 2, 4, 6, 8, 10, 12, 16, 24, 28, 46, 48, 50, 52, 68, 72, 90, 92, 96 |
| 178 | 2, 90 |
| 180 | 2, 4, 6, 8, 10, 12, 16, 20, 24, 32, 40, 46, 60, 64, 92 |
| 182 | 2, 8, 14, 92 |
| 184 | 2, 4, 6, 8, 48, 50, 94, 96 |
| 186 | 2, 4, 32, 94 |
| 188 | 2, 4, 96 |
| 190 | 2, 6, 20, 96 |
| 192 | 2, 4, 8, 12, 16, 18, 20, 24, 28, 32, 34, 36, 40, 44, 48, 50, 52, 56, 60, 64, 66, 68, 72, 76, 80, 84, 88, 96, 98, 100, 104, 112, 128 |
| 194 | 2, 98 |
| 196 | 2, 4, 16, 50, 64, 100 |

Table E.3: (continued)

| $\|G\|$ | Realized $j(G)$s |
|---|---|
| 198 | 2, 4, 10, 12, 34, 100 |
| 200 | 2, 4, 6, 8, 12, 14, 22, 24, 26, 36, 46, 52, 54, 62, 72, 102, 104 |
| 202 | 2, 102 |
| 204 | 2, 4, 8, 18, 36, 72, 104 |
| 206 | 2, 104 |
| 208 | 2, 4, 6, 8, 10, 12, 16, 28, 32, 54, 56, 58, 60, 80, 84, 106, 108, 112 |
| 210 | 2, 4, 6, 8, 16, 22, 36, 106 |
| 212 | 2, 4, 54, 108 |
| 214 | 2, 108 |
| 216 | 2, 4, 6, 8, 10, 14, 16, 20, 22, 26, 32, 34, 38, 40, 46, 56, 58, 62, 64, 74, 80, 110, 112 |
| 218 | 2, 110 |
| 220 | 2, 4, 6, 12, 24, 72, 112 |
| 222 | 2, 4, 38, 112 |
| 224 | 2, 4, 8, 10, 12, 16, 18, 20, 24, 32, 36, 40, 44, 48, 52, 58, 60, 64, 66, 68, 72, 76, 80, 88, 92, 96, 114, 116, 120, 128 |
| 226 | 2, 114 |
| 228 | 2, 4, 8, 40, 80, 116 |
| 230 | 2, 6, 24, 116 |
| 232 | 2, 4, 6, 8, 60, 62, 118, 120 |
| 234 | 2, 4, 10, 14, 40, 118 |
| 236 | 2, 4, 120 |
| 238 | 2, 8, 18, 120 |
| 240 | 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 32, 34, 36, 40, 42, 44, 48, 52, 56, 60, 62, 64, 66, 68, 72, 74, 76, 80, 82, 84, 88, 92, 96, 122, 124, 128 |
| 242 | 2, 12, 122 |
| 244 | 2, 4, 62, 124 |
| 246 | 2, 4, 42, 124 |
| 248 | 2, 4, 6, 8, 64, 66, 126, 128 |
| 250 | 2, 6, 26, 126 |
| 252 | 2, 4, 8, 10, 16, 20, 32, 44, 52, 80, 88, 128 |
| 254 | 2, 128 |
| 256 | 2, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, 48, 52, 56, 60, 64, 66, 68, 72, 76, 80, 84, 88, 92, 96, 100, 104, 108, 112, 120, 128, 130, 132, 136, 144, 160, 192, 256 |

Table E.3 was created using the following GAP code:

```
jfunc := function(a,b) local jofg,lj,og,i,x,q,sq;
jofg := [];
og := 2*Int(a/2);
```

```
repeat
 lj := [];
 for x in AllSmallGroups(og) do
  if not (jinvol(x) in lj) then Append(lj,[jinvol(x)]); fi;
 od;
 Sort(lj);
 Add(jofg,lj);
 og := og + 2;
until og > b;
for i in [1..Length(jofg)] do
 Print(2*i+2*Int(a/2)-2,":    ",jofg[i],"\n");
od;
PrintTo("/GAP_TeX/Table of js",
"\\begin{longtable}{|c|p{4in}}\n",
"\\caption{Every realized $j(G)$ for every even $|G| \leq 256$.
 \\label{realizedjs}}\\\\\n",
"\\hline\n",
"$|G|$ & Realized $j(G)$s\\\\\\hline\\endfirsthead\n",
"\\caption[]{(continued)}\\\\\n",
"\\hline\n",
"$|G|$ & Realized $j(G)$s\\\\\\hline\\endhead\n",
"\\hline\\endfoot\n");
for q in [1..Length(jofg)] do
 AppendTo("/GAP_TeX/Table of js","\\hline\n",2*q+2*Int(a/2)-2,"  &  ");
 for sq in [1..Length(jofg[q])-1] do
  AppendTo("/GAP_TeX/Table of js",jofg[q][sq],", ");
 od;
 AppendTo("/GAP_TeX/Table of js",jofg[q][Length(jofg[q])],"\\\\\n");
od;
AppendTo("/GAP_TeX/Table of js",
"\\hline\n",
"\\end{longtable}");
end;
jfunc(2,256);
```

Merely adding an argument to `jfunc` and changing the 2 in line 11 to that local variable would turn this function into a 'for loop' to generate tables like Table E.3.

Table E.4: The conjectured "2-almost-maximal" groups of order less than 256.

| $\|G\|$ | $j(G)^6$ | Isomorphism Type |
|---|---|---|
| 8 | 6 | $D_8$ |
| 16 | 12 | $C_2 \times D_8$ |
| 24 | 14 | $D_{24}$ |
| 32 | 24 | $C_2 \times C_2 \times D_8$ |

Table E.4: (continued)

| $\|G\|$ | $j(G)$ | Isomorphism Type |
|---|---|---|
| 40 | 22 | $D_{40}$ |
| 48 | 28 | $C_2 \times D_{24}$ |
| 56 | 30 | $D_{56}$ |
| 64 | 48 | $C_2 \times C_2 \times C_2 \times D_8$ |
| 72 | 38 | $D_{72}$ |
| 72 | 38 | $(C_{12} \times C_3) \rtimes C_2$ |
| 80 | 44 | $C_2 \times D_{40}$ |
| 88 | 46 | $D_{88}$ |
| 96 | 56 | $C_2 \times C_2 \times D_{24}$ |
| 104 | 54 | $D_{104}$ |
| 112 | 60 | $C_2 \times D_{56}$ |
| 120 | 62 | $D_{120}$ |
| 128 | 96 | $C_2 \times C_2 \times C_2 \times C_2 \times D_8$ |
| 136 | 70 | $D_{136}$ |
| 144 | 76 | $C_2 \times D_{72}$ |
| 144 | 76 | $C_2 \times ((C_{12} \times C_3) \rtimes C_2)$ |
| 152 | 78 | $D_{152}$ |
| 160 | 88 | $C_2 \times C_2 \times D_{40}$ |
| 168 | 86 | $D_{168}$ |
| 176 | 92 | $C_2 \times D_{88}$ |
| 184 | 94 | $D_{184}$ |
| 192 | 112 | $C_2 \times C_2 \times C_2 \times D_{24}$ |
| 200 | 102 | $D_{200}$ |
| 200 | 102 | $(C_{20} \times C_5) \rtimes C_2$ |
| 208 | 108 | $C_2 \times D_{104}$ |
| 216 | 110 | $D_{216}$ |
| 216 | 110 | $(C_{36} \times C_3) \rtimes C_2$ |
| 216 | 110 | $(C_{12} \times C_3 \times C_3) \rtimes C_2$ |
| 224 | 120 | $C_2 \times C_2 \times D_{56}$ |
| 232 | 118 | $D_{232}$ |
| 240 | 124 | $C_2 \times D_{120}$ |
| 248 | 126 | $D_{248}$ |

Constructing Table E.4 required use of the following two GAP functions. The `2nm` function factors the a number in our standard way: $a = 2^n m$, $m$ odd. Calling this function to select the groups of conjectured "2-almost-maximal" order, `almostmax` creates (as above) a text file containing code for Table E.4. The `emacs` commands described above will again format this table for LaTeX.

```
2nm := function(a) local dummy,x;
dummy := [];
```

[6] Note that if $|G| = 2^n m$, $m$ odd, then $j(G) = 2^{n-2}(2m+1)$.

```
for x in FactorsInt(a) do
 if x <> 2 then
  Append(dummy,[x]);
 fi;
od;
return [Length(FactorsInt(a))-Length(dummy),
 Product(dummy)];
end;

almostmax := function(b)
 local c,m,n,y;
 c := 0;
 PrintTo("/Users/zachnorwood/GAP_TeX/almostmax",
  "\\begin{longtable}{|c|c|p{2.5in}|}\n",
   "\\caption{The conjectured $2$-almost-maximal\
    groups of order less than $256$.\n",
   "\\label{almostmax}}\\\\\n",
   "\\hline\n",
   "$|G|$ & $j(G)$\\footnotemark & Isomorphism Type\
    \\\\\\hline\\endfirsthead\n",
   "\\caption[]{(continued)}\\\\\n",
   "\\hline\n",
   "$|G|$ & $j(G)$ & Isomorphism Type\\\\\n",
   "\\hline\\endhead\n",
   "\\hline\\endfoot\n");
  repeat
   c := c+8;
   n := 2nm(c)[1];
   m := 2nm(c)[2];
   for y in AllSmallGroups(c) do
    if jinvol(y) = 2^(n-2)*(2*m+1) then
     AppendTo("/Users/zachnorwood/GAP_TeX/almostmax",
      c," & ",jinvol(y)," & $",StructureDescription(y),
      "$\\\\\n","\\hline\n");
    fi;
   od;
  until c >= b;
  AppendTo("/Users/zachnorwood/GAP_TeX/almostmax",
   "\\end{longtable}\n",
   "\\footnotetext{Note that if $|G| = 2^{n}m$,\
    $m$ odd, then $j(G) =  2^{n-2}(2m+1)$.}");
 end;
```

The command `almostmax(248)` outputs the text file used to create Table E.4.

## Acknowledgements

For invaluable guidance, insight, and patience I would like to thank Allan Edmonds, without whom this project would not have been realized; I don't know how I could have had more fun with an REU project. Thanks also to Kevin Pilgrim for committing time, enthusiasm, and a thought-provoking attitude to the program. And for many a tasty PB&J and countless other behind-the-scenes services, Mandie McCarty deserves our thanks. Also deserving of recognition are Indiana University and the NSF for housing and funding, respectively. I can't take credit for this beautifully typeset paper and all the included GAP data without recognizing those responsible for the LaTeX documentation and the writers of the GAP Manual; both groups unknowingly corrected many errors and simplified many tasks. Finally, I would like to express my appreciation for my fellow REUers, whose quirks, enthusiasm, and as yet unexplained obsession with ethnic food completed this experience.

## Bibliography

1. Allan L. Edmonds, *The Partition Problem for Equifacetal Simplices.* Contributions to Algebra and Geometry, **50** (2009), 195-213.
2. David Dummit, Richard Foote, *Abstract Algebra.* John Wiley and Sons, 2004.
3. Michael Aschbacher. *Finite Group Theory.* Cambridge University Press, Cambridge, 1986.
4. W.R. Scott, *Group Theory.* Dover Publications, New York, 1987.
5. The GAP Group, *GAP—Groups, Algorithms, and Programming*, Version 4.4.12; 2008, `www.gap-system.org`.
6. J. G. Rainbolt and J. A. Gallian, *Abstract Algebra with GAP*; 2010, `http://math.slu.edu/~rainbolt/manual2.html`.

# On Quadratic Mappings With and Attracting Cycle

Jacek Skryzalin
Indiana University

Indiana University REU Summer 2009
Advisor: Eric Bedford

# F.1 Introduction

Let $f_c\colon \mathbb{C} \to \mathbb{C}$ be defined by $f_c(z) = z^2 + c$. Let $f_c^n$ denote $f_c$ composed with itself $n$ times, and let $f_c^{-n}(z_0)$ denote the $n^{\text{th}}$ inverse images of $z_0$ under $f_c$. The *forward orbit* of a point $z_0 \in \mathbb{C}$ is defined to be the set of points

$$O^+(z_0) = \{f_c^n(z_0)\colon n \geq 0\}$$

and the *backward orbit* of $z_0$ is the set

$$O^-(z_0) = \{f_c^n(z_0)\colon n \leq -1\}$$

Moreover, the *total orbit* is the set of points

$$O(z_0) = O^-(z_0) \cup O^+(z_0)$$

When considering the dynamics of the map $f_c$, two types of points will be of interest: fixed points and periodic points.

**Definition F.1.** A point $z_0$ is a fixed point iff $f_c(z_0) = z_0$. Moreover, for a fixed point $z_0$, we have $f_c(z_0) = {z_0}^2 + c = z_0$, so an application of the quadratic formula shows that $z_0 = \frac{1 \pm \sqrt{1-4c}}{2}$.

**Definition F.2.** A point $z_0$ is a periodic point of period $n$ iff $(\exists n)$ such that $f_c^n(z_0) = z_0$.

Thus, a fixed point can be thought of as a period point of period 1. Furthermore, we can classify fixed points and periodic points according to the following:

**Definition F.3.** Let $z_0$ be a periodic point of period $n$. Then $z_0$ is

$$\left\{ \begin{array}{ll} \text{superattracting} & \text{if } |(f_c^n)'(z_0)| = 0 \\ \text{attracting} & \text{if } 0 < |(f_c^n)'(z_0)| < 1 \\ \text{neutral} & \text{if } |(f_c^n)'(z_0)| = 1 \\ \text{repelling} & \text{if } |(f_c^n)'(z_0)| > 1 \end{array} \right\}$$

Throughout the paper, we will use the term *attracting* to refer to both attracting and superattracting periodic points. The reasoning for the terminology in definition F.3 is the following: assume $z_0$ is an attracting periodic point of period $n$. Near $z_0$, the function $f_c^n(z)$ behaves linearly. Furthermore,

$$|(f_c^n)'(z_0)| = \lim_{h \to 0} \left| \frac{f_c^n(z_0 + h) - f_c^n(z_0)}{h} \right| = |\lambda| < 1$$

Thus, let $z_1$ be sufficiently close to $z_0$, so that:

$$|f_c^n(z_1) - z_0| = |f_c^n(z_1) - f_c^n(z_0)| \approx |\lambda z_1 - \lambda z_0| = |\lambda(z_1 - z_0)| < |z_1 - z_0|$$

We can see that if $z_0$ is an attracting period point of period $n$, and $z$ is sufficiently close to $z_0$, then $\lim_{m \to \infty} f_c^{mn}(z) = z_0$. A similar argument shows that if $z_0$ is a repelling periodic point of period $n$ and $|z - z_0|$ is sufficiently small, then $|f_c^n(z) - z_0| > |z - z_0|$.

Although the behavior of periodic points is complicated, we can describe the existence of fixed points in some detail by the following:

**Theorem F.1.1.** *$f_c(z) = z^2 + c$ has at most one attracting fixed point, and there exists an attracting fixed point iff $c$ is in the interior of the region given by $g(\theta) = -\left(\frac{e^{i\theta}}{2} + \left(\frac{e^{i\theta}}{2}\right)^2\right)$.*

*Proof.* Let the two fixed points of $f_c(z)$ be $z_0 = \frac{1-\sqrt{1-4c}}{2}$ and $z_1 = \frac{1+\sqrt{1-4c}}{2}$. Consider the fixed point $z_1 = \frac{1+\sqrt{1-4c}}{2}$. Since $(f_c)'(z) = 2z$, we have

$$|(f_c)'(z_1)| = |1 + \sqrt{1-4c}| \leq |1| + |\sqrt{1-4c}| = 1$$

Thus shows that $z_1$ cannot be an attracting fixed point.

However, $z_0$ is an attracting fixed point iff $|(f_c)'(z_0)| < 1$. This will happen when $|(f_c)'(z_0)| = |1 - \sqrt{1-4c}| < 1$. That is, $\sqrt{1-4c}$ must be inside the circle with radius 1 centered at $1 + 0i$. Parameterizing this circle as $\gamma(\theta) = (1 + \cos(\theta)) + i(\sin(\theta))$, squaring $\gamma$, and applying trigonometric identities, we see that $1 - 4c$ must lie inside the region enclosed by $\gamma^2(\theta) = (2\cos(\theta) + \cos(2\theta) + 1) + i\,(2\sin(\theta) + \sin(2\theta))$. Further computation shows that $c$ must lie on the interior of the region enclosed by

$$\widetilde{\gamma}(\theta) = \left(-\frac{1}{2}\cos(\theta) - \frac{1}{4}\cos(2\theta)\right) + i\left(-\frac{1}{2}\sin(\theta) - \frac{1}{4}\sin(2\theta)\right) = -\left(\frac{e^{i\theta}}{2} + \left(\frac{e^{i\theta}}{2}\right)^2\right)$$

Q.E.D.

Furthermore, if $c$ can be written in the form $c = -\left(\frac{e^{i\theta}}{2} + \left(\frac{e^{i\theta}}{2}\right)^2\right)$, then $f_c$ contains a neutral fixed point, since $\left|(f_c)'\left(\frac{1-\sqrt{1-4c}}{2}\right)\right| = \left|1 - \sqrt{1 + 2e^{i\theta} + e^{i(2\theta)}}\right| = \left|e^{i\theta}\right| = 1$.

Finally, we define the natural numbers to be the set $\mathbb{N} = \{0, 1, 2, \dots\}$, equivalent to the nonnegative integers. Additionally, we define the *extended complex plane* to be the set $\widehat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$.

## F.2 The Julia and Fatou Sets

We may also consider how the iterates of $f_c$ behave in a neighborhood of a point $z_0$. This leads to the definition of equicontinuity:

**Definition F.4.** A family $\mathcal{F}$ of functions in the complex plane is equicontinuous at $z_0$ iff

$$(\forall \varepsilon > 0)(\exists \delta > 0)(\forall z \in \mathbb{C})(\forall f \in \mathcal{F})(|z - z_0| < \delta \longrightarrow |f(z) - f(z_0)| < \varepsilon)$$

Furthermore, $\mathcal{F}$ is equicontinuous on a set $X \subset \mathbb{C}$ iff it is equicontinuous for all $z \in X$.

In this paper, the family $\mathcal{F}$ will refer to the iterates of $f_c$, and the family will be denoted $\{f_c^n\}_{n=0}^{\infty}$.

**Definition F.5.** The Fatou set of $f_c$ is the maximal open subset of $\widehat{\mathbb{C}}$ on which $\{f_c^n\}$ is equicontinuous. Furthermore, the Julia set of $f_c$, denoted $J(f_c)$ is the complement of the Fatou set.

To make better sense of this definition, we introduce local uniform convergence and normality:

**Definition F.6.** A sequence $\{g_n\}$ of functions converges locally uniformly to $g$ on an open set $U$ iff $(\forall K \subset U)$, where $K$ is compact, $\{g_n\}$ converges uniformly to $g$ on $K$.

**Definition F.7.** A family $\mathcal{F}$ is normal on $X \subset \mathbb{C}$ if every infinite sequence of functions from $\mathcal{F}$ contains a subsequence which converges locally uniformly on $X$.

The following result, known as the Arzela-Ascoli Theorem, equates the definitions of equicontinuity and normality. A proof may be found in [1, p.222]:

**Theorem F.2.1.** ***Arzela-Ascoli Theorem.*** *Let $X \subset \mathbb{C}$ be open, and let $\mathcal{F}$ be a family of continuous functions $f\colon X \to \mathbb{C}$. Then $\mathcal{F}$ is equicontinuous on $X$ iff it is a normal family on $X$.*

Thus, points in the Fatou set are "well behaved", in the sense that given any point $z_0$ in the Fatou set and any value $\beta > 0$, we can choose a sufficiently small neighborhood $U$ of $z_0$ such that for all $n$ and for all $z_1 \in U$, the distance between $f_c^n(z_1)$ and $f_c^n(z_0)$ is less than $\beta$. On the contrary, points in the Julia set are not so well behaved, in the sense that given a point $z_0 \in J(f_c)$, and any neighborhood $U$ of $z_0$, we can choose an $n$ such that the diameter of $f_c^n(U)$ is arbitrarily large. The points in the Fatou set are also "well behaved" due to the following:

**Theorem F.2.2.** *The Fatou set and the Julia set are invariant.*

*Proof.* Since the Julia set is defined to the be compliment of the Fatou set, we must only show that the Fatou set is invariant. That is, we must show that, for any $z_0$, it is true that $z_0$ is in the Fatou set iff $f_c(z_0)$ is in the Fatou set. Equivalently (because of the Arzela-Ascoli theorem), we must show that every infinite sequence of functions $\{f_c^{n_k}(z_0)\}$ has a subsequence which converges locally uniformly on a set $X$ iff every infinite sequence of functions $\{f_c^{m_k+1}(z_0)\}$ has a subsequence which converges locally uniformly on $f_c(X)$.

For the forward direction, assume that every infinite sequence of functions $\{f_c^{n_k}(z_0)\}$ has a subsequence which converges locally uniformly on a set $X$. Let $\{f_c^{m_k+1}(z_0)\}$ be an infinite sequence of functions. Then we can take $n_k = m_k + 1$ to prove the existence of a convergent subsequence on $f(X)$. Now, assume that every infinite sequence of functions $\{f_c^{m_k+1}(z_0)\}$ has a subsequence which converges locally uniformly on a set $f(X)$. Let $\{f_c^{n_k}(z_0)\}$ be an infinite sequence of functions. Let $m_k = n_{k+1} - 1$ to prove the existence of a convergent subsequence on $X$. Q.E.D.

There are a number of alternate characterizations of the Julia set, and we will introduce a few useful ones here:

**Definition F.8.** The filled Julia set of $f_c$, denoted $K(f_c)$, is the set

$$K(f_c) = \{z \in \mathbb{C}\colon f_c^n(z) \not\to \infty \text{ as } n \to \infty\}$$

Additionally, the basin of attraction of infinity for $f_c$, denoted $A_c(\infty)$, is the set

$$A_c(\infty) = \{z \in \mathbb{C}\colon f_c^n(z) \to \infty \text{ as } n \to \infty\}$$

**Lemma F.2.3.** *$K(f_c)$ is infinite.*

*Proof.* It has been shown that for $c \neq \frac{1}{4}$, $f_c$ has a repelling fixed point $z_1 = \frac{1-\sqrt{1-4c}}{2}$. Furthermore, $z_1$ has two inverse images, which are given by $f_c^{-1}(z_1) = \pm\sqrt{\frac{1-\sqrt{1-4c}}{2} - c}$. By taking successive inverses, and since each point has two inverses, we can construct an infinite set of points whose orbits do not tend to infinity. If $c = \frac{1}{4}$, we can take the inverses of the neutral fixed point $z_0 = \frac{1}{2}$ to construct an infinite set of points in $K(f_c)$. Q.E.D.

In other words, the filled Julia set consists of those points whose forward orbits do not approach infinity, and the basin of attraction of infinity consists of those points whose forward orbits do approach infinity. As its name implies, the filled Julia set is related to the Julia set, as stated in [6, p.91]:

**Theorem F.2.4.** *$J(f_c) = \partial K(f_c)$. That is, the Julia set is the boundary of the filled Julia set.*

Another interesting result relates periodic points and the Julia set. A proof may be found in [2, p.70,109]:

**Theorem F.2.5.** *The Julia set $J(f_c)$ is equal to the closure of the set of repelling periodic points.*

By combining a few results, we have the following:

**Corollary F.2.6.** *There are an infinite number of repelling periodic points.*

*Proof.* By Lemma F.2.3, $K(f_c)$ is infinite. Thus, $J(f_c) = \partial K(f_c)$ must be infinite, and by applying Theorem F.2.5, we see that the set of repelling periodic points is infinite. Q.E.D.

## F.3 The Structure of the Fatou Set

In this section, we will explore $A_c(\infty)$, and then use it to describe the Fatou set. First, however, we need to provide a few results from complex analysis:

**Theorem F.3.1.** ***Maximum Principle.*** *Let $g\colon \mathbb{C} \longrightarrow \mathbb{C}$ be nonconstant and differentiable on an open set $U$. Then $|g(z)|$ does not attain a maximum on $U$.*

*Proof.* Let $z_0 \in U$, and let $V \subset U$ be a neighborhood around $z_0$. Since the image of an open set is an open set, $g(V)$ must be open, and $g(z_0) \in g(V)$. Then there must exist a point $g(z_1) \in g(V)$ such that $|g(z_1)| > |g(z_0)|$. Q.E.D.

The next result is also useful when characterizing the Fatou set. It's proof can be found in [1, §3.3]:

**Theorem F.3.2.** ***Montel's Theorem.*** *Let $U \subset \widehat{\mathbb{C}}$ be open, and let $\mathcal{F}$ be a family of functions such that $(\forall f \in \mathcal{F})(\exists$ distinct $a_f, b_f, c_f \in \widehat{\mathbb{C}})$ such that $f$ does not take the values $a_f$, $b_f$, and $c_f$ in $U$. Then $\mathcal{F}$ is normal in $U$.*

Now, we consider $A_c(\infty)$ in some detail.

**Theorem F.3.3.** *$A_c(\infty)$ is an open subset of the Fatou set.*

*Proof.* First, it is obvious that $A(\infty)$ is completely invariant. Now, let $U_R = \{z\colon |z| > R\}$, and let $R$ be sufficiently large such that $|f_c(z)| = |z^2 + c| > 2|z|$ on $U_R$, so that $\{f_c^n(z)\}$ converges uniformly to $\infty$ on $U_R$. Thus, $U_R$ is a subset of both the Fatou set and $A_c(\infty)$. Because both $A_c(\infty)$ and the Fatou set are completely invariant, we have that

$$A_c(\infty) = \bigcup_{i=0}^{\infty} f_c^{-i}(U_R)$$

is an open subset of the Fatou set (open because it is the union of the inverse images of open sets, which are open). Q.E.D.

Next, we wish to describe the interior of the filled Julia set. In particular, we wish to prove that the interior of the filled Julia set consists of open components which are simply connected. First, however, we must show that $A_c(\infty)$ is connected and that $\partial A_c(\infty) = J(f_c)$.

**Lemma F.3.4.** *$A_c(\infty)$ is connected.*

*Proof.* Since $A_c(\infty)$ is completely invariant, we must have that $\partial A_c(\infty)$ is also completely invariant. Let $\widetilde{A}$ be an (open) bounded component of $\mathbb{C} - \partial A_c(\infty)$. Then, by the maximum principle, $\sup\{|z|\colon z \in f_c(\widetilde{A})\}$ occurs on $\partial A_c(\infty)$, so $f_c(\widetilde{A})$ is an open bounded component of $\mathbb{C} - \partial A_c(\infty)$. It follows that $\sup\{|z|\colon z \in f_c^n(\widetilde{A})\}$ occurs on $\partial A_c(\infty)$. This means that no bounded component of $\mathbb{C} - \partial A_c(\infty)$ maps to a neighborhood of $\infty$. Thus, $A_c(\infty)$ consists of one component, the unbounded component of $\mathbb{C} - \partial A_c(\infty)$. Therefore, $A_c(\infty)$ is connected. Q.E.D.

**Lemma F.3.5.** $\partial A_c(\infty) = J(f_c)$

*Proof.* First, let $z_0 \in J(f_c)$. Then $\{f_c^n\}$ is not equicontinuous on any neighborhood $U$ of $z_0$. By Montel's theorem, this implies that $\{f_c^n(U)\}$ omits at most two points in $\widehat{C}$. Hence, $f(U) \cup A_c(\infty) \neq \emptyset$, and $z_0 \in \partial A_c(\infty)$, so $J(f_c) \subset \partial A_c(\infty)$.

Now, let $z_0 \in \partial A_c(\infty)$. Let $U$ be a neighborhood of $z_0$. Then $\{f_c^n(z)\}$ converges to $\infty$ on $A_c(\infty) \cup U$. However, $\{f_c^n(z_0)\}$ is bounded, so no subsequence of $\{f_c^n(z)\}$ can converge on $U$. Thus, another application of Montel's theorem shows that $z_0 \in J(f_c)$, so $\partial A_c(\infty) \subset J(f_c)$. Q.E.D.

**Theorem F.3.6.** *The interior of the filled Julia set consists of open components, each of which is simply connected. That is, they are topologically equivalent to an open disk.*

*Proof.* By definition, the interior of the filled Julia set is open. Because $A_c(\infty)$ is connected, we must have that $A_c(\infty) \cup \partial A_c(\infty) = A_c(\infty) \cup J(f_c)$ is connected. Finally, since $\widehat{\mathbb{C}}$ is the disjoint union of the interior of the filled Julia set, $J(f_c)$, and $A_c(\infty)$, we must have that the components of the interior of the filled Julia set are simply connected. Q.E.D.

For values of $c$ such that $f_c$ has an attracting cycle, we can further characterize the Fatou set of $f_c$ by the following:

**Definition F.9.** The immediate basin of attraction of an attracting cycle $\{z_1, z_2, \ldots, z_n\}$ of period $n$ is the subset of components $B_i$ in the Fatou set such that $z_i \in B_i$, where $i \in \{1, 2, \ldots, n\}$. Moreover, the basin of attraction of an attracting cycle $\{z_1, z_2, \ldots, z_n\}$ of period $n$ is the set of all components $B$ in the Fatou set such that $\lim_{m\to\infty} f_c^{mn}(B) = B_i$, where $i \in \{1, 2, \ldots, n\}$.

The next result is quite important, and has far-reaching consequences. For a proof, see [2, p.195].

**Theorem F.3.7.** *The immediate basin of each attracting cycle of $f_c$ contains a critical point.*

This leads to the following corollary:

**Corollary F.3.8.** *$f_c$ has at most one attracting cycle, and 0 must be contained in the immediate basin of attraction for this cycle.*

*Proof.* The only critical point (a point at which the derivative vanishes) for the map $f_c(z) = z^2 + c$ is $z = 0$. Thus, if $f_c$ has an attracting cycle, Theorem F.3.7 requires that $z = 0$ must be in its immediate basin. Q.E.D.

We can use Corollary F.3.8 to begin to construct a model of the filled Julia set. If $f_c$ has an attracting cycle, then we know that there is a component of $K_c$ which contains the origin. We will hereafter denote this particular component as $B_*$. Additionally, Theorem F.3.6 suggests that the closures of the components of $K(f_c)$ meet at points belonging to $J(f_c)$. Specifically, we will focus on maps $f_c$ whose filled Julia set contains an attracting cycle with period $q$ such that the closures of the components $B_*, f_c(B_*), \ldots, f_c^{q-1}(B_*)$ (in other words, the closures of the components of the immediate basin of attraction for the attracting cycle $\{z_0, f_c^1(z_0), \ldots, f_c^{q-1}(z_0)\}$) meet at a single point $P^0$. Interestingly, these components $B_*, f_c(B_*), \ldots, f_c^{q-1}(B_*)$ behave predictably, in the sense that each component will move counterclockwise around $P^0$ by a predetermined number of components. We thus define:

**Definition F.10.** The rotation number for $f_c$ is a fraction $\frac{p}{q}$, where $q$ is the period of the attracting orbit, and each component of the immediate basin of attraction of $f_c$ rotates $p$ components counterclockwise under iteration of $f_c$.

Figure F.1: Julia sets with rotation number 1/4 (left) and 2/5 (right)

What we will focus on now is the process of finding values of $c$ that meet the above criteria.

## F.4 External Rays

This section will provide the tools necessary to show where the components of $K(f_c)$ map under $f_c$. This is done partially by comparing $f_0$ and $f_c$. First, note that the dynamics for $f_0(z) = z^2$ are very well-behaved. In fact, 0 is one of only two values of $c$ (the other being $-2$) such that $J(f_c)$ can be described explicitly. It should be noted that the squaring function $f_0(z) = z^2$ squares the modulus and doubles the argument of $z$, since if we represent a point $z$ as $z = re^{i\theta}$, then $z^2 = r^2e^{i(2\theta)}$. Thus, the forward orbit of any point $z$ such that $|z| > 1$ approaches $\infty$, while the forward orbit of a point $z$ such that $|z| < 1$ approaches 0. Finally, for any point $z_0$ such that $|z_0| = 1$, any point $w \in O^+(z_0)$ in the forward orbit of $z_0$ will also have $|w| = 1$. Thus, we have ascertained the following:

$$K(f_0) = \{z \colon |z| \leq 1\}$$
$$J(f_0) = \partial K(f_0) = \{z \colon |z| = 1\}$$
$$A_0(\infty) = \{z \colon |z| > 1\}$$

The next theorem helps to relate $A_c(\infty)$ and $A_0(\infty)$.

**Theorem F.4.1.** *If $f_c$ has an attracting cycle, then there exists an invertible mapping $\Phi_c\colon \mathbb{C} - K(f_c) \longrightarrow \{z\colon |z| > 1\}$ such that $\Phi_c(f_c(z)) = (\Phi_c(z))^2$.*

*Proof.* Although a detailed proof may be found as a special case of a theorem stated in [6, p.91], we will nonetheless give an explanation of how to find such a map $\Phi_c$. We can let

$$\Phi_c(z) = \lim_{n\to\infty} (f_c^n(z))^{1/2^n}$$

We refer to [6] here in order to solve the problem of determining how to choose a $(2^n)^{\text{th}}$ root of $f_c^n$ that is one-to-one and onto. However, if we assume that we can choose such a map $\Phi_c$ that is invertible, then

$$\Phi_c(f_c(z)) = \lim_{n\to\infty} (f_c^n(f_c(z)))^{1/2^n} = \lim_{n\to\infty} (f_c^{n+1}(z))^{1/2^n} = \lim_{n\to\infty} (f_c^n(z))^{1/2^{n-1}} = (\Phi_c(z))^2$$

Q.E.D.

Hence, the following diagram is commutative:

$$\begin{array}{ccc} A_c(\infty) & \xrightarrow{f_c} & A_c(\infty) \\ {\scriptstyle \Phi_c}\downarrow & & \downarrow{\scriptstyle \Phi_c} \\ A_0(\infty) & \xrightarrow[f_0]{} & A_0(\infty) \end{array}$$

Now, we will describe external rays and how they allow us to understand the $K_c$.

**Definition F.11.** The external rays $R_t$ for the map $f_0$ are defined by

$$R_t = \{re^{2\pi it}\colon r > 1\}$$

Note that $R_t = R_{t+k}$, where $k \in \mathbb{Z}$. Therefore, we will limit $t$ to the interval $[0, 1)$. Furthermore, since $R_t$ consists only values whose modulus is greater than 1, $R_t \subset A_0(\infty)$.

**Definition F.12.** The external ray $R_t$ lands at a point $z_0$ if

$$\lim_{r\to 1^+} R_t = z_0$$

Hence, given a point $z_0 = e^{2\pi it} \in J(f_0)$, the external ray $R_t$ lands at $z_0$. We can use the map $\Phi_c$ to define the external rays $R_t$ for an arbitrary value $c$.

**Definition F.13.** The external ray $R_t$ for a map $f_c$ is defined by

$$R_t = \Phi_c^{-1}(re^{2\pi it}), \quad r > 1$$

It has been shown that, if $f_c$ has an attracting orbit, then every external ray will land (see [6, p.176,195]). Therefore, since $e^{2\pi it} \in J(f_0)$, then $\lim_{r\to 1^+} \Phi_c^{-1}(re^{2\pi it}) \in J(f_c)$. This allows us to understand the dynamics of $f_c$ in terms of the dynamics of the more-easily-understood dynamics of $f_0$.

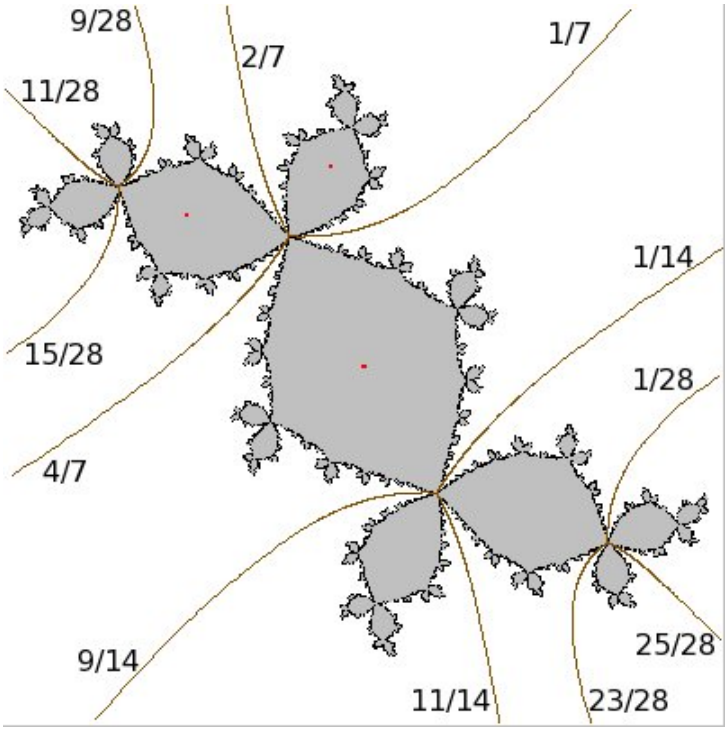



Figure F.2: External rays for $f_{-0.122561+0.744862i}$ (rotation number 1/3)

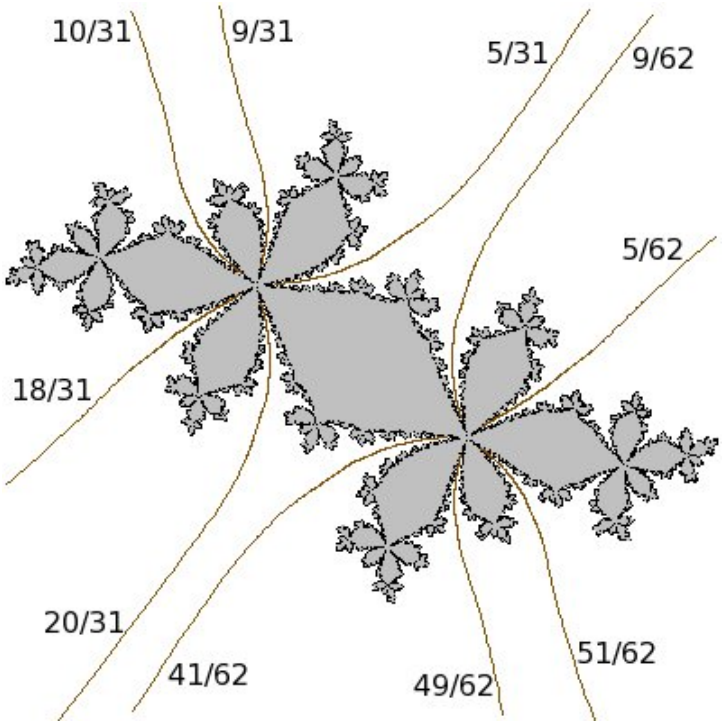



Figure F.3: External rays for $f_{-0.50434+0.562766i}$ (rotation number 2/5)

## F.5 The Mandelbrot Set

We wish to find values $c \in \mathbb{C}$ such that closures of the components of the immediate basin of attraction for the attracting cycle $\{z_0, f_c^1(z_0), \ldots, f_c^{q-1}(z_0)\}$ meet at a single point. To do this, we introduce the Mandelbrot set (pictured in Figure F.4).

**Definition F.14.** The Mandelbrot set $\mathcal{M}$ is the set

$$\mathcal{M} = \{c \in \mathbb{C} \colon f_c^n(0) \not\to \infty \text{ as } n \to \infty\}$$

The mapping $\Phi_c$, defined in Theorem F.4.1, is helpful here also. The next theorem compares $A_0(\infty)$ and $\mathbb{C} - \mathcal{M}$, and is presented in [3].

**Theorem F.5.1.** *There exists an invertible mapping* $\Psi \colon \mathbb{C} - \mathcal{M} \longrightarrow \{z \colon |z| > 1\}$. *Moreover,* $\Psi(c) = \Phi_c(c)$.

We can use this to define external rays in the parameter space:

**Definition F.15.** The external ray $R_t$ of $\mathcal{M}$ with angle $t$ is defined as

$$R_t = \Psi^{-1}(re^{2\pi it}), \quad r > 1$$

The largest component of the Mandelbrot set is a cardioid. For all values of $c$ in this main cardioid, $f_c$ has an attracting fixed point. Thus, the main cardioid is bounded by the curve $g(\theta) = -\left(\frac{e^{i\theta}}{2} + \left(\frac{e^{i\theta}}{2}\right)^2\right)$, as proven in Theorem F.1.1. We can then use Schleicher's algorithm (as given in [3]) to name the bulbs on the Mandelbrot set and to determine which external rays land at the base of each bulb. Here, the external rays will be given in their binary expansion. Thus, the ray $\overline{r_1 r_2 \ldots r_n}$, where $r_i \in \{0, 1\}$, represents the unending binary decimal $0.r_1 r_2 \ldots r_n r_1 r_2 \ldots r_n r_1 r_2 \ldots$. For example, the external ray $R_{1/7}$ is equivalent to $\overline{001}$. Figure F.5 shows the result of a few applications of Schleicher's algorithm.

**Schleicher's Algorithm:** The main cardioid is defined as the 0/1 or 1/1 bulb. The rays $\overline{0}$ and $\overline{1}$ land at its cusp. Furthermore, the largest bulb connected to the main cardioid is the 1/2 bulb. The ray $\overline{01}$ lands at the 1/2 bulb from above, the ray $\overline{10}$ lands from below.

Locate the largest bulb between two already-named bulbs $p_1/q_1$ and $p_2/q_2$. This is the $\frac{p_1}{q_1} \oplus \frac{p_2}{q_2} = \frac{p_1+p_2}{q_1+q_2}$ bulb. (The operation $\oplus$ is known as Farey addition.)

Find the rays closest to the $\frac{p_1+p_2}{q_1+q_2}$ bulb. One ray, $\overline{r_1}$, will be connected to the $\frac{p_1}{q_1}$ bulb, and the other, $\overline{r_2}$, will be connected to the $\frac{p_2}{q_2}$ bulb. Then, the ray landing on the $\frac{p_1+p_2}{q_1+q_2}$ bulb closest to the $\frac{p_1}{q_1}$ bulb is $\overline{r_1 r_2}$, and the ray landing on the $\frac{p_1+p_2}{q_1+q_2}$ bulb closest to the $\frac{p_2}{q_2}$ bulb is $\overline{r_2 r_1}$.

For values of $c$ within the components of $\mathcal{M}$ connected to the main cardioid, $f_c$ has an attracting cycle. Moreover, the immediate basin of attraction for this cycle consists of components whose closures meet at a single point. Finally, if $c$ is within the $p/q$ bulb of the Mandelbrot set, then the rotation number for $f_c$ is $p/q$. If the external rays $R_{t_1}$ and $R_{t_2}$ land at the $p/q$ bulb of $\mathcal{M}$, then $R_{t_1}$ and $R_{t_2}$ will land at the base of the component $f_c(B_*)$. By taking forward and inverse image of these rays, we can discover which rays land at the base of any component. Now, we may use this information to construct an algorithm to determine the structure of $K(f_c)$.

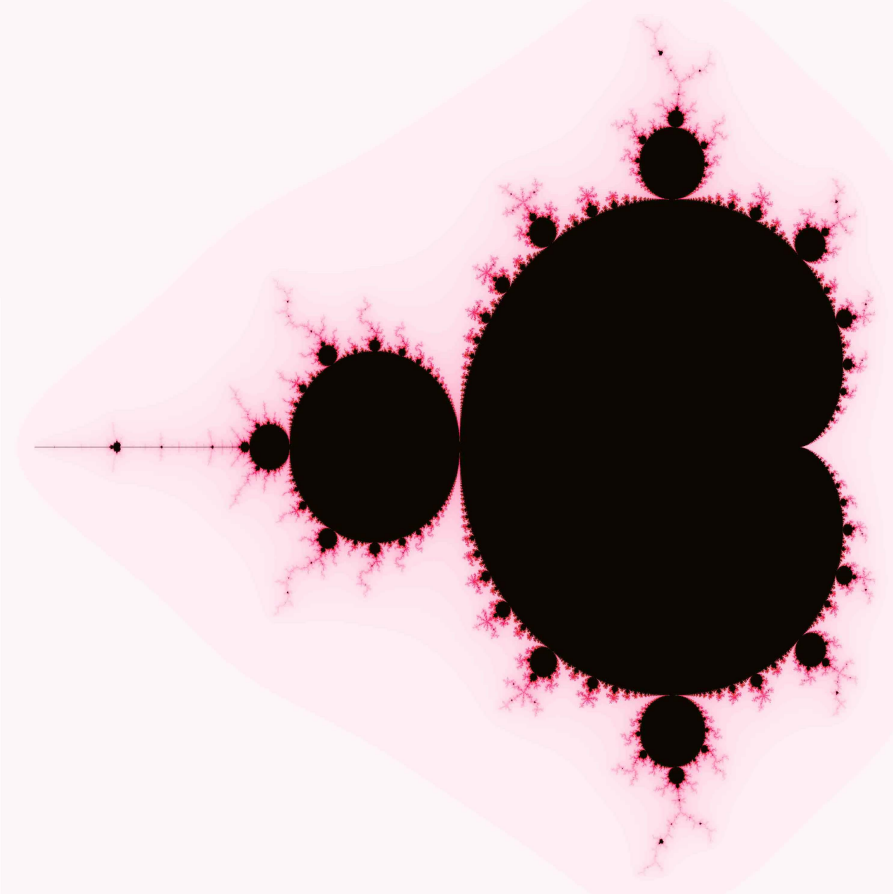


Figure F.4: Mandelbrot Set

## F.6 Constructing the Graph of $K(f_c)$

The following algorithm will construct a sequence of graphs which will represent successive approximations of $K(f_c)$ for values of $c$ specified above. The graph completed to the $n^{\text{th}}$ stage will be denoted $G^n$, for $n \in \mathbb{N}$. There will be two different types of vertices in the graph: those labeled $C^n_{i_\phi}$, and those labeled $P^{j_\theta}_{C^n_{i_\phi}}$, although some of these indices may be omitted occasionally either when the value of the index is not important, or when the value of the index is not known. Whereas the $C$ vertices represent the components of $K(f_c)$, the $P$ vertices represent the points in $J(f_c)$ at which the components in $K(f_c)$ meet. Therefore, each edge in $G^n$ will connect a $C$ vertex and a $P$ vertex. Finally, throughout this algorithm, angles (represented by $\phi$ and $\theta$) will be measured in *turns*. Turns may be converted into radians using the following: 1 turn $= 2\pi$ radians. Therefore, we will constrain $\phi$ and $\theta$ to the interval $[0, 1)$. All angles are measured counterclockwise from the horizontal.

This algorithm is used to give a sequence of planar graphs. The indices on $C^n_{i_\phi}$ and $P^{j_\theta}_{C^n_i}$ will help to embed $G^n$ in the plane, and will now be explained in more detail. For a vertex $C^n_{i_\phi}$, the number $n \in \mathbb{N}$ signifies that a vertex $C$ has been created at the $n^{\text{th}}$ stage. The number $i \in \mathbb{N}$ is an index which will distinguish the various vertices created at the $n^{\text{th}}$ stage; it will be used for showing which components represented by a vertex $C^n$ map to which component represented by $C^{n-1}$. Now, we define the *base point* of a component $C^n_{i_\phi}$, where $n \geq 1$, to be the unique vertex $P^{j_\theta}_{C^m_{i_\phi}}$ connected to $C^n_{i_\phi}$ such that $m < n$. Then, $\phi$ represents the angle that the vector from the base point of $C^n_{i_\phi}$ to $C^n_{i_\phi}$ makes with the horizontal. Vertices $P^{j_\theta}_{C^n_i}$, henceforth called *junctures*, are created so that each $P^{j_\theta}_{C^n_i}$ is adjacent to the vertex $C^n_{i_\phi}$. The vector from $C^n_{i_\phi}$ to $P^{j_\theta}_{C^n_i}$ makes an angle $\theta$ with the horizontal. The particular use of the index $j \in \mathbb{N}$ will become apparent throughout the course of the algorithm.

To create the graph $G$ for $f_c$, first we must construct the initial graph $G^0$ consisting of the one vertex $C^0_{0_0}$. Next, we give a procedure to construct $G^n$ from $G^{n-1}$. This procedure is split in three parts. First, we discuss how to add the components $C^n$. Then, we provide instructions for indexing the $C^n_{i_\phi}$. Then, we determine where to add junctures. Afterwards, we show how to use the index $i$ to determine which components of $G^n$ map to which component of $G^{n-1}$.

### F.6.1 Creating $G^0$

First, create the graph $G^0$, consisting of a single vertex labeled $C^0_{0_0}$. Draw an edge from $C^0_{0_0}$ to two junctures labeled $P^{0_0}_{C^0_{0_0}}$ (at 0 radians) and $P^{0_{1/2}}_{C^0_{0_0}}$ (at $\pi$ radians).

### F.6.2 Step 1: adding vertices $C^n$ to $G^{n-1}$

To each vertex $C^m$ such that $n \not\equiv m \pmod q$, add a $C^n$ to each $C^m$ at each $P^{j_\theta}_{C^m}$ such that $j = \lfloor \frac{n-m}{q} \rfloor$. These points should be labeled $C^n_{i_\phi}$ (keep the $i$ as an undetermined constant for now), where $\phi = \theta + \frac{1}{2} - \left(\frac{p}{q}\right)(n-m) \pmod 1$. Also, the vector from the base point $P^{j_\theta}_{C^m}$ of $C^n_{i_\phi}$ to $C^n_{i_\phi}$ should have angle $\phi$.

### F.6.3 Step 2: labeling the $C^n$

For $n = 1$, go to the unique component $C^1$ connected to $P^{0_0}_{C^0_{0_0}}$. Otherwise, go to the unique component $C^n$ connected to $P^0_{C^{n-1}_0}$. Label this component as $C^n_{0_\phi}$. Let $l = 1$. From $C^n_{0_\phi}$, and facing $P^0_{C^{n-1}_0}$, make a clockwise traversal around the graph $G^n$ (as in figure F.6). This is equivalent to traversing the entire graph $G^n$ by only making left turns. Whenever a component $C^n$ is encountered, label this component $C^n_l$, and then increase the value of $l$ by 1. This processes mimics the clockwise traversal of $\bigcup_{k=0}^n f^{-k}(B_*)$ starting at the fixed point $z_0 = \frac{1+\sqrt{1-4c}}{2}$. For the moment, we treat this procedure as intuitively clear, but we will give a precise algorithm for the clockwise traversal of $G^n$ in section F.8.

### F.6.4 Step 3: adding junctures $P^{j\theta}_{C^m_{i_\phi}}$

Consider all the vertices $C^m_i$ such that $m \equiv n \pmod q$. Let $j = \frac{m-n}{q}$. Attach $2^j$ junctures $P^j_{C^m_i}$ to vertices $C^m_i$. The vectors from $C^m_{i_\phi}$ to $P^{j_\theta}_{C^m_{i_\phi}}$ should have angles $\theta = \phi + \frac{1}{2} + \frac{2k+1}{2^{j+1}} \pmod 1$, where $k = \{0, 1, 2, \ldots, 2^j - 1\}$.

These angles $\theta$ ensure that no two edges from $C^m_i$ to the points $P^j_{C^m_i}$ are consecutive. Since there will be $1 + 2^0 + 2^1 + \cdots + 2^{j-1} = 2^j$ points $P_{C^i_m}$ already connected to $C^m_i$, there is a unique way to add the new points $P^j$ such that no two edges from $C$ to $P^j_C$ are adjacent.

### F.6.5 An Example

The following graphs show successive iterations of the algorithm applied to a map $f_c$ with rotation number 1/3 (as in figure F.2).

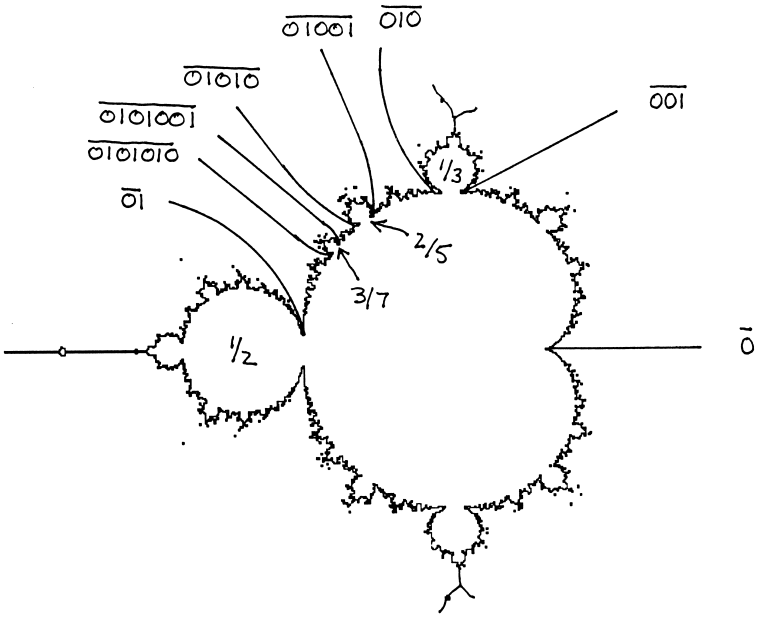



Figure F.5: Schleicher's Algorithm

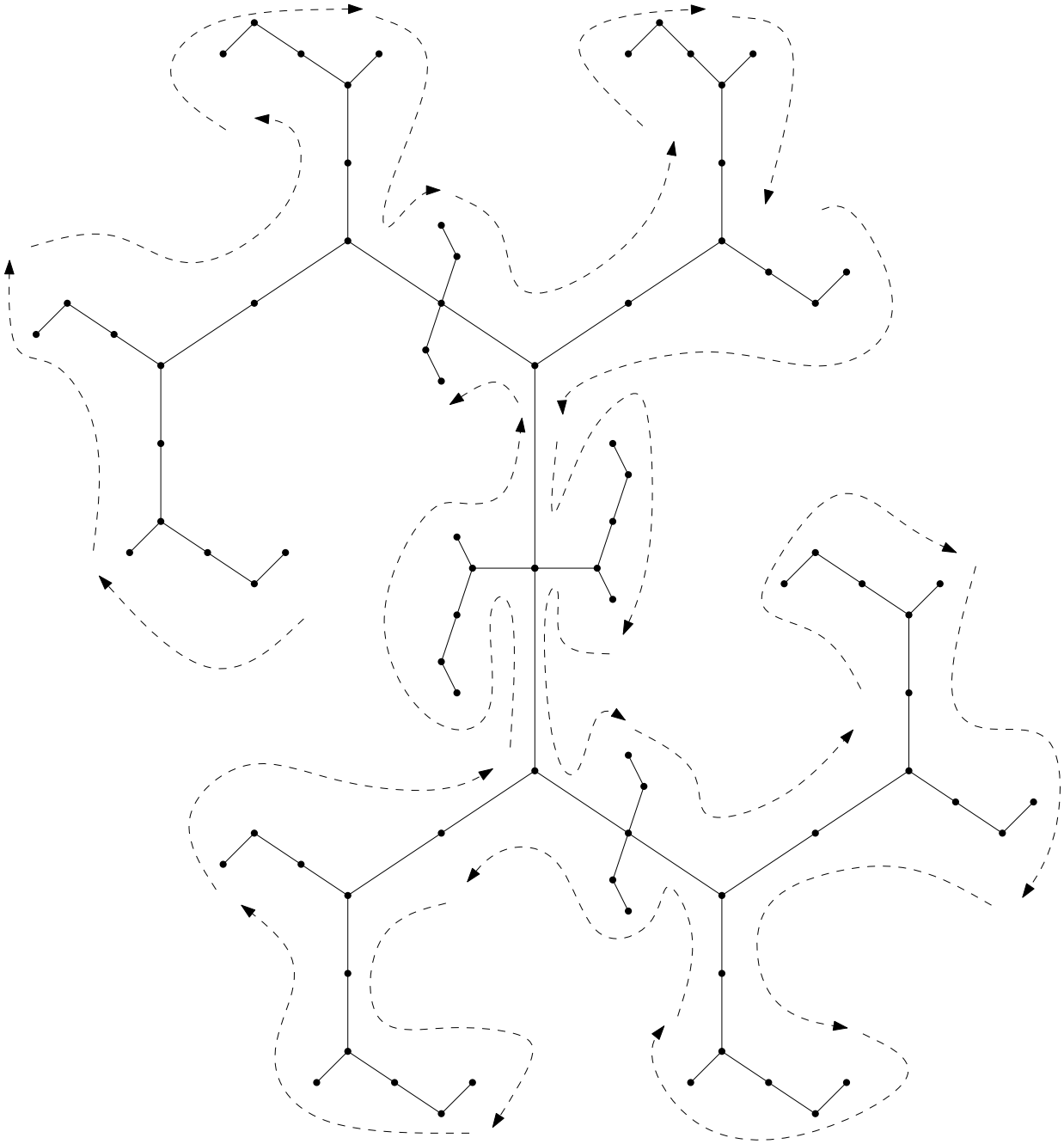


Figure F.6: $G^0$

Figure F.7: $G^0$

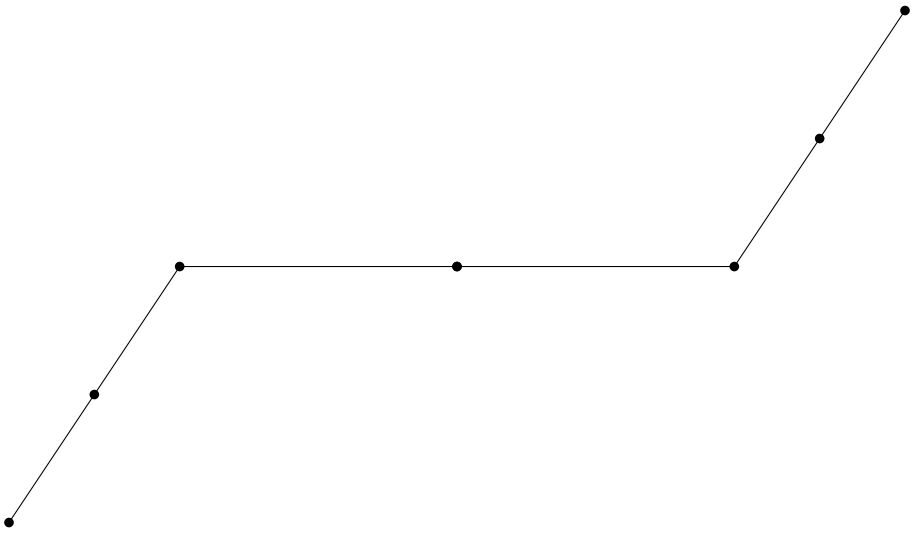


Figure F.8: $G^1$

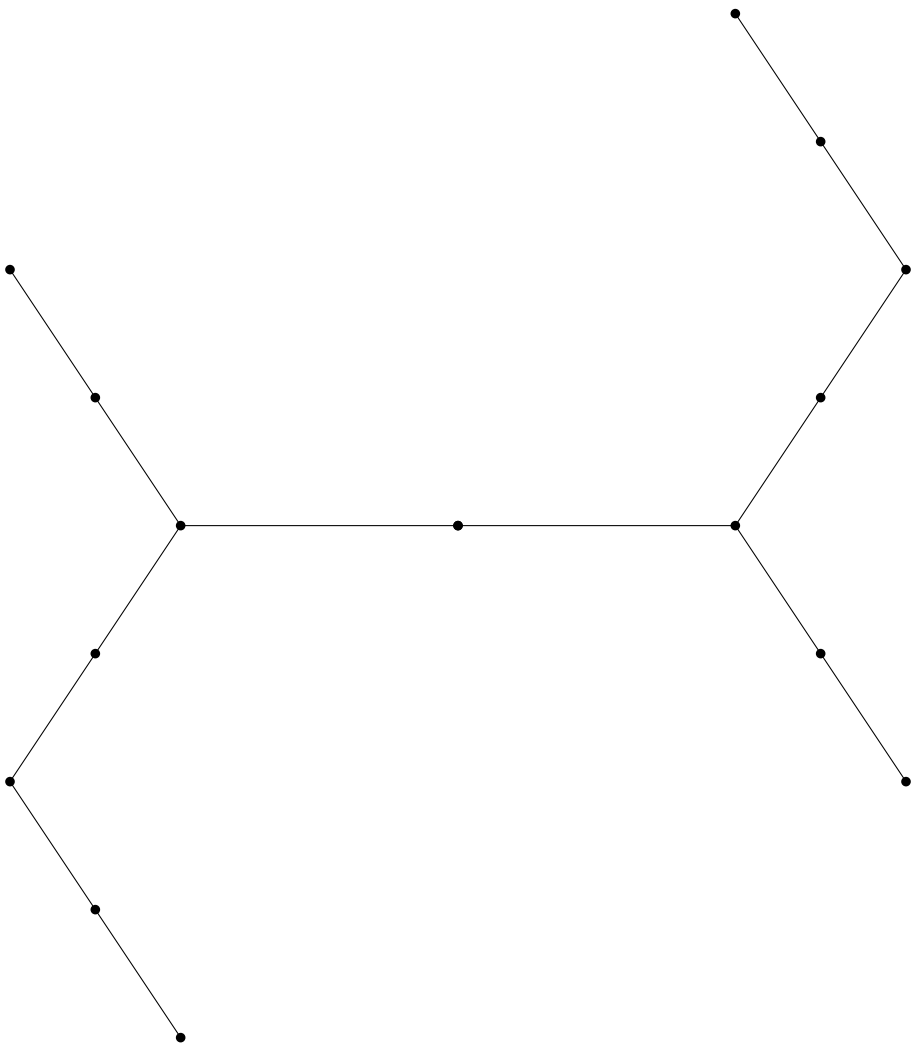


Figure F.9: $G^2$

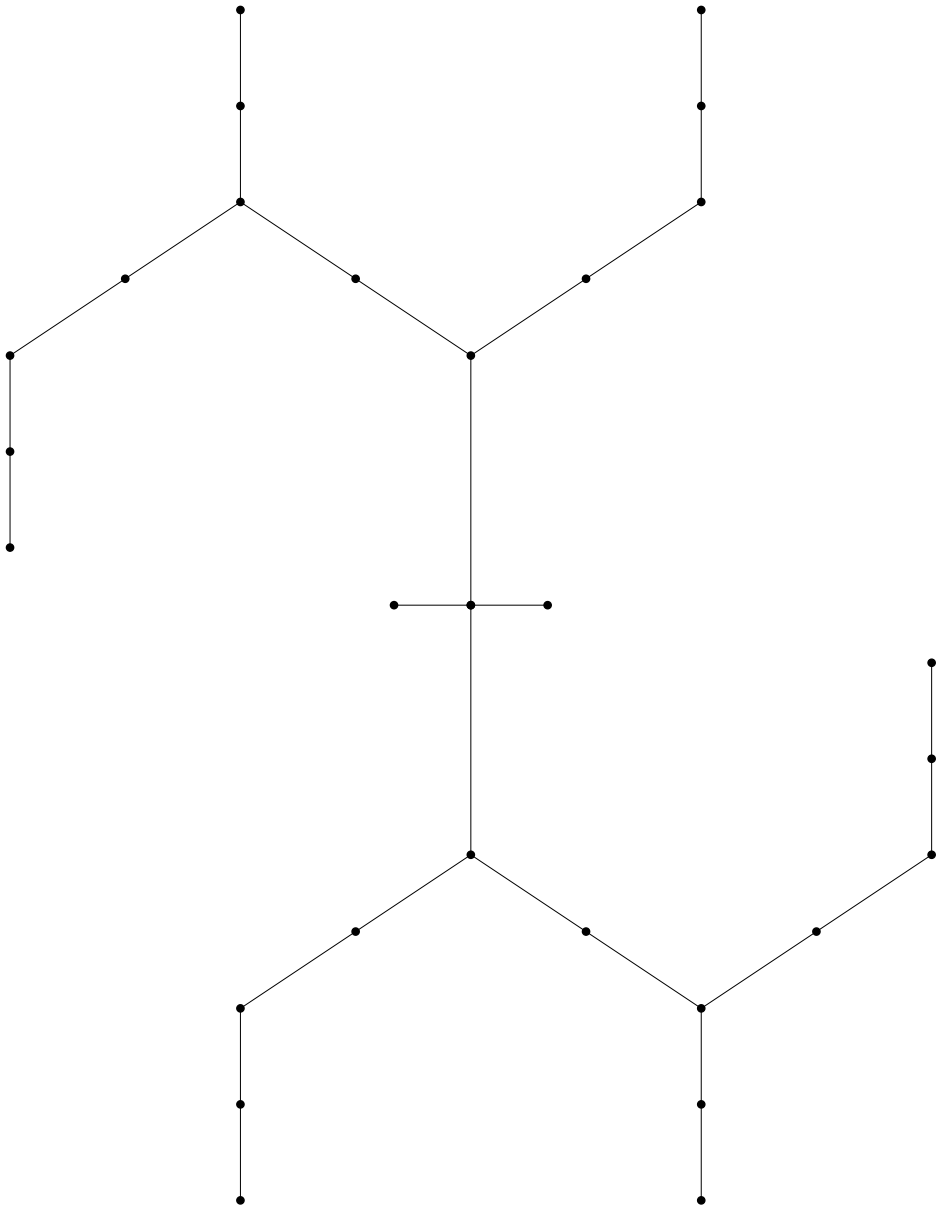


Figure F.10: $G^3$

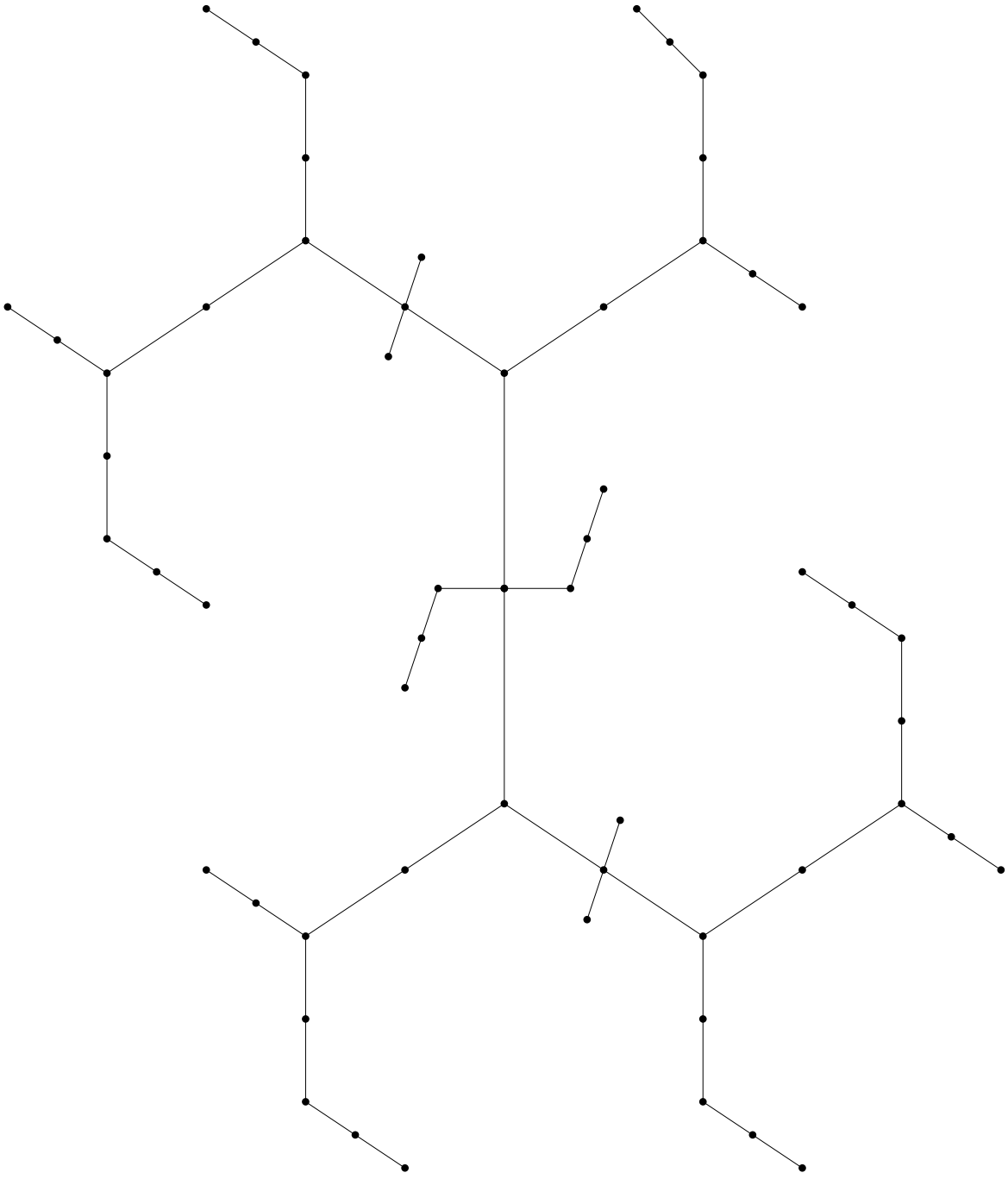


Figure F.11: $G^4$

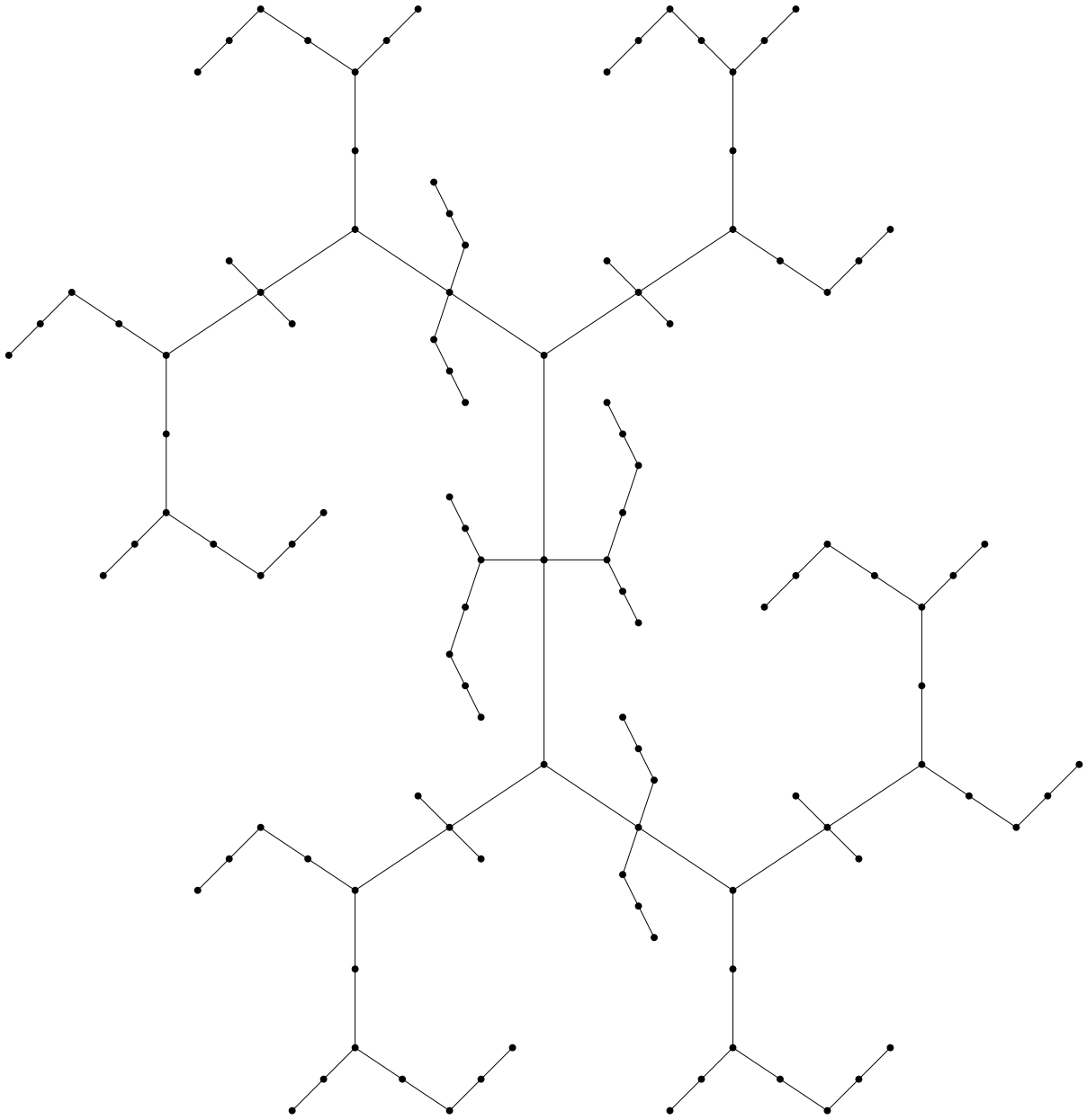


Figure F.12: $G^5$

## F.7 Interpreting the Algorithm

Our algorithm begins by creating a vertex representing the component $B_*$ containing the origin (Corollary F.3.8 requires that this exists). The construction of $C^n$ from $C^{n-1}$ replicates taking the $n^{\text{th}}$ inverse image of $B_*$. The following theorems show that our algorithm constructs an accurate approximation of $K(f_c)$. For a proof of theorem F.7.1, consult [6, p.234]

**Theorem F.7.1.** ***No Wandering Domains.*** *Let $g\colon \widehat{\mathbb{C}} \to \widehat{\mathbb{C}}$ be a rational map. Then every component in the Fatou set of $g$ is eventually periodic.*

**Corollary F.7.2.** $O^-(B_*) = K(f_c)$

*Proof.* Let $B \in K(f_c)$ be a component of the interior of $K(f_c)$. By theorem F.7.1, the forward orbit of $B$ is eventually periodic. Therefore, the forward orbit of $B$ must eventually meet the immediate basin of attraction for either an attracting cycle or a parabolic cycle (a cycle which contains neutral periodic points). As stated on [6, p.112], the total number of attracting orbits and parabolic orbits for $f_c$ is 2. We know that $\infty$ is an attracting fixed point for $f_c$, and by hypothesis, there is an attracting orbit. By definition, $O^+(B) \cap A_c(\infty) = \emptyset$, and therefore, $O^+(B) \cap B_* \neq \emptyset$. Thus, $O^-(B_*) = K(f_c)$ Q.E.D.

We now define a function, $N\colon \mathbb{N} \to \mathbb{N}$. Let $N(n)$ give the number of components $C^n$ created during the construction of $G^n$. If $p/q$ be the rotation number for $f_c$, then we can give an explicit formula for $N(n)$ as

$$N(n) = \left\{ \begin{array}{ll} 2^n & n \leq q-1 \\ \left(2^{q-1}-1\right)\left(2^{n-q+1}\right) & n > q-1 \end{array} \right\}$$

By constructing the graph up to $G^n$, for $n$ sufficiently high ($n = 20$ seems to be sufficient), and then by connecting all of the points $P$ surrounding a vertex $C$, we can create a picture which is topologically equivalent to $K(f_c)$. Thus, each vertex $C$ can now represent a component of $K(f_c)$. Furthermore, we can use the indices $i$ to show where the components $C^m_{i_\phi}$ map under $f_c$.

**Theorem F.7.3.** *Let $C^{q-1}_{I_{p/q}}$ represent the component attached to $P^{0_{1/2}}_{C^0_{0_0}}$ whose angle is $\phi = p/q$. Furthermore, define $\widetilde{i}$ as:*

$$\widetilde{i} = \left\{ \begin{array}{lll} i \pmod{N(q-1)} & & i < I \\ i+1 \pmod{N(q-1)} & & I \leq i < I + \frac{N(q)}{2} \\ i+2 \pmod{N(q-1)} & & i \geq I + \frac{N(q)}{2} \end{array} \right\}$$

*Then the following is true:*

$$f_c(C^m_i) = \left\{ \begin{array}{ll} C^{q-1}_I & m = 0 \\ C^{q-1}_{\widetilde{i}} & m = q \\ C^{m-1}_{i \pmod{N(m-1)}} & m \neq 0 \text{ and } m \neq q \end{array} \right\}$$

*Proof.* Recall the map $\Phi_c$, defined in Theorem F.4.1. For the maps $f_c$ which we are considering, every external ray lands. We can define a surjective map $\Upsilon_c \colon J(f_0) \to J(f_c)$ given by

$$\Upsilon_c(e^{2\pi it}) = \lim_{r \to 1^+} \Phi_c^{-1}(re^{2\pi it})$$

Thus, $\Upsilon_c$ associates a point in $J(f_c)$ to each point in $J(f_0)$. Moreover, since $\Phi_c(f_c(z)) = (\Phi_c(z))^2$, it follows that $\Upsilon_c(f_0(z)) = f_c(\Upsilon_c(z))$. That is, the following diagram is commutative:

$$\begin{array}{ccc} J(f_0) & \xrightarrow{f_0} & J(f_0) \\ {\scriptstyle \Upsilon_c}\downarrow & & \downarrow{\scriptstyle \Upsilon_c} \\ J(f_c) & \xrightarrow[f_c]{} & J(f_c) \end{array}$$

Now, we parameterize $J(f_0)$ by $\gamma(t) = e^{-2\pi it}$. As $t$ goes from 0 to 1, then while $\gamma(t)$ goes clockwise around $J(f_0)$ once, $f_0(\gamma(t)) = e^{-4\pi it}$ goes around the $J(f_0)$ twice. Similarly, as $t$ goes from 0 to 1, $\Upsilon_c(\gamma(t)) = \Upsilon_c(e^{-2\pi it})$ goes clockwise around $J(f_c)$ once, and

$$f_c(\Upsilon_c(\gamma(t))) = \Upsilon_c(f_0(\gamma(t))) = \Upsilon_c(e^{-4\pi it})$$

goes clockwise around $J(f_c)$ twice. The clockwise traversal of $G^n$, given in the second step of the algorithm, mimics the clockwise traversal of $J(f_c)$ by the mapping $\Upsilon_c(\gamma(t))$. Furthermore, the double clockwise cover of $J(f_c)$ by $f_c(\Upsilon_c(\gamma(t)))$ justifies the formula $f_c(C_i^m) = C_{i \pmod{N(m-1)}}^{m-1}$ for $m \neq 0$ and $m \neq q$. The main problem for $m = 0$ and $m = q$ is that the components in both $C^q$ and $C^0$ map to components in $C^{q-1}$. So, we isolate the one component to which $C_0^0$ maps ($C_I^{q-1}$), and "skip over" this component when constructing a formula for $f_c(C_i^q)$ (this results in our formula for $\widetilde{i}$). Q.E.D.

# F.8 The Labeling Algorithm

In Section F.6.3, we give instructions to "traverse $G^n$ clockwise". We realize that, for programming purposes, this is surprisingly unhelpful. Therefore, we provide a relatively detailed outline, written in pseudo-code, of a program that traverses $G^n$ clockwise. First, however, we will give answers to several issues that had to be considered when writing this program and give guidelines that should be followed when writing a computer program.

This program requires the rotation number $(p/q)$ and the completed (fully indexed) graph $G^{n-1}$. One can begin the program by calling the `label_components (p q n)` procedure. The program starts by locating the path

$$\mathcal{P} = C_0^0 P_{C_0^0}^{0_0} C_0^1 \dots C_0^{n-1}$$

Then, beginning on $C_0^0$, and facing the next vertex in $\mathcal{P}$ (this is $P_{C_0^0}^0$), the program applies recursive procedures which mimic a clockwise traversal of each edge connected to the current vertex.

The following facts needed to be ascertained before writing the program. Let

$$[m \neq 0] = \left\{ \begin{array}{ll} 1 & m \neq 0 \\ 0 & m = 0 \end{array} \right\}$$

Then there are $2^{\lceil \frac{n-m}{q} \rceil} - [m \neq 0]$ vertices $P^{j_\theta}_{C^m_{i_\phi}}$ which are connected to $C^m_{i_\phi}$. The vectors from $C^m_{i_\phi}$ to the $P^{j_\theta}_{C^m_{i_\phi}}$s will have angles

$$\theta = \frac{1}{2} + \phi - \frac{k}{2^{\lceil \frac{n-m}{q} \rceil}} \pmod 1, \quad k = \{1, \ldots, 2^{\lceil \frac{n-m}{q} \rceil} - [m \neq 0]\}$$

Furthermore, there are $\max\{q-1, n-m-qj\}$ vertices $C^\mu_{\varepsilon_\Omega}$ which are connected to $P^{j_\theta}_{C^m_{i_\phi}}$ other than $C^m_{i_\phi}$. The angles of the vectors from $P^{j_\theta}_{C^m_{i_\phi}}$ to the $C^\mu_{\varepsilon_\Omega}$s will be in the set

$$\left\{\Omega = \frac{1}{2} + \theta - \frac{k}{q} \pmod 1\right\}, \quad k = \{1, 2, \ldots, q-1\}$$

Lastly, we will need to explain some peculiarities unique to this program. Whenever fractions appear (they will be represented by $\lambda$ and $\omega$), they should kept in lowest terms. Whenever "`if` $\exists\, C^\mu_{\varepsilon_\Omega}$" appears in the program, the program should check if there exists a vertex $C$ adjacent to the current juncture $P$ such that the vector from $P$ to $C$ makes an angle $\Omega$ with the horizontal. If so, the remainder of the indices ($\varepsilon$ and $\mu$) can be uniquely determined from $\Omega$, unless $\mu = n$, in which case an instruction to determine $\varepsilon$ is on the next line. Likewise, whenever "$\exists\, P^{j_\omega}_{C^m_{i_\phi}}$" appears in the program, the program should determine the value of $j$ (this will be unique) from the values of $i$, $\phi$, $m$, and $\omega$, which will always be known. Finally, "`stop`" ends the entire program; that is, by the time "`stop`" is encountered, all vertices $C^n$ will have been labeled.

**Labeling Algorithm**

```
begin label_components (p q n)
   l := 0
   α := (N(n))/2
   β := N(n)
   do label_1
end label_components

begin label_1
   for (λ = 0) to (1/2) by (1/2^⌈n/q⌉)
      if λ = 1/2
         do next_component_2 (P^{0_λ}_{C^0_{0_0}})
      else
         ∃ P^{j_λ}_{C^0_{0_0}}
         do next_component_1 (P^{j_λ}_{C^0_{0_0}})
      end if
   end for
```

```
end label_1

begin label_2
   for (λ = 1/2) to (1) by (1/(2^⌈n/q⌉))
      if λ = 1
         do next_component_4 (P^{0_0}_{C^0_{0_0}})
      else
         ∃ P^{j_λ}_{C^0_{0_0}}
         do next_component_3 (P^{j_λ}_{C^0_{0_0}})
      end if
   end for
end label_2

begin next_component_1 (P^{j_θ}_{C^m_{i_φ}})
   if j = 0 and if i = 0
      for (λ = 1/2 + θ − p/q (mod 1)) to (1/2 + θ + 1/q (mod 1)) by (−1/q)
         Ω := λ
         if ∃ C^μ_{ε_Ω}
            if μ = n
               label C^n_{l_Ω}
               l := l + 1
            else
               next_branch_point_1 (C^μ_{ε_Ω})
            end if
         end if
      end for
   else
      for (λ = 1) to (q − 1) by (1)
         Ω := 1/2 + θ − λ/q (mod 1)
         if ∃ C^μ_{ε_Ω}
            if μ = n
               label C^n_{l_Ω}
               l := l + 1
            else
               next_branch_point_1 (C^μ_{ε_Ω})
            end if
         end if
      end for
   end if
end next_component_1

begin next_component_2 (P^{j_θ}_{C^m_{i_φ}})
   if m = 0
```

```
      a := 0
   else
      a := (N(m))/2
   end if
   if j = 0 and if i = a
      for (λ = 1/2 + θ − 1/q (mod 1)) to (1/2 + θ − p/q (mod 1)) by (−1/q)
         Ω := λ
         if ∃ C^μ_{ε_Ω}
            if μ = n
               label C^n_{l_Ω}
               l := l + 1
               if l = α
                  do label_2
               end if
            else
               next_branch_point_2 (C^μ_{ε_Ω})
            end if
         end if
      end for
   else
      for (λ = 1) to (q − 1) by (1)
         Ω := 1/2 + θ − λ/q (mod 1)
         if ∃ C^μ_{ε_Ω}
            if μ = n
               label C^n_{l_Ω}
               l := l + 1
            else
               next_branch_point_2 (C^μ_{ε_Ω})
            end if
         end if
      end for
   end if
end next_component_2

begin next_component_3 (P^{j_θ}_{C^m_{i_φ}})
   if m = 0
      a := 0
   else
      a := (N(m))/2
   end if
   if j = 0 and if i = a
      for (λ = 1/2 + θ − p/q (mod 1)) to (1/2 + θ + 1/q (mod 1)) by (−1/q)
         Ω := λ
         if ∃ C^μ_{ε_Ω}
```

```
            if μ = n
               label C^n_{l_Ω}
               l := l + 1
            else
               next_branch_point_3 (C^μ_{ε_Ω})
            end if
         end if
      end for
   else
      for (λ = 1) to (q − 1) by (1)
         Ω := 1/2 + θ − λ/q (mod 1)
         if ∃ C^μ_{ε_Ω}
            if μ = n
               label C^n_{l_Ω}
               l := l + 1
            else
               next_branch_point_3 (C^μ_{ε_Ω})
            end if
         end if
      end for
   end if
end next_component_3

begin next_component_4 (P^{j_θ}_{C^m_{i_φ}})
   if j = 0 and if i = 0
      for (λ = 1/2 + θ − 1/q (mod 1)) to (1/2 + θ − p/q (mod 1)) by (−1/q)
         Ω := λ
         if ∃ C^μ_{ε_Ω}
            if μ = n
               label C^n_{l_Ω}
               l := l + 1
               if l = β
                  stop
               end if
            else
               next_branch_point_4 (C^μ_{ε_Ω})
            end if
         end if
      end for
   else
      for (λ = 1) to (q − 1) by (1)
         Ω := 1/2 + θ − λ/q (mod 1)
         if ∃ C^μ_{ε_Ω}
            if μ = n
               label C^n_{l_Ω}
```

```
            l := l + 1
         else
            next_branch_point_4 (C^μ_{ε_Ω})
         end if
      end if
   end for
end if
end next_component_4

begin next_branch_point_1 (C^m_{i_φ})
   if m = 0 and if i = 0
      for (λ = φ) to (1/2 + φ + 1/2^⌈(n−m)/q⌉ (mod 1)) by (−1/2^⌈(n−m)/q⌉)
         ω := λ
         ∃ P^{j_ω}_{C^m_{i_φ}}
         do next_component_1 (P^{j_ω}_{C^m_{i_φ}})
      end for
   else
      for (λ = 1) to (2^⌈(n−m)/q⌉ − 1) by (1)
         ω := 1/2 + φ − λ/2^⌈(n−m)/q⌉ (mod 1)
         ∃ P^{j_ω}_{C^m_{i_φ}}
         do next_component_1 (P^{j_ω}_{C^m_{i_φ}})
      end for
   end if
end next_branch_point_1

begin next_branch_point_2 (C^m_{i_φ})
   a := (N(m))/2
   if i = a
      for (λ = 1/2 + φ − 1/2^⌈(n−m)/q⌉ (mod 1)) to (φ) by (−1/2^⌈(n−m)/q⌉)
         ω := λ
         ∃ P^{j_ω}_{C^m_{i_φ}}
         do next_component_2 (P^{j_ω}_{C^m_{i_φ}})
      end for
   else
      for (λ = 1) to (2^⌈(n−m)/q⌉ − 1) by (1)
         ω := 1/2 + φ − λ/2^⌈(n−m)/q⌉ (mod 1)
         ∃ P^{j_ω}_{C^m_{i_φ}}
         do next_component_2 (P^{j_ω}_{C^m_{i_φ}})
      end for
```

```
   end if
end next_branch_point_2

begin next_branch_point_3 (C^m_{i_φ})
   a := (N(m))/2
   if i = a
      for (λ = φ) to (1/2 + φ + 1/2^⌈(n−m)/q⌉ (mod 1)) by (−1/2^⌈(n−m)/q⌉)
         ω := λ
         ∃ P^{j_ω}_{C^m_{i_φ}}
         do next_component_3 (P^{j_ω}_{C^m_{i_φ}})
      end for
   else
      for (λ = 1) to (2^⌈(n−m)/q⌉ − 1) by (1)
         ω := 1/2 + φ − λ/2^⌈(n−m)/q⌉ (mod 1)
         ∃ P^{j_ω}_{C^m_{i_φ}}
         do next_component_3 (P^{j_ω}_{C^m_{i_φ}})
      end for
   end if
end next_branch_point_3

begin next_branch_point_4 (C^m_{i_φ})
   if m = 0 and if i = 0
      for (λ = 1/2 + φ − 1/2^⌈(n−m)/q⌉ (mod 1)) to (φ) by (−1/2^⌈(n−m)/q⌉)
         ω := λ
         ∃ P^{j_ω}_{C^m_{i_φ}}
         do next_component_4 (P^{j_ω}_{C^m_{i_φ}})
      end for
   else
      for (λ = 1) to (2^⌈(n−m)/q⌉ − 1) by (1)
         ω := 1/2 + φ − λ/2^⌈(n−m)/q⌉ (mod 1)
         ∃ P^{j_ω}_{C^m_{i_φ}}
         do next_component_4 (P^{j_ω}_{C^m_{i_φ}})
      end for
   end if
end next_branch_point_4
```

## F.9 Acknowledgments

I would like to thank my advisor, Eric Bedford, for introducing me to complex dynamical systems. Further thanks goes to Kevin Pilgrim, who coordinated this REU. I also thank the Indiana University mathematics department for its enthusiastic support of my education.

Finally, I would like to acknowledge the sources of my graphics. Figures F.1 and F.5 are from [3]. Figures F.2 and F.3 were made using the OTIS applet on the Tomoki Kawahira's website; the applet can be found at http://www.math.nagoya-u.ac.jp/~kawahira/. Figure F.4 was taken from Derek Dreier's webpage, http://www.cs.ucr.edu/~ddreier/. The remainder of the figures were created using the drawing program Ipe, which can be downloaded from http://tclab.kaist.ac.kr/ipe/.

## Bibliography

1. Ahlfors, L. V. *Complex Analysis* (third edition), McGraw-Hill, 1979.

2. Beardon, A. F. *Iteration of Rational Functions*. Springer-Verlag, New York, 1991

3. Devaney, R. L. The Complex Dynamics of Quadratic Polynomials. In *Complex Dynamical Systems*. American Mathematical Society. (1994), 1-27.

4. Devaney, R. L. *An Introduction to Chaotic Dynamical Systems*, Second Edition. Addison-Wesley, Co., Reading, MA, 1992.

5. Gamelin, T. W. *Complex Analysis*. Springer, New York, 2001.

6. Milnor, J. *Dynamics in One Complex Variable*. Vieweg, Braunschweig/Wiesbaden, 1999.

# The Construction of a Complete, Bounded, Negatively Curved Surface in $\mathbb{R}^3$

Joseph Thurman
Vanderbilt University

Indiana University REU Summer 2009
Advisor: Chris Connell

# G.1 Introduction

Although the differential geometry of surfaces, and negatively curved surfaces in particular, is a well-studied field, there still remain a number of open problems and conjectures, even in the more basic case of surfaces in $\mathbb{R}^3$. In this paper, we examine one such question, the existence of a complete, bounded, negatively curved surface in $\mathbb{R}^3$. In this section, we give the background information necessary to state and understand the problem, which we begin investigating in Section 2. With the exception of Definition G.6, all definitions given in this section are taken from [2]. We start with the basic definition of a surface.

**Definition G.1.** A subset $S \subset \mathbb{R}^3$ is a *regular surface* if, for each $p \in S$, there exists a neighborhood $V \subset \mathbb{R}^3$ and a map $\mathbf{x} : U \to V \cap S$ of an open set $U \subset \mathbb{R}^2$ onto $V \cap S$ such that $\mathbf{x}$ is a diffeomorphism; that is, $\mathbf{x}$ is differentiable and has differentiable inverse $\mathbf{x}^{-1}$. The mapping $\mathbf{x}$ is called a *parameterization* of the surface at $p$, and the neighborhood $V \cap S$ is called a *coordinate neighborhood* at $p$.

*Notation.* We will use $(x, y, z)$ as our coordinates in $\mathbb{R}^3$, and $(u, v)$ as coordinates in $\mathbb{R}^2$. For convenience, partial derivatives will usually be written as subscripts throughout this paper. For example, $\mathbf{x}_u = \frac{\partial \mathbf{x}}{\partial u}$.

The conditions of the definition guarantee that a surface $S$ does not self-intersect, as a self-intersection would violate the bijectivity of $\mathbf{x}$. It also guarantees the existence of a tangent plane at every point, defined below.

**Definition G.2.** Let $\mathbf{x} : U \subset \mathbb{R}^2 \to \mathbb{R}^3$ be a parameterization of a regular surface $S$ at a point $p$. Let $q \in U$ such that $\mathbf{x}(q) = p$. Then the vectors $\left\{ \frac{\partial \mathbf{x}}{\partial u}(q), \frac{\partial \mathbf{x}}{\partial v}(q) \right\}$ form a basis for a 2-dimensional vector subspace called the *tangent plane* to $S$ at $p$ and denoted by $T_p(S)$.

In order to define the properties of completeness and negative curvature, we must first have some notion of distance on the surface. We can use the natural inner product of $\mathbb{R}^3$ to define an inner product on the tangent space at each point of $S$, $T_p(S)$. Let $\langle \cdot, \cdot \rangle_p$ denote the inner product on $T_p(S)$, and $\langle \cdot, \cdot \rangle$ denote the natural inner product of $\mathbb{R}^3$. Then, for $\mathbf{w}_1, \mathbf{w}_2 \in T_p(S)$, define $\langle \mathbf{w}_1, \mathbf{w}_2 \rangle_p := \langle \mathbf{w}_1, \mathbf{w}_2 \rangle$. This inner product on $T_p(S)$ leads naturally to the definition of the following quadratic form.

**Definition G.3.** Let $I_p : T_p(S) \to \mathbb{R}$ be defined as

$$I_p(\mathbf{w}) = \langle \mathbf{w}, \mathbf{w} \rangle_p = |\mathbf{w}|^2 \geq 0 \tag{G.1}$$

This quadratic form is called the *first fundamental form* of $S$ at $p$.

This form can be given in terms of $(u, v)$ for the basis $\{\mathbf{x}_u, \mathbf{x}_v\}$ of $T_p(S)$, where $\mathbf{x}(u, v) = p$. Then

$$I_p(\mathbf{w}) = \mathbf{w}^T \begin{pmatrix} E & F \\ F & G \end{pmatrix} \mathbf{w}$$

with

$$\begin{array}{rcl} E(u,v) & = & \langle \mathbf{x}_u, \mathbf{x}_u \rangle \\ F(u,v) & = & \langle \mathbf{x}_u, \mathbf{x}_v \rangle \\ G(u,v) & = & \langle \mathbf{x}_v, \mathbf{x}_v \rangle \end{array} . \tag{G.2}$$

$E, F$, and $G$ are called the *coefficients of the first fundamental form.*

The first fundamental form can be used to define arc length on a surface. Let $\alpha(t) : (0, t_0) \subset \mathbb{R} \to S$ be the equation of a parameterized curve on $S$. Then the arc length of the curve from $t = 0$ to $t = t_0$, denoted by $\ell(\alpha)$, is given by

$$\ell(\alpha) = \int_0^{t_0} \sqrt{I(\alpha'(t))}dt.$$

We can therefore define distance on a surface as follows.

**Definition G.4.** Let $S$ be a regular surface with first fundamental form $I_p$ at point $p$, and let $p, q \in S$ be given. Then the *distance* from $p$ to $q$ on $S$, $d(p, q)$, is

$$d(p, q) = \inf_A \ell(\alpha) \tag{G.3}$$

where $A$ is the set of all $C^1$ curves $\alpha$ on $S$ from $p$ to $q$.

We use this definition of distance in the next definition, when we require that a sequence of points on the surface is Cauchy with respect to the surface distance $d(p, q)$.

**Definition G.5.** A surface $S \subset \mathbb{R}^3$ is *intrinsically complete* if, for every Cauchy sequence $\{p_i\}_{i=1}^{\infty}$ with $p_i \in S$ and limit point $p \in \mathbb{R}^3$, we have $p \in S$.

**Definition G.6.** A surface $S$ is *bounded* in $\mathbb{R}^3$ if there exists a constant $M \in \mathbb{R}$ such that $S$ is completely contained in the open ball in $\mathbb{R}^3$ of radius $M$ centered at the origin.

Note that a negatively curved, bounded, noncompact surface cannot be extrinsically complete, that is, complete with respect to the usual Euclidean distance in $\mathbb{R}^3$, as such a surface would necessarily have a Cauchy sequence of points that converges outside the boundary. The surfaces we consider are noncompact, as discussed at the start of Section G.2.

Now that we have defined two of the properties of the desired surface, we move on to the definition of Gauss curvature. First, we place one more requirement on the type of surface we consider.

**Definition G.7.** A regular surface $S$ is called *orientable* if it is possible to cover it with a family of coordinate neighborhoods in such a way that if a point $p \in S$ belongs to two neighborhoods of this family, then the change of coordinates has positive Jacobian at $p$. A choice of such a family is called an *orientation* of $S$. If such a choice is not possible, the surface is called *nonorientable*.

*Notation.* For the reminder of this paper, we use *surface* when we mean *regular, orientable surface*.

**Proposition G.1.1.** *Let $S$ be a surface with parameterization $\mathbf{x} : U \subset \mathbb{R}^2 \to S$. Then for each $q \in \mathbf{x}(U)$, we can choose a unit normal vector at $q$ by*

$$N(q) = \frac{\mathbf{x}_u \times \mathbf{x}_v}{|\mathbf{x}_u \times \mathbf{x}_v|}(q), \tag{G.4}$$

*where $\times$ denotes the usual cross product in $\mathbb{R}^3$.*

*Proof.* The proof of Proposition G.1.1 is given in [2]. Q.E.D.

Using the definition of the normal field, we may now define another quadratic form on $S$, the so-called *second fundamental form*.

**Definition G.8.** The *second fundamental form*, denoted by $II_p(\mathbf{w}) : T_p(S) \to \mathbb{R}$ is given by

$$II_p(\mathbf{w}) = \mathbf{w}^T \begin{pmatrix} e & f \\ f & g \end{pmatrix} \mathbf{w}, \tag{G.5}$$

where

$$\begin{aligned} e(u,v) &= \langle N, \mathbf{x}_{uu} \rangle \\ f(u,v) &= \langle N, \mathbf{x}_{uv} \rangle \\ g(u,v) &= \langle N, \mathbf{x}_{vv} \rangle \end{aligned} \tag{G.6}$$

With these fundamental forms defined, we are now able to give the definition of one of the most important concepts in differential geometry, that of Gaussian curvature of a surface. There are many equivalent definitions of Gaussian curvature. It can be defined in terms of the differential of the normal map given in equation G.1.1, in terms of the curvatures of arcs through points on a surface, or in terms of the first and second fundamental forms. We use the last definition, as this definition leads more easily to the computation of the curvature for a surface.

**Definition G.9.** Let $S \subset \mathbb{R}^3$ be a surface with parameterization $\mathbf{x} : U \subset \mathbb{R}^2 \to \mathbb{R}^3$, and let $E, F, G, e, f$, and $g$ be as defined in (G.2) and (G.6). Let $q \in U$ be given such that $\mathbf{x}(q) = p$. Then the *Gaussian curvature* of $S$ at point $p$, denoted by $K(p)$, is given by

$$K(p) = \frac{eg - f^2}{EG - F^2}, \tag{G.7}$$

with each function evaluated at $q$.

Note that $EG - F^2 > 0$, so the sign of the curvature is always determined by the coefficients of the second fundamental form.

Just as Gaussian curvature can be defined in many ways, it also has a number of useful geometric interpretations. The following proposition gives a property of surfaces around points of negative curvature, which will be useful later. The proof can be found in [2].

**Proposition G.1.2.** *Let $p \in S$ be a point on a surface $S$ such that $K(p) < 0$. Then for any neighborhood $V$ around $p$, there are points of $S \subset V$ on each side of $T_p(S)$.*

## G.2 Previous Results

Armed with this basic understanding of surfaces, we now discuss some past results that will serve as motivation for the problem considered in this paper. In particular, we present theorems from Efimov, Connell, and Ullman that place restrictions on the end behavior of complete, negatively curved surfaces. We also briefly review Rozendorn's construction of a bounded, complete surface with nonpositive curvature. It will be the goal in later sections of this paper to examine how Rozendorn's surface can be modified to create a complete, bounded surface with negative curvature.

We are first motivated by the observation that a compact, negatively curved surface cannot exist in $\mathbb{R}^3$. If such a surface existed, it would be possible to take a flat plane and, beginning at infinity, move it toward the surface until it is just tangent to the surface. However, by Proposition G.1.2, the surface would exist on both sides of the tangent plane at this point, contradicting the assumption that this is first point of contact between the surface and the plane.

Since the construction of a compact, negatively curved surface is not possible, we consider the construction of negatively curved surfaces with weaker properties than compactness. This leads naturally to the problem under consideration. Theorems from Efimov, Connell, and Ullman give further insight into the possibility of constructing such a surface. We begin with Efimov's Theorem.

**Theorem G.2.1** ([4])**.** *Let $S$ be a complete surface in $\mathbb{R}^3$, with $K(p) \leq 0$ for all $p \in S$. Then the least upper bound for $K$ is* 0*.*

Essentially, this theorem states that the Gaussian curvature of a complete surface cannot be bounded away from zero. The curvature always approaches zero down at least one end of a surface.

Each "end" of a complete surface is topologically equivalent to a punctured disc, i.e., an annulus, that is unbounded in the intrinsic metric, the distance given in definition G.4. These ends can be classified into the following types.

**Definition G.10.** Let $\Sigma$ denote an end of a surface $S$. We call a simple closed curve $\Gamma$ on $\Sigma$ a *belt curve* if it is homotopic to the boundary of the closure of $\Sigma$. Then $\Sigma$ is called a *horn end* if there is no belt curve of shortest length on the closure of $\Sigma$. Otherwise, $\Sigma$ is called a *bowl end*. Further, a horn end is called a *cusp* if the infimum of the lengths of the belt curves is 0.

With these definitions, we can give the following theorem from Connell and Ullman.

**Theorem G.2.2** ([3])**.** *Given nonnegative integers $n_c$, $n_b$, and $g$, with $n_b > 0$, there exists a negatively curved $C^\infty$ surface with genus $g$ in $\mathbb{R}^3$ with $n_c$ cusp ends and $n_b$ bowl ends.*

Theorem G.2.2 essentially states that there are many types of surfaces that can be embedded in $\mathbb{R}^3$ so long as they have at least one bowl end. This raises an interesting secondary problem also considered throughout the project. Are there any negatively curved, complete surfaces in $\mathbb{R}^3$ with only cusp ends? Currently, only one such surface is known. It has four ends and is due to Vaigant [1]. This lead us to the following conjecture.

**Conjecture G.2.3.** *Let $n$ be an integer such that $n \geq 4$. Then there exists a complete, negatively curved surface in $\mathbb{R}^3$ with $n$ cusp ends.*

*Remark* G.1. A theorem from Osserman [7] requires that a bounded subset $R$ of a nonpositively curved surface be a contained in the convex hull of $\partial R$. On a surface $S$ with one, two, or three cusp ends, the convex hull of the ends approaches a point, a line, or a plane, since each end approaches a point. Since the surface is not contained in any plane, a nonpositively curved surface with $\leq 3$ cusp ends is not possible. We therefore require that $n \geq 4$.

*Remark* G.2. The proof of this conjecture would actually follow as a special case of the overall question of this paper. We discuss attempts to prove this conjecture in Section G.4.

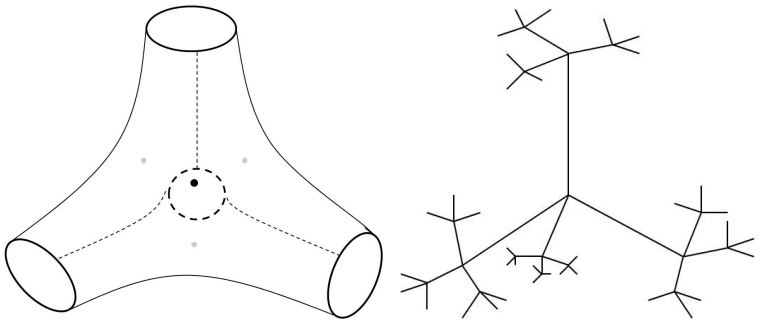
The following result is the most relevant to the main problem considered in this paper.

**Theorem G.2.4.** *There exists a complete, bounded, nonpositively curved surface in $\mathbb{R}^3$.*

*Proof.* This result was proved by Rozendorn in [8], in which he gave a method to construct such a surface. Q.E.D.

We use Rozendorn's surface as the starting point in our investigation. Specifically, it is our goal to find a modification of Rozendorn's surface to create a negatively curved surface. The following details of Rozendorn's construction are therefore relevant.

- Rozendorn's construction begins with a 4-horned sphere of non-positive curvature. See Figure G.1(a) on page G-5 for an illustration of the building block. Note that this figure has the ends of the building block removed. At each boundary circle, a tapering tube extending out to infinity is attached. This starting surface has negative curvature everywhere except at 4 discrete points of zero curvature, marked in the figure.
- At each stage of the construction, a new building block surface is attached to the end of the surface. Each new surface attached is modified through a linear transformation, which shrinks and bends the surface, but does not affect the sign of the curvature. Infinitely many building blocks are attached in this way. The surfaces are joined together by gluing the circle boundaries shown in Figure G.1(a) together with negatively curved tubes. It therefore resembles a thickening of an an infinite tree with 4 branches at each vertex. Three levels of such a tree are shown in Figure G.1(b).
- The transformations are such that, as the construction continues, the surface remains bounded inside a ball in $\mathbb{R}^3$, but the path length along the surface is unbounded. The surface is therefore complete.
- Rozendorn's surface therefore has infinitely many ends, and infinitely many discrete points of zero curvature.



(a) Rozendorn's Building Block (Ends Removed)

(b) Tree Construction

Figure G.1: Roxendorn's Construction

Our main goal is to find a modification of this surface that removes the points of zero curvature. Such a modification would therefore yield a complete, bounded surface of negative curvature in $\mathbb{R}^3$. We began the search for such a modification by examining the topological properties of such a surface.

## G.3 Topology of Negatively Curved Surfaces

The simplest and most desirable modification would be a local deformation of Rozendorn's surface around each point of zero curvature. We could hope to find a transformation that would simply bend the surface from zero to negative curvature. Applied to every zero point on Rozendorn's surface, such a deformation would quickly give the existence of the desired surface.

Unfortunately, the topological properties of negatively curved surfaces make such a local transformation impossible. In order to give the reason why, we require the following definitions and theorems.

**Definition G.11.** Let $\vec{v} : S \to \mathbb{R}^3$ be a differentiable vector field on a surface $S$. Then $p_i \in S$ is called a *critical point* of the vector field $\vec{v}$ if $\vec{v}(p_i) = \mathbf{0}$.

**Definition G.12.** Let $\vec{v} : S \to \mathbb{R}^3$ be a differentiable vector field on a surface $S$, and let $p \in S$ be a critical point of $\vec{v}$. Choose a coordinate neighborhood around $p$, $V \cap S$, such that $p$ is the only critical point of the field $\vec{v}$ in $V \cap S$. Use the inverse parameterization function $\mathbf{x}^{-1}$ to project the vector field $\vec{v}$ onto $\mathbb{R}^2$, creating a vector field $\vec{v}^* : U \subset \mathbb{R}^2 \to \mathbb{R}^2$, where $\mathbf{x}^{-1}(V \cap S) = U$. Let $\Gamma$ be a simple closed curve in $U$ surrounding the point $q$ with $\mathbf{x}(q) = p$. Then the *index* of the point $p$, denoted $ind(p)$, is the winding number of the vector field $\vec{v}^*$ along $\Gamma$.

*Remark* G.3. Since the vector field $\vec{v}$ is differentiable, and therefore continuous, the index of a critical point must be an integer.

*Remark* G.4. Although the above is defined for vector fields, an analogous definition is possible for line fields. In the case of line fields, the index of a critical point is an integer multiple of $\frac{1}{2}$.

Finally, we give a simple definition of the Euler characteristic, a basic notion from topology. Although there are other, more precise definitions, we give the following more intuitive definition, which will be better suited to our situation.

**Definition G.13.** Let $S$ be a closed, orientable surface. Then the *Euler characteristic* of $S$, denoted by $\chi(S)$ is given by

$$\chi(S) = 2 - 2g,$$

where $g$ is the genus, or "number of holes," of the surface.

*Remark* G.5. Although the genus of a surface also has a precise definition, we use the intuitive notion of the number of "holes" in the surface. For example, the sphere has no holes, and therefore has genus 0, while the torus has one hole, or genus 1, and so on.

With these definitions, we can now state an important and extremely useful theorem connecting the two ideas, the classical Poincaré-Hopf Theorem.

**Theorem G.3.1.** *Let $S$ be a closed, orientable surface, and let $\vec{v}$ be a differentiable vector field defined on $S$, with critical points $p_i \in P$, where $P$ is some index set. Then,*

$$\sum_P ind(p_i) = \chi(S) \tag{G.8}$$

*Remark* G.6. Although we give this theorem here for a vector field on a surface, the conclusion of the theorem also applies, unchanged, to the case of a line field defined on the surface. (See Remark G.4.)

The following proposition allows us to apply Theorem G.3.1 to nonpositively curved surfaces.

**Proposition G.3.2.** *Let $S$ be a surface of nonpositive curvature. Then we can define a differentiable line field $L$ on $S$ as follows: Let $p \in S$ be given. Let $\mathbf{k}_1$ and $\mathbf{k}_2$ be the principal curvature directions, the vectors in $T_p(S)$ where $II_p(S)$ reaches its maximum and minimum, respectively. Then*

$$L(p) = \begin{cases} 0 & K(p) = 0 \\ \{-c, c\}\,\mathbf{w} : II_p(\mathbf{w}) = 0 & K(p) < 0 \end{cases}, \tag{G.9}$$

*where* $\mathbf{w} \in T_p(S)$ *is chosen so that when* $\{\mathbf{k}_1, \mathbf{k}_2\}$ *is positively ordered with respect to the normal field on* $S$*, then* $\{\mathbf{k}_1, \mathbf{w}\}$ *and* $\{\mathbf{w}, \mathbf{k}_2\}$ *are positively ordered as well, and* $\{-c, c\}$ *is a interval of scalars such that*

$$c = \prod_{\vec{v}(q_i)=0} d^2(p, q_i)$$

.

*Proof.* This follows from the properties of the asymptotic directions as given in [2] Q.E.D.

*Remark* G.7. The direction $\mathbf{w}$ is called an *asymptotic direction,* and is the direction where the so called "normal curvature" of $S$, which we have not defined, is zero. We multiply $\mathbf{w}$ by the scalars in $\{-c, c, \}$ to turn the vector into a line extending in both directions. $c$ is chosen so that the length of the lines vanish as the points reach a critical point. Additionally, there are two such asymptotic directions at a give point, bisecting the principal directions. The conditions on the ordering of the principal directions with $\mathbf{w}$ guarantees that we choose only one of these asymptotic direction and that our choice gives a continuous field.

We can now use Proposition G.3.2 and Theorem G.3.1 to examine the topology of Rozendorn's building block surface. We can cut off each cusp end and cap each end with a hemisphere. This creates a closed, complete surface of non-positive curvature with 8 critical points - the four points of zero curvature originally on the surface, and four new critical points at each capped-off end. This surface is topologically equivalent to a sphere, and therefore has Euler characteristic 2. We can calculate the index at each critical point. The index on each of the original zero curvature points is $-\frac{1}{2}$, while the index on each of the new critical points is 1. Therefore, the sum of the indices is $4 + 4 \cdot -\frac{1}{2} = 2$, as required by Theorem G.3.1.

The points of zero curvature on Rozendorn's building block are therefore topologically necessary. A local deformation that removed those points without modifying the indices on each end would give a surface with $\chi(S) = 4$, which is not possible. Rather, these points can only be removed by a global transformation that also changes the indices of the ends.

Index restrictions are also relevant when joining the ends of two surfaces. In Rozendorn's construction, each pair of ends are joined by a negatively curved, tubular surface. This joining section can be capped on each end, giving 2 critical points on the joining piece that have indices identical to the indices of the joined ends. This capped tube is also homeomorphic to a sphere, with $\chi = 2$, so the sum of the indices must also be 2. Thus, we say that two ends of index $a$ and $b$ are *complementary,* and therefore able to be joined together, if $a + b = 2$.

**Definition G.14.** A surface $S$ is called *self-perpetuating* if for every end of index $a \neq 1$ there is another end of the surface with index $2 - a$.

*Remark* G.8. We call such a surface self-perpetuating because each end could be joined to another copy of the surface without violating the index conditions on the surface. We consider ends with index $\neq 1$ because an end with index 1 can always be joined to a copy of itself. Note that Rozendorn's building block surface is self-perpetuating. Thus, many copies of a self-perpetuating surface could possibly be joined together in a construction method similar to Rozendorn's method.

We investigated the possibility of constructing other building block surfaces that are self-perpetuating, but also negatively curved. The sum of the indices on the ends of such a surface is therefore required to be 2, as there can be no critical points elsewhere on the surface. For any number of ends greater than 3, such a combination of index values is in fact possible. We found the following index combinations are possible for surfaces with $n = 4, 5, 6$.

*Example* G.1. For $n = 4$, the only self-perpetuating surfaces of genus 0 possible have indices of $\{1, -1, -1, 3\}$ or $\{0, 0, 0, 2\}$ on the ends. For genus $g$, the only ends possible have indices of $\{-g, -g, -g, 2 + g\}$.

*Example* G.2. For $n = 5$ and genus 0, there are an infinite number of possible index combinations on the ends of the surface. The combination $\{-2, -2, 4, a, 2 - a\}$ is self-perpetuating for any index $a$. $\{0, 0, 0, 0, 2\}$ is also self-perpetuating. For genus $g$,
$\{-2 - g, 4 + 2g, -2 - 2g, a, 2 - a\}$ is self-perpetuating for any index $a$, and $\{-\frac{2g}{3}, -\frac{2g}{3}, -\frac{2g}{3}, -\frac{2g}{3}, 2 - \frac{2g}{3}\}$ is self-perpetuating for any $g$ divisible by 3.

*Example* G.3. For $n = 6$ and genus 0, there are an infinite number of possible index combinations on the ends of surface. The combination $\{-3, -3, 1, 5, a, 2 - a\}$ is self-perpetuating for any index $a$. Other possible index combinations include $\{-1, -1, -1, -1, 3, 3\}$,
$\{-2, -2, 0, 0, 2, 4\}$, and $\{-4, -4, 0, 2, 2, 6\}$. Other possible index combinations continue to have increasing index on each end, making such a surface even harder to construct.

Even with a self-perpetuating, negatively-curved surface, it is not guaranteed that copies of the surface could be joined together while maintaining their negative curvature. Rozendorn's method of joining the building block surfaces with negatively curved tubes require that the cross-sections of the ends being joined together be convex curves. The following proposition gives a relationship between the index of an end and the convexity of its cross section.

**Proposition G.3.3.** *Let $S$ be a nonpositively curved, complete surface with a finite set of discrete zero curvature points. Consider an end of this surface with index $a$. Let $\Gamma$ be the closed, continuous curve given by the intersection of the end with a plane. Let $n$ be the number of points of inflection of $\Gamma$, that is, the number of points where $\Gamma$ changes convexity. Then,*

$$4\,|1 - a| = n \tag{G.10}$$

*Proof.* By inspection we see that if an end has index $a$, then the asymptotic line field must be tangent to $\Gamma$ at $2\,|1 - a|$ points. Since there are two asymptotic line fields on any surface (see Remark G.7), there are a total of $4\,|1 - a|$ points on $\Gamma$ where an asymptotic direction is tangent to $\Gamma$. We must therefore show that these points, and only these points, are inflection points for the curve.

Let $\Gamma$ be parameterized by arc length, so that the curve is given by $\Gamma(s) : \mathbb{R} \to \mathbb{R}^3$. Then, since $\Gamma$ is a planar curve, its inflection points are exactly those points where $\Gamma''(s)$ changes sign. A geometric interpretation of $II_p$ given in [2] shows that $\Gamma''(s)$ changes sign only where $II_p = 0$, that is, when the asymptotic line field is tangent to $\Gamma$. Q.E.D.

**Corollary G.3.4.** *$\Gamma$ is a convex curve if and only if $a = 1$.*

Thus, even if we could construct a negatively curved surface with self-perpetuating ends, we would also have to devise a method to join the pieces together while maintaining negative curvature. Considering the difficulty of these tasks, we decided to approach the problem from a different direction.

## G.4 Point Sliding and the Monge-Ampère Equation

As shown in Section G.3, a local deformation of Rozendorn's surface at each point of zero curvature is not sufficient to create an everywhere negatively curved surface. In this section we attempt to find a global deformation instead. However, our desired global deformation will actually consist of an infinite number of repeated, local deformations.

Specifically, we note that although an index point cannot be simply created or destroyed, two index points can usually be combined together. For example, two critical points on a sphere, each of index 1, could be moved together and combined into a single critical point of index 2. We attempt to find a similar sliding procedure on Rozendorn's surface. However, instead of combining the points of zero curvature, we hope to move each point out an end and to infinity along the surface, still maintaining negative curvature elsewhere on the surface. Such a procedure will change the index at each end, and therefore be a global transformation of the surface.

If such a sliding procedure could be constructed on Rozendorn's surface, we also note that Conjecture G.2.3 would follow quite easily. After constructing a surface with $n$ ends and non-positive curvature, the points of zero curvature could simply be moved down each end, deforming the surface into a new complete, negatively curved surface. In fact, although Vaigant's surface has been given explicitly, it could likely be constructed through such a procedure. Beginning with Rozendorn's building block, a four-horned sphere with four points of nonzero curvature, each point could be moved along an end to give a negatively curved surface the would likely be extremely similar to Vaigant's.

Finally, we note that a such a sliding procedure, although it ultimately results in a global deformation of the surface, could be accomplished by repeating a local sliding procedure. If a zero curvature point can be moved just a small distance from its original location without creating positive curvature, this small sliding procedure could be repeated to drag the point to infinity.

With this intuitive understanding of the desired sliding procedure, we formalize this problem, starting with the following basic theorem from [2].

**Theorem G.4.1.** *Let $S$ surface with $p \in S$ given. Then there exists an open set $V \subset \mathbb{R}^3$ with $p \in V$ such that $S \cap V$ can be parameterized as a graph. That is, there exists a function $\mathbf{f} : U \subset \mathbb{R}^2 \to \mathbb{R}$ such that the coordinates of a point $q \in S \cap V$ can be expressed as $(u, v, \mathbf{f}(u, v))$*

We now consider a nonpositvely curved surface $S$ with a point of zero curvature $p$. Using Proposition G.4.1, we parameterize the surface on a neighborhood $V$ around $p$ as a graph, and associating the point $p$ with the origin of $\mathbb{R}^2$. Let $z : U \subset \mathbb{R}^2 \to \mathbb{R}$ be the graph function of the surface, and let $K(u, v)$ denote the Gaussian curvature at the point $(u, v, z(u, v))$. Then, as given, $K(0, 0) = 0$. Let us call the direction in which we want to move the zero curvature point the $+u$-direction, and the distance we wish to move the point $a$. That is, if $K^*$ is the curvature function for the new graph, $K^*(a, 0) = 0$. At every other point in $U$, we desire $K^* < 0$. Finally, we must ensure that this transformation does not affect the rest of the surface. Therefore, using $z^*$ to denote the graph function for the new surface, we desire $z^*|_{\partial U} = z|_{\partial U}$, $z^*_{\vec{n}}|_{\partial U} = z_{\vec{n}}|_{\partial U}$, and $K^*|_{\partial U} = K|_{\partial U}$, where $z^*_{\vec{n}}$ denotes the derivative in the direction of the normal.

Our goal is therefore to find a new graph function $z^*$ on the same domain that gives a surface with the curvature problems stated above. Using the definitions of Gaussian curvature and the first and second fundamental forms given in the introduction, this problem is equivalent to solving the Dirichlet problem for the Monge-Ampère equation. That is, we attempt to find a function $z^*$ such that

$$(z^*_{uu} z^*_{vv} - {z^*_{uv}}^2) = K(u, v) \left(1 + {z^*_u}^2 + {z^*_v}^2)\right)^2, \tag{G.11}$$

with,

$$K^* \leq 0,\ K^* = 0 \text{ at only one point in } U,\ z^*|_{\partial U} = z|_{\partial U},\ z^*_{\vec{n}}|_{\partial U} = z_{\vec{n}}|_{\partial U}, \text{ and } K^*|_{\partial U} = K|_{\partial U}.$$

Monge-Ampère equations have many applications, and have therefore been extensively studied. However, much of this study concerns the case of the elliptic Monge-Ampère equation, which occurs when $K > 0$. In our case, $K \leq 0$, the equation is hyperbolic, and therefore far more difficult to solve.

We performed a thorough search of the literature to find methods to solve the hyperbolic Monge-Ampère equation, ultimately focusing on two results given by Han and Hong in [5] and Khuri in [6]. Both of these results use similar methods to solve the equation. They begin by linearizing the equation and obtaining weak solutions to the new linear equation. Using various estimates on the derivatives of the equation, they modify the weak solutions to give regular solutions to the linear equation. Then, using a Nash-Moser iteration procedure, these regular solutions to the linear problem are used to find a solution to the original equation.

Both of these papers, however, address a problem different from our own. They are not generally concerned with the domain over which the solution is obtained, and allow this domain to become arbitrarily small to simplify the estimates they use to solve the equation. In general, the larger the curvature and its derivatives are in a certain domain, the smaller the subset of the domain on which the solution can be found.

In our problem, we attempt to find a solution over a given domain, and therefore we must ensure that the curvature and its derivatives are sufficiently bounded to guarantee that the solution domain is large enough to cover our given domain. In general, these curvature bounds are to restrictive to allow us to find solutions for most surfaces, as they can place very small bounds on the curvature and its derivatives to orders greater than 10.

Additionally, the very act of moving these points of zero curvature increases the curvature and its derivatives on the domain considered. Thus, even if the point could be moved once a certain distance, the increase in curvature caused by that move could mean that the next move could only move a smaller distance. Thus, instead of being able to draw the points out to infinity, each zero curvature point could end up being bounded inside a region of prohibitively high curvature.

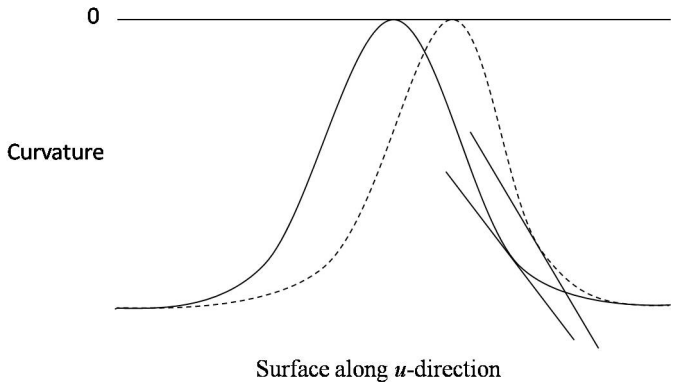



Figure G.2: Changing Derivative of Curvature

Figure G.2 illustrates why the increase in the derivative of the curvature is necessary. The figure is a graph of the Gaussian curvature of a nonpositively curved surface along the curve $v = 0$, where the zero point is moved in the $u$-direction. The solid curve represents the original curvature of the surface, and the dashed curve represents the curvature after the zero point is moved. Because the dashed curve must coincide with the original curve outside of the domain on which the point is being moved, the surface must reach the same level of negative curvature in a shorter distance, meaning the derivative of the curvature must increase.

Unable to find a solution to the Monge-Ampère in the general case, and therefore unable to find a general point-sliding method, we began investigating an explicit moving procedure for a special case, the so-called

"monkey saddle," given as the graph

$$z(u, v) = u^3 - 3uv^2. \tag{G.12}$$

The monkey saddle is shown in Figure G.3.

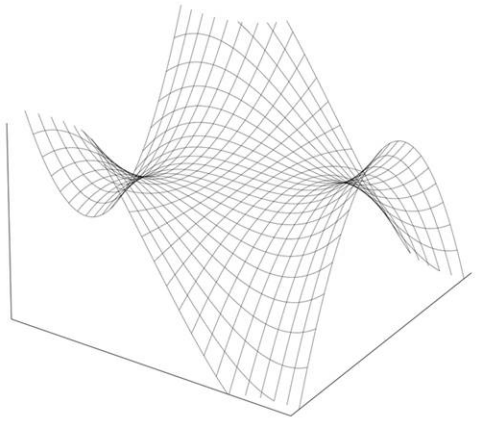


Figure G.3: The Monkey Saddle

The monkey saddle is negatively curved except at $(0, 0, 0)$, where $K = 0$, and is in fact very similar to Rozendorn's surface in the neighborhood of a zero curvature point. We then attempt to find a modification of this surface in a small neighborhood around the origin that moves the zero curvature point without changing the rest of the function. Such a new function could be constructed of the form $z(u + b_1(u)b_1(v), v)$, where $z$ is the equation of the monkey saddle given in G.12 and $b_1$, $b_2$ are "bump functions," more precisely functions of compact support. These bump functions would be zero outside the domain on which the point is moved, and thus would not modify the rest of the surface. However, they would have non-zero values within the boundary, therefore modifying the monkey saddle and moving the point of zero curvature. Such bump functions must be carefully constructed to ensure that the modified part of the surface still smoothly matches the surface on the boundary of the domain while still maintaining negative curvature on the surface. We have not yet been able to give an explicit form for these bump functions. Still, even with the proper bump functions, this process would only be enough to move the zero curvature once, and only in the $u$-direction. Further investigation is required to find a point sliding procedure in any direction, or to be able to move the point repeatedly.

## G.5 Conclusions and Further Research

Although we have been unable to solve the main problem considered in this paper, our research suggests a number of possible avenues for further research. We hope that continued work on the monkey saddle could yield a general sliding procedure on that surface. Such a procedure could possibly be generalized to give an explicit sliding procedure on some of the non-positively curved surfaces we have considered.

Examples G.1, G.2, and G.3 give another possible method to construct a bounded, complete, negatively-curved surface. Ideally, we could find explicit equations for a negatively-curved surface with indices given

in one of those examples. Then we must construct a gluing method that smoothly joins non-convex ends without creating positive curvature. Combining those possible results would yield the desired surface.

Of course, it may also be possible that such a surface cannot be constructed. Further investigations of negatively curved surfaces and their topology could yield such a result.

## G.6 Acknowledgments

I would like to thank my advisor, Chris Connell, for his guidance throughout the project. I would also like the thank the faculty and staff of the Indiana University Mathematics Department, especially Kevin Pilgrim and Mandie McCarty.

## Bibliography

1. Yu.D. Burago and S.Z. Shefel. *Geometry III*,Chapter 1, Encyclopedia of Mathematical Sciences, V. A, Zalgaller (ed.), **48**, Springer-Verlag, Berlin (1992). Trans. E. Primrose

2. M. do Carmo. *Differential Geometry of Curves and Surfaces*, Prentice Hall, Englewood Cliffs, NJ (1976).

3. C. Connell and J. Ullman. *Ends of Negatively Curved Surfaces in Euclidean Space*, to appear.

4. N.V. Efimov. *Appearance of Singularities on Surface of Negative Curvature*, Trans. J. Danskin, Amer. Math. Soc. Transl. **66**, 154-190 (1968).

5. Q.Han and J-X. Hong. *Isometric Embedding of Riemannian Manifolds in Euclidean Spaces*, American Mathematical Society, Providence, RI (2006).

6. M. A. Khuri. *Local Solvability of Degenerate Monge-Ampère Equations and Applications to Geometry*, Electron. J. Differential Equations. **65**, 1-37 (2007).

7. R. Osserman. *The Convex Hull Property of Immersed Manifolds*, J. Differential Geom. **6**, 267-270 (1971-2).

8. E.R. Rozendorn. *The Construction of a Bounded, Complete Surface of Nonpositive Curvature*, Uspehi Mat. Nauk **16** (1961), no. 2 (98), 149-156. (Russian)

# Simple, Closed Geodesics on Polyhedra

Leah Wolberg
Bowdoin College

Indiana University REU Summer 2009
Advisor: Matthias Weber

# H.1 Introduction

We studied *geodesics*—simple, locally straight curves—on surfaces with polygonal metrics.

Fuchs and Fuchs [2] studied geodesics, both simple and non-simple, on the Platonic solids. Their results only covered the regular tetrahedron and cube in detail, however, which makes their paper an excellent starting point.

This paper begins with a discussion of curvature and how it is measured on and relates to the polygonal metric. We then set up the background for our methods: the developing map (Section H.1.2), the dual graph (Section H.2), and quotient spaces (Section H.3). It continues with sections on specific polyhedra we worked on, such as zonohedra (Section H.4), the rhombic dodecahedron (Section H.4.2), and $\{6, 4|4\}$, an infinite skew polyhedron which has four hexagons meeting at every vertex (Section **??**).

## H.1.1 Curvature

Curvature of a surface falls into one of two categories: *extrinsic* and *intrinsic*. The easiest way to understand the distinction is to imagine a Flatlander that lives on the given surface and ask whether she could detect the curvature. If so, the curvature is intrinsic; if not, it is extrinsic.

One type of extrinsic curvature which will come up quite often in the study of polyhedra is the curvature represented by edges. A Flatlander passing across an edge would remain unchanged from her own perspective. At least locally, it is possible to flatten out the edge without changing the metric, just as it is possible to unfold a piece of paper without changing the intrinsic distance between two points on it.

Similarly, a common type of intrinsic curvature we will need to deal with is manifested as the corners of polyhedra, which are locally equivalent to a cone and known as *cone points*. They are single points with highly concentrated curvature. They are best envisioned as a paper wedge with an angle $\alpha$ on which the two straight edges have been identified. $\alpha$ is called the *cone angle*.

**Definition H.1.** The curvature $\kappa_{p_i}$ at a cone point $p_i$ with cone angle $\alpha$ is $\kappa_{p_i} = 2\pi - \alpha$.

For simplicity, we will avoid discussing geodesics that pass through cone points; they can be regarded as limit cases. The cone points of a polyhedron act like punctures: a Flatlander would have to avoid them. Two continuous curves are said to be *homotopic* if one can be continuously deformed into the other. If one of the curves encloses a cone point, so must the other, since there is no way to pass over a cone point by continuous deformation.

A geodesic is a curve of constant curvature since a Flatlander on a polyhedron would see it as a straight line. In order to be intrinsically straight, when a geodesic crosses an edge the four angles between the two lines must behave similarly to the angles of two intersecting lines in the plane. In other words (moving counterclockwise around the intersection) the sum of the angle between the line segment in the first face and the edge and the angle formed by the edge and the line segment in the second face must be $\pi$.

## H.1.2 The Polyhedron Developing Map

The developing map of a polyhedron is useful in determining the homotopy class of a given curve. Take a curve, not necessarily closed, on a polyhedron. Starting at one end of the curve and following it, roll the polyhedron on a plane so that the curve always touches the plane.

The three lines shown in FigureH.1 are homotopically equivalent.

**Observation H.1.1.** *If two homotopic curves on a polyhedron connect the same points, the developed image of both curves connects the points.*

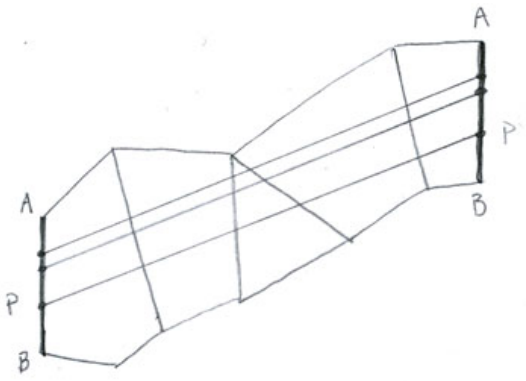



Figure H.1: Polyhedron developing map

Constructing the developing map is not an efficient process, but it clearly distinguishes homotopy classes which contain geodesics from those which do not:

**Observation H.1.2.** *Take the developing map of a polyhedron along a closed curve; the curve begins and ends at the same point, so the first and last polygons in the developing map will be congruent.*
*The curve is homotopic to a geodesic if:*

1. *it is possible to choose a point on one face and connect it to the associated point on the associated face by a straight line that does not exit the developing map, and*
2. *the faces are translations of one another.*

## H.2 Dual Graphs

The graph of the dual of a polyhedron $P$ turns out to be another useful method for finding geodesics on $P$ since distinct cycles on the dual graph correspond to distinct homotopy classes of curves on $P$. Here, we present some preliminary definitions.

The edge graph of $P$ is a graph in which $P$'s edges are the graph's edges and $P$'s vertices the graph's vertices. Every simple, spherical (genus zero) polyhedron has a connected, planar edge graph. The dual graph $G_P$ is derived from $P$'s edge graph by constructing a single vertex in each face (including the outside of the graph) and connecting two vertices by an edge if the faces they lie on share an edge.

For example, let $P$ be the cube, shown in Figure H.2(a). Its edge graph is shown in Figure H.2(b) and the dual graph is constructed in Figure H.2(c). The dual graph has been redrawn for clarity in Figure H.2(d).

The set of faces $\{f_i\}$ of $G_P$ is bijectively related to the set of cone points of $P$, and thus each face represents—and can be assigned—curvature equivalent to that of the corresponding cone point. We will generally refer to faces by the number of sides they possess—for instance, a triangular or a hexagonal face—since in all the polyhedra we will be working with, two corners where the same number of faces meet will be congruent. Note, however, that the curvature depends only on the original polygon $P$ and cannot be derived from $G_P$ or even the edge graph of $P$.

A *cycle* is a set of consecutive edges of $G_P$ beginning and ending at the same point. It may pass through the same vertex more than once.

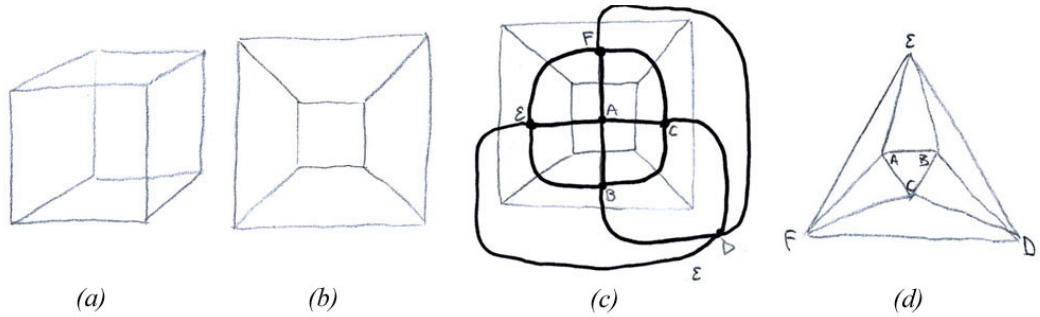



Figure H.2: Constructing the dual graph of a cube.

**Observation H.2.1.** *If $c$ is a cycle on $G_P$, $c$ corresponds to a homotopy equivalence class $[p_c]$ of continuous, closed curves $p_c$ on $P$.*

**Theorem H.2.2** (Jordan Curve Theorem)**.** *Every simple, closed curve in the plane divides the plane into an inside and outside region (and any path connecting a point in the inside to a point in the outside intersects the curve in at least one point).*

This is also true on the sphere. The plane is homeomorphic to a sphere with one point removed. Since a plane-filling curve can't be simple, for every simple, closed curve on the sphere there is at least one point that is not part of the curve. Remove that point and flatten the sphere, and the Jordan Curve Theorem applies.

Assume that $P$ is finite and has genus 0, and that $p_c$ is a curve thereon. Since $p_c$ is simple and closed, by the Jordan Curve Theorem it is the boundary of a disk. There are a finite number of cone points $p_i$ contained in the disk, and we define the *enclosed curvature* of $p_c$ to be the sum of their individual curvatures $\kappa_{p_i}$.

Note that the Jordan Curve Theorem only applies for surfaces with genus zero—a curve that encircles the torus' hole, for instance, does not divide the torus into two regions. We will mainly deal with spherical polygons, but later in this paper we will also deal with the general notion.

**Definition H.2.** Choose $p_c \subset P$ and construct the corresponding cycle $c \subset G_P$. If a cone point $p_i$ is in the interior of $p_c$, then $c$ can be said to *enclose* the corresponding face $f_i$ in $G_P$, and the enclosed curvature of $c$ is equal to the enclosed curvature of $p_c$.

**Theorem H.2.3** (The Gauss-Bonnet Theorem)**.** *If there exists a geodesic bounding a disk on $P$, then its enclosed curvature is $2\pi$.*

$c$ and the equivalent cycle of the opposite orientation (for which the "interior" and "exterior" sets are switched) can be regarded as equivalent—if either of them has curvature of $2\pi$, then both are geodesics. We will refer to both orientations with the same notation of $c$.

The Gauss-Bonnet Theorem gives our first necessary condition for geodesics: in order to find geodesics on $P$, we will look first for curves enclosing curvature of $2\pi$. The enclosed curvature of a curve depends only on its homotopy class, so the dual graph useful in finding candidate $[p_i]$.

One homotopy class is shown by the path on the dual graph of the cube in Figure H.3. Without loss of generality, we can choose a side of the closed path to call the interior and shade it. The path encloses four

of the eight triangles; each of the triangles corresponds to one of the curvature-$\pi$ cone points of the cube, for a total enclosed curvature of $2\pi$. The corresponding region is shown, shaded, on the cube in Figure H.3.

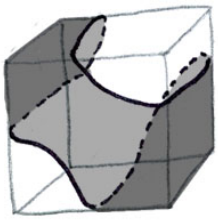


Figure H.3: A homotopy class of curves which does not contain a geodesic.

But, as shown in the developing map (figure H.4), there is no straight geodesic in this homotopy class.

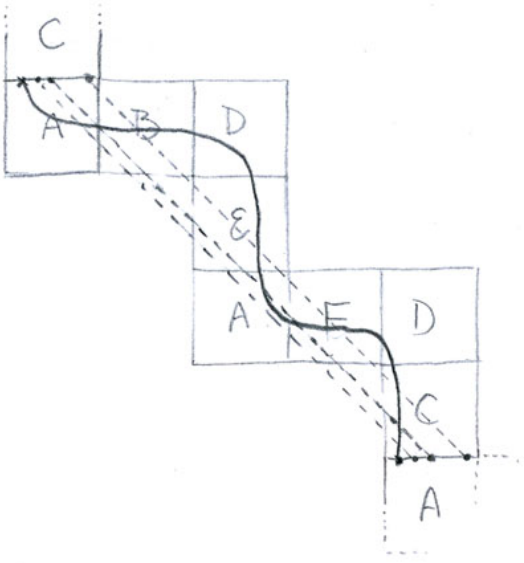



Figure H.4: Developing map of the curve in Figure H.3.

Next, we must develop methods for detecting homotopy equivalence classes which contain geodesics.

## H.3 Quotient Spaces

In some circumstances, it's possible to use a geodesic on one surface to find a geodesic on another. For example, all geodesics on the tetrahedron correspond to at least one and no more than two geodesics on the torus. This is a consequence of the fact that the torus covers the tetrahedron twice over.

What is a quotient space? A typical representation of a torus is a parallelogram with its opposite edges identified, as shown in Figure H.5.

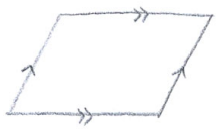


Figure H.5: Torus quotient space

A torus can also be seen as a quotient of the plane, however. Define a group of isometries of $\mathbb{R}_2$, $T = \left\{\iota, \tau_{(p,q)}, \tau_{(r,s)}\right\}$ (the identity and two translations along vectors $(\vec{p,q})$ and $(\vec{r,s})$). Also define an equivalence relation $x \sim x' \Leftrightarrow \phi x = x'$ (where $x, x' \in \mathbb{R}_2$ and $\phi \in T$). Then the torus, expressed as a *quotient space* of $\mathbb{R}_2$, is $\mathbb{R}_2/T$.

The tetrahedron is a quotient space of the torus (and thus, also of $\mathbb{R}_2$).

**Lemma H.3.1.** *Suppose $G$ and $H$ are subsets of the same group. Then $H$ is covered by (or* acts on*) $S/G$ if and only if $h^{-1}gh \in G \ \forall g \in G, h \in H$.*

*Proof.* In the group $S/G$, when $x, y \in S$ and $g \in G$, $x \sim y \Leftrightarrow g(x) = y$.

1. ($\Rightarrow$ :) Assume that $S/G$ covers $H$. Then $x \sim y$ implies that $h(x) \sim h(y)$ for all $h \in H$. $h(x) \sim h(y)$ means that $h(x) = g(h(y))$ for some $g \in G$; since $x \sim y$ implies that there exists $g' \in G$ such that $x = g'(y)$, then $g' = h^{-1}gh$. So $h^{-1}gh \in G$ for all $g \in G, h \in H$.
2. ($\Leftarrow$ :) Assume $hgh^{-1} \in G$ for all $g \in G, h \in H$. $hgh^{-1} \in G$ implies that there exists some $g' \in G$ such that $hgh^{-1} = g'$. In $S/G$ $x = g'(y) \Leftrightarrow x \sim y$ for any $g' \in G$. So $g'(y) = h^{-1}gh(y)$, and then $x = h^{-1}gh(y)$, and $h(x) = g(h(y))$. Thus $h(x) \sim h(y)$ and $S/G$ covers $H$.

Q.E.D.

Thanks to Lemma H.3.1, we only need for $\{\iota, \nu\}$, the tetrahedron group, to commute with $\{\tau_1, \tau_2\}$ to check that the torus, $\mathbb{R}^2/T$ covers the tetrahedron.

**Theorem H.3.1.** *$\{\iota, \nu\}$ acts on the torus group.*

*Proof.* 1. *Identity* Clearly, $\iota\tau = \tau\iota \ \forall \tau \in T$.

2. *Point Inversion* We must show that $\nu\tau = \tau\nu \forall \tau \in T$.

   Choose $(x, y) \in \mathbb{R}_2$ and $\tau_{(p,q)} \in T$.

   $\nu\tau_{(p,q)} = \tau_{(p,q)}\nu \Rightarrow \tau_{(p,q)} = \nu^{-1}\tau_{(p,q)}\nu$, so:

   (a) $\tau_{(p,q)}((x, y)) = (x + p, y + q)$

   (b)
   $$\begin{aligned} \nu^{-1}\tau_{(p,q)}\nu((x, y)) &= \nu^{-1}\tau_{(p,q)}((-x, -y)) \\ &= \nu^{-1}((-x + p, -y + q)) \\ &= (x - p, y - q) \end{aligned}$$

   But $\tau^2_{(p,q)}((x - p, y - q)) \sim (x + p, y + q)$, so by the equivalence relation on the torus, these two points are the same.

Thus, $\{\iota, \nu\}$ commutes with the torus group. Q.E.D.

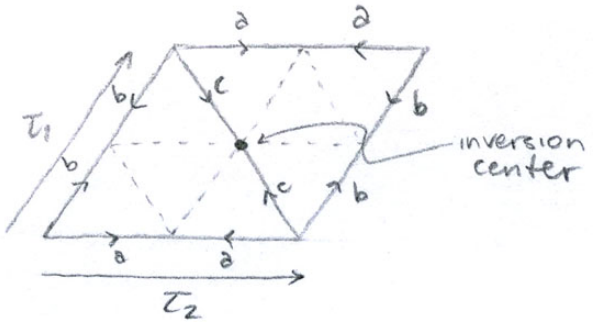



Figure H.6: A torus plus a point inversion equals a tetrahedron

Folding up the torus in Figure H.6 so that the associated points touch yields a tetrahedron with each side two layers thick. This is why the torus is said to *cover* the tetrahedron. The torus is a *twofold cover* of the tetrahedron because equivalence of points under $\nu$ ensures that two points $(x, y)$, $(-x, -y)$ which are not equivalent under $T$ are mapped by $\nu$ to a single point in the tetrahedron.

The idea of covering suggests an interesting way to transfer geodesics from the tetrahedron to the torus. Imagine drawing a geodesic on the tetrahedron in ink so that, as it is rolled across the torus, it leaves an imprint of the geodesic. More precisely, define a mapping $\Psi$ from the original space to the quotient space. Using our example, which takes the torus to the tetrahedron, $\Psi(x) = \Psi(y)$ if and only if $x = \phi(y)$ where $\phi \in \{\iota, \nu\}$. Then, with a slight abuse of notation, $p_c = \Psi(q_c)$ when $\forall x \in q_c$, $\Psi(x) \in p_c$. In this case, $\Psi$ is a two-to-one function because the order of an element in $\phi \in \{\iota, \nu\}$ is at most two.

This leads to a very useful theorem:

**Theorem H.3.2.** *If $Q$ is a finite cover of $P$ where $P/\Phi = Q$, then for any simple, closed $p_c$ on $P$ there is at least one corresponding simple, closed $q_c$ on $Q$.*

*Proof.* 1. *$p_c$ is simple $\Rightarrow$ $q_c$ is simple*: Assume toward contradiction that $q_c$ is not simple. Then there is a point where it intersects itself. Since covering is locally one-to-one, $p_c$ would also need to have an intersection, which is a contradiction.

2. *$p_c$ is closed $\Rightarrow$ $q_c$ is closed*: Suppose there is a curve segment with endpoints $q \neq q'$ where $\Psi^{-1}(q) = \Psi^{-1}(q')$; then the curve on P is closed, but the curve on Q is not. Then we must continue the geodesic.

   Suppose $\Psi(q) = \Psi(q')$ but $q \neq q'$. $\Psi(q) = \Psi(q') \Rightarrow \exists \phi \in \Phi : \phi q = q'$. Then let $\phi^2 q = q''$; since $\phi^2 \in \Phi$, $\Psi(q) = \Psi(q'')$—in other words, the geodesic has been traced out again from $P$ onto $Q$ and now begins at $q$, passes through $q'$, and ends at $q''$.

   Assume toward contradiction that $\nexists n \in \mathbb{N} : \phi^n q = q$. Then $q \neq \phi q \neq \phi^2 q \neq ....$ But $\Psi(q) = \Psi(\phi q) = \Psi(\phi^2 q) = ...$, meaning that an infinite number of points in $Q$ map to a single point in $P$. This contradicts our assumption that $Q$ is a finite cover of $P$.

   Therefore, the transferred geodesic $q_c$ must eventually close as long as $p_c$ does.

3. *$p_c$ is smooth $\Rightarrow$ $q_c$ is smooth*: The tangent vector at $q$ is mapped to the tangent vector at $\phi q = q'$ by $\phi$, so the curve between $q$ and $q''$ is at least $C^1$. See Figure H.7.

Q.E.D.

Note that mapping a geodesic $q_c$ on $Q$ to $P$ does not always preserve simplicity of the geodesic.

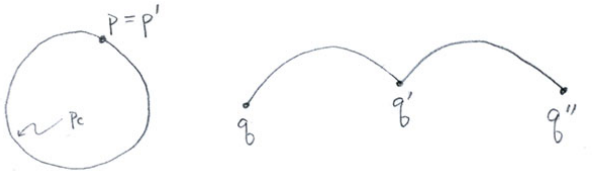



Figure H.7: If the curve is $C^1$ at $p'$, it must be $C^1$ at $q'$.

## H.4 Geodesics on Zonohedra

A zonohedron is a finite, convex polyhedron with parallelogram faces. A zonohedron is uniquely defined by its *star*, a set of $n$ vectors $e_1, e_2, ..., e_n$: take convex hull of the points

$$x_1\vec{e_1} + x_2\vec{e_2} + ... + x_n\vec{e_n} \ (x_i \text{ either 0 or 1})$$

[1]

We are interested in zonohedra because there is an algorithmic way of generating them, which implies there might be an algorithmic way of generating geodesics on them.

### H.4.1 Geodesics on Zonohedra with 3 Star Vectors

The most obvious geodesics on a zonohedron are called *zone geodesics* because they lie in *zones*. There are $n$ zones in a zonohedron with $n$ star vectors; the zone is the "encircling band" of faces which each have two edges equal and parallel to the given star vector, as shown in Figure H.8. [1]

Not every zone allows a zone geodesic; one of our first questions is whether there are zonohedra with *no* zone geodesics. To reduce this problem to a manageable size, we will limit our exploration to the set of zonohedra with three star vectors $\vec{e_1}$, $\vec{e_2}$, and $\vec{e_3}$, all of unit length.

We would like to determine when the $\vec{e_1}$-zone contains a closed geodesic. First, note that the upper vertices of the $e_1$-zone are given by 0, $e_3$, $e_2 + e_3$, and $e_2$, while the lower vertices are given by $e_1$, $e_1 + e_3$, $e_1 + e_2 + e_3$, and $e_1 + e_2$, as shown in Figure H.9.

**Lemma H.4.1.** *The $\vec{e_1}$-zone contains a simple geodesic if and only if*

$$|\vec{e_1} \cdot \vec{e_2}| + |\vec{e_1} \cdot \vec{e_3}| < 1$$

*Proof.* Take the dot product of all the zone vertices with $\vec{e_1}$. Then the minimum value for the upper zone needs to be above the maximum value for the lower zone. Inspection of these conditions proves the claim. Q.E.D.

Now, which zonohedra have no zone geodesics? Let

$$\begin{pmatrix} 1 & a & b \\ a & 1 & c \\ b & c & 1 \end{pmatrix}$$

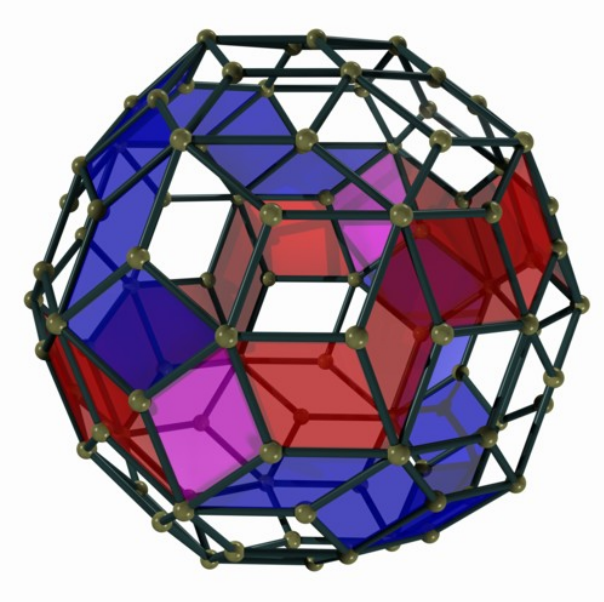


Figure H.8: Two zones (highlighted) on a zonohedron.

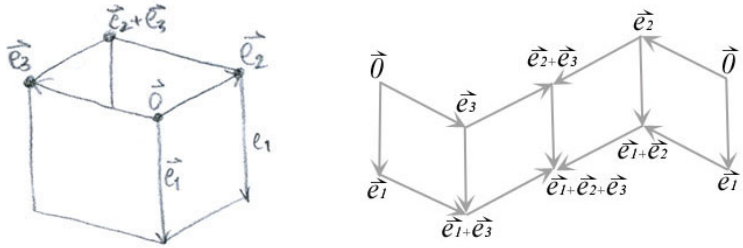



Figure H.9: A zone from a generic zonohedron with three star vectors, and its developing map.

be the Gram matrix of the dot products of the $\vec{e_i}$. Then there are no zone geodesics if and only if $|a| + |b|$, $|a| + |c|$, and $|b| + |c|$ are all less than 1. For example, let $1 > a = b = c > \frac{1}{2}$. This suffices to make the matrix positive definite, so it is invertible and thus the columns are a basis.

Using the Gram-Schmidt process, we can find an orthonormal basis of $\mathbb{R}^3$ with respect to the dot product given by that Gram matrix. By changing the basis, we can thus find 3 vectors $\vec{e_i}$ in $\mathbb{R}^3$ that have precisely the above Gram matrix.

Clearly there are zonohedra without any zone geodesics, but there may also be other closed, non-zone geodesics. The cube is a specific example of a zonohedron with three star vectors, and [2] gives the only three simple, closed geodesics on the cube, shown in Figure H.10.

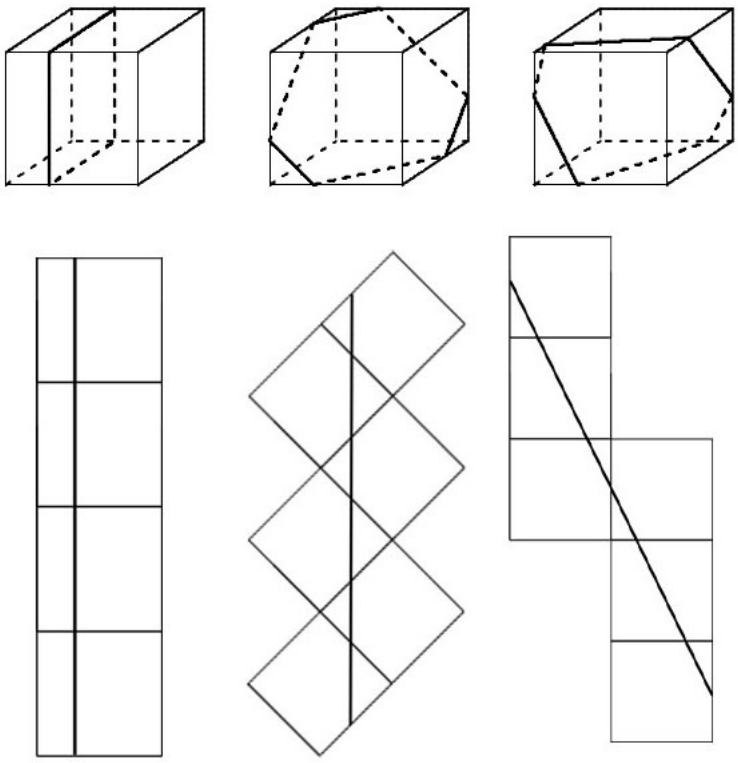


Figure H.10: The three cube geodesics and their developing maps. Figure adapted from [2].

Their first, denoted $(0, 4)$, is the zone geodesic on the cube. It would be similarly possible, although not quite as straightforward, to characterize the other types of geodesic on an arbitrary zonohedron $P$ in $\mathcal{M}_3$.

It is important to note that the geodesics shown are not the *only* possible geodesics on an arbitrary $\mathcal{M}_3$ zonohedron. The degenerate case where $\vec{e_1}$, $\vec{e_2}$, and $\vec{e_3}$ are coplanar and evenly spaced at 120°apart can be described as a doubled hexagon (that is, a hexagon with both a "back" and a "front" side) with two additional punctures at the points $O$ and $O'$. This polyhedron has a geodesic shown in Figure H.11 which doesn't correspond to any of the geodesics on the cube.

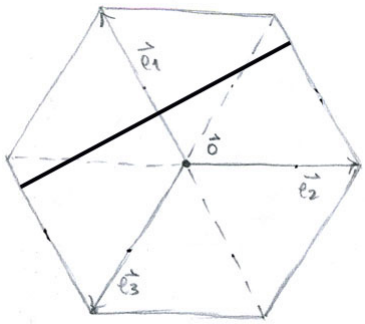



Figure H.11: A degenerate zonohedron with three coplanar star vectors has a geodesic which doesn't fit into any of the categories in Figure H.10.

### H.4.2 Geodesics on the Rhombic Dodecahedron

The problem with the methods used above for $\mathcal{M}_3$ is that they do not extend easily to more complicated zonohedra. For instance $\mathcal{M}_4$, the space of zonohedra with four unit-length star vectors, has five degrees of freedom and $\mathcal{M}_5$ has seven.

$\mathcal{M}_1$, the space of zonohedra with one unit-length star vector, has 0 degrees of freedom: no matter where on the unit sphere a vector lies, it is equivalent (under rotation and reflection) to every other vector. In $\mathcal{M}_2$, the only feature that distinguishes zonohedra is the angle between the two vectors: thus, zero degrees of freedom for placing the first vector plus one for placing the second gives one total degree of freedom. In order to fix the arrangement of three unit vectors ($\mathcal{M}_3$), it is necessary to specify the angles between the third vector and both of the previous vectors, giving $0 + 1 + 2 = 3$ degrees of freedom. As long as the zonohedron remains embeddable in $\mathbb{R}^3$, additional vectors can be fixed with only two additional specified angles, but even for the four star vector case, this gives $0 + 1 + 2 + 2 = 5$ degrees of freedom. This makes the space of more complicated zonohedra considerably more difficult both to visualize and to work with in the way we did above.

Once we have the dimension, we still must find the simple, non-zone geodesics, which isn't straightforward: the case for the cube used above had previously been completed by [2]. In order to extend the approach to $\mathcal{M}_4$, we would first have to find all the geodesics on the hypercube.

As further generalizations were not practical, we decided to explore specific cases in hope of larger insights. One of those we studied was the the rhombic dodecahedron (Figure H.12). The rhombic dodecahedron's star consists of four unit vectors that point to the top four corners of a cube centered at the origin:

$$\left\{ \begin{array}{ll} \vec{e_1} = & (1,1,1) \\ \vec{e_2} = & (1,-1,1) \\ \vec{e_3} = & (-1,1,1) \\ \vec{e_4} = & (-1,-1,1) \end{array} \right\}$$

We found all of the paths $c$ enclosing $2\pi$ by an exhaustive search using the dual graph $G_{rd}$ (Figure H.13). Excluding isomorphisms, there were 20. Table H.4.2 gives a breakdown of these cycles.

Of the three $c$ for which $[p_c]$ contains a geodesic, the geodesic lengths vary. All the vectors in the rhombic dodecahedron's star have the same magnitude; for simplicity, we will assume they are unit vectors. The zone geodesic (listed in Table H.4.2 as the one path $c$ which passes through 6 faces) is the shortest, with a length

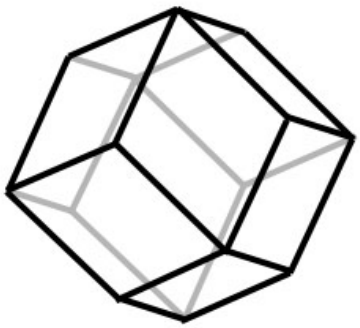


Figure H.12: Rhombic dodecahedron.

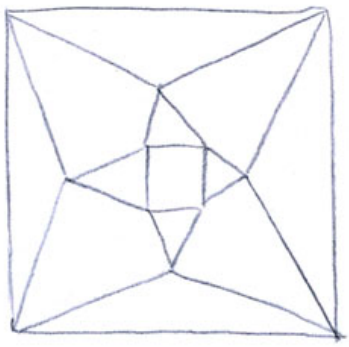


Figure H.13: $G_{rd}$, the net of the rhombic dodecahedron's dual.

Table H.1: Curves on the Rhombic Dodecahedron

| No. of faces $c$ passes through | Distinct $c$ | $[p_c]$ contains a geodesic |
|---|---|---|
| 6 | 1 | 1 |
| 8 | 3 | 2 |
| 10 | 9 | 0 |
| 12 | 6 | 0 |

(relative to the unit vector) of about 5.66. The other two geodesics, which both pass through eight faces, have relative lengths of about 5.81 and 6.16.

We made several interesting observations:

**Theorem H.4.1.** *There must be an equal number of left and right turns in c.*

**Theorem H.4.2.** *There are exactly three simple geodesics that pass through any face at most once on a rhombic dodecahedron.*

*Proof.* By exhaustion, using the dual. Q.E.D.

**Conjecture H.1.** *There is no zonohedron with a geodesic that passes through the same face more than once.*

If true, this would prove that the geodesics listed above are all possible geodesics.

**Observation H.4.3.** *The shortest geodesic is the zone geodesic.*

This seems to be true for all zonohedra.

## H.5 Geodesics on 6, 4, 4

Based on our work, it seems unlikely that it is possible to find arbitrarily long geodesics on any convex polyhedron except the tetrahedron. Then why not try polyhedra with cone angles of greater than $2\pi$?

The infinite skew polyhedra have several interesting characteristics: all their faces are congruent, as are all their vertices, but the curvature at the vertices is more than $2\pi$. One such polyhedron is known as $\{6, 4|4\}$ (Figure H.14). Its faces are regular hexagons, four of which meet at each vertex, giving each vertex a cone angle of $2\pi - 4(\frac{\pi}{3}) = -\frac{2\pi}{3}$.

Unbounded polyhedra like $\{6, 4|4\}$ are difficult to work with using the methods detailed in Section H.2. A quotient space would be much easier, especially a quotient space which is homeomorphic to the sphere, since we want to guarantee a connected, planar edge graph and dual graph.

One finite quotient space of $\{6, 4|4\}$ (by three equal, mutually orthogonal translations $\tau_{e_1}$, $\tau_{e_2}$, and $\tau_{e_3}$) is the object shown in Figure H.15, which has genus 3.

This is easier to visualize, but still rather complicated, so we will introduce a further quotient space: the stella octangula (Figure H.16).

**Observation H.5.1.** *The involution through the center point of a hexagonal face is well-defined on* $\{6, 4|4\} / \langle \tau_1, \tau_2, \tau_3 \rangle$. *It has eight fixed points (the center of every hexagon is preserved by the involution) and the quotient space is the stella octangula. Proof is by inspection.*

*The covering is twofold since an involution identifies at most two points of the original.*

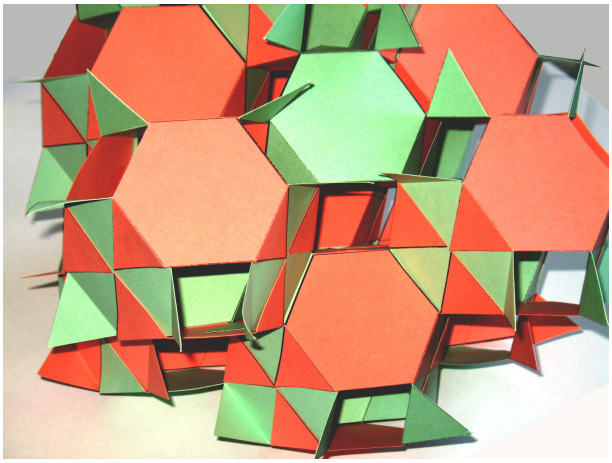


Figure H.14: A paper model of $\{6, 4|4\}$.

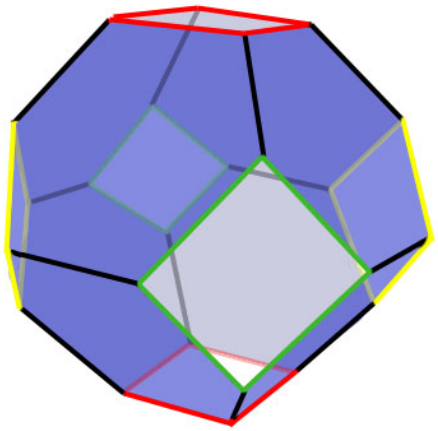


Figure H.15: $\{6, 4|4\} / \langle \tau_1, \tau_2, \tau_3 \rangle$. Red, green, and yellow edges show associations.

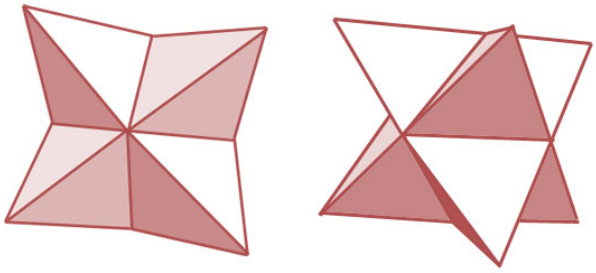


Figure H.16: Two views of the Stella octangula.

So by Theorem H.3.2, geodesics on the stella octangula correspond to potential geodesics on $\{6,4|4\}\,/\,\langle\tau_1,\tau_2,\tau_3\rangle$, and thus on $\{6,4|4\}$, as long as they remain simple when lifted from the stella octangula to $\{6,4|4\}$. Since the covering is twofold (when lifting to $\{6,4|4\}\,/\,\langle\tau_1,\tau_2,\tau_3\rangle$) or infinite (when lifting to $\{6,4|4\}$), simplicity on these is not guaranteed.

### H.5.1 Geodesics on the Stella Octangula

Let $G_{so}$ (Figure H.17) be the graph of the stella octangula dual. $G_{so}$ has six 8-loops and eight 3-loops.

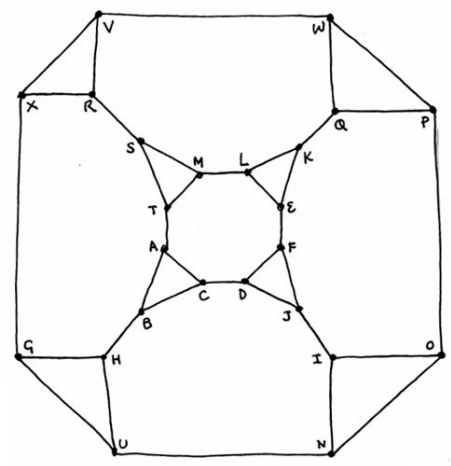



Figure H.17: $G_{so}$, the graph of the stella octangula dual.

**Theorem H.5.2.** *Each 8-loop enclosed by c represents an added curvature of* $-\frac{2\pi}{3}$.

*Proof.* The faces of the stella octangula are congruent equilateral triangles, so an octahedral loop represents a cone point where eight equilateral triangles meet. Since each triangle has an angle of $\frac{\pi}{6}$, the total curvature at the cone point is $\kappa_8 = 2\pi - 8(\frac{\pi}{6}) = 2\pi - \frac{4\pi}{3} = -\frac{2\pi}{3}$. Q.E.D.

Similarly, each 3-loop adds a curvature of $2\pi - 3(\frac{\pi}{6}) = \pi$ to $c$.

**Theorem H.5.3.** *If a path class* $[p_c]$ *contains a geodesic, then c encloses either zero or three 8-loops.*

*Proof.* By the Gauss-Bonnet Theorem (Theorem H.2.3), if $[p_c]$ contains a geodesic then it encloses curvature of $2\pi$, and by Definition H.2, the curvature of $c$ is equal to the curvature of $[p_c]$. Since each 8-loop represents $-\frac{2\pi}{3}$ curvature, in order to get a curvature which is an integer multiple of $\pi$, $p_c$ must enclose $3n$ 8-loops (where $n \in \mathbb{Z}$). Since there are only six 8-loops in $G_{so}$, $c$ must enclose zero, three, or six of them.

It is clear by inspection that for any cycle $c$ which encloses six 8-loops is equivalent, there is a corresponding cycle of the opposite orientation that encloses zero 8-loops. Since the two are equivalent for our purposes, the case where $c$ encloses six 8-loops can be ignored. Q.E.D.

Further, if $[p_c]$ contains a geodesic and $c$ encloses no 8-loops, $c$ must enclose two 3-loops; and if $c$ encloses three 8-loops, $c$ encloses four 3-loops.

Unfortunately, these restrictions still allow many equivalence classes of curves which do not contain a geodesic. Fortunately, there are a few other restrictions on $c$ which help to narrow the field.

**Theorem H.5.4.** *If $c$ contains $\geq 5$ consecutive points of an 8-loop, then there is no geodesic in $[p_c]$.*

*Proof.* Assume toward contradiction that $c$ contains 5 consecutive points of an 8-loop. Then $p_c$ must pass through five equilateral triangles which share a point, as shown in Figure H.18, but *not* through O, the cone point.

$\bigtriangleup AOB$ refers to the union of the interior and the boundary of $\bigtriangleup AOB$, minus $\{O\}$.

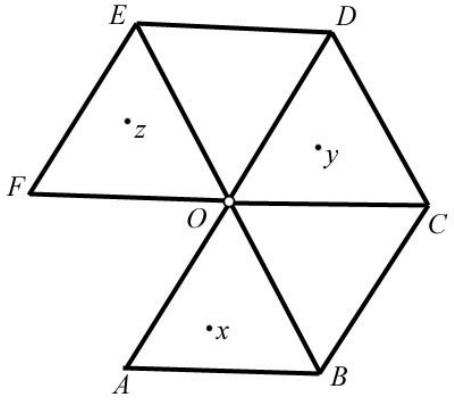



Figure H.18: Developing map of a curve which passes through five consecutive equilateral triangles.

Since $p_c$ must pass through all five triangles in order, it is possible to choose three points

$$\left\{ \begin{array}{c} x \in \bigtriangleup AOB \cap p_c \\ y \in \bigtriangleup COD \cap p_c \\ z \in \bigtriangleup EOF \cap p_c \end{array} \right\}$$

Then $p_c \supset \overline{xyz}$, where $y$ is between $x$ and $z$.

Let $\mathcal{H}$ be the half-plane defined by $\overleftrightarrow{BOE}$ and containing $x$. ($B, O, E$ are collinear since $\measuredangle BOC = \measuredangle COD = \measuredangle DOE = 60°$ and thus $\measuredangle BOE = 180°$.) Since $x \in \bigtriangleup AOB$ and $x \in \mathcal{H}$, $A \in \mathcal{H}$. $AB = BO = OF = FA \Rightarrow \square ABOF$ is a rhombus and thus $\overleftrightarrow{BO} || \overleftrightarrow{AF}$; then since $A \in \mathcal{H}$, $F \in \mathcal{H}$. $O, E, F \in \mathcal{H} \Rightarrow \bigtriangleup EOF \in \mathcal{H} \Rightarrow z \in \mathcal{H}$, so $\overline{xz} \in \mathcal{H}$.

$A \neq O \neq D$ and $\measuredangle AOD = 180°$, so $O$ is between $A$ and $D$. $\overline{AOD}$ intersects $\overleftrightarrow{BOE}$ at $O$, and therefore $A \in \mathcal{H} \Rightarrow D \notin \mathcal{H}$. Similarly, $F \in \mathcal{H}$ and $C \notin \mathcal{H}$. Thus $\bigtriangleup COD - \{O\} \notin \mathcal{H}$ and $y \notin \mathcal{H}$. Then $y \notin \overline{xz}$, a contradiction. So $\nexists c \in C$ which passes through five or more points of an 8-loop consecutively. Q.E.D.

We have not yet carried out the full analysis of geodesics on the stella octangula. There are a few interesting examples which I will note, however. For one, some geodesics have cycles $c$ which trace the same edge twice (Figure H.19).

Once all the geodesics on the stella octangula have been classified, the next step would be to lift geodesics on the stella octangula to $\{6, 4|4\} / \langle \tau_1, \tau_2, \tau_3 \rangle$ and $\{6, 4|4\}$ and check whether they are simple.

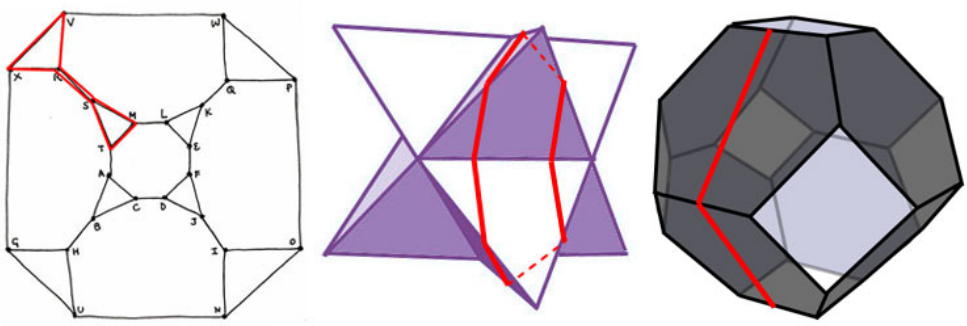


Figure H.19: One example of a geodesic on $G_{so}$, the stella octangula, and $\{6, 4|4\}$.

## H.6 Conclusion and Further Work

We discovered some interesting things about geodesics on polyhedra. We completely classified geodesics on the rhombic dodecahedron, the $\mathcal{M}_3$ zonohedra which had zone geodesics, and got a significant start on classifying geodesics on much more complicated polyhedra, the stella octangula and $\{6, 4|4\}$. We also developed new methods that may help lay the foundation for using a computer to find geodesics.

Of course, there is a lot of room for further work. One large open question is whether all polyhedra have at least one simple, closed geodesic. Our work with $\mathcal{M}_3$ might make it possible to find such a zonohedron if it exists. This would require finding non-zone geodesics on the hypercube and beyond, as well as accounting for degenerate cases.

The case of the rhombic dodecahedron shows that there are some interesting possibilities for computing geodesics. As preliminary work, it could be useful to quickly run through other test cases and seek out other patterns. There is also potential in finding better algorithms to search for cycles in dual graphs and deciding which path classes contain geodesics.

## H.7 Acknowledgments

I would like to thank Prof. Weber for his unflagging commitment of time and energy to working with me this summer and giving me a taste of what it means to be a mathematician. Without Prof. Pilgrim and Mandie McCarty, this program could never have happened (and we might have all starved)! Finally, but not least, I would like to thank my fellow REU students for fun times and fascinating discussions—and helping one another over hurdles, of course.

## Bibliography

1. Coxeter. *Regular Polytopes*, Dover Publications, Inc., New York, 1978
2. Fuchs and Fuchs. Closed Geodesics on Regular Polyhedra, *Moscow Mathematical Journal* **Vol. 7, No. 2** (2001) 265-79
3. Harary. *Graph Theory*, Narosa Publishing House, New Delhi, 1998

# The Erdös Box Problem

Chengcheng Yang
Rice University

Indiana University REU Summer 2009
Advisor: Nets Katz

## I.1 introduction

Given an $N \times N \times N$ grid, A is a subset that does not contain the eight corners of any nontrivial box. Let $I_{ijk}$ be an $N \times N \times N$ tensor of $1's$ and $0's$, then

$A = \{I_{ijk} : \prod_{l=1}^{2} \prod_{m=1}^{2} \prod_{n=1}^{2} = 0 \ for \ any \ i_1 \neq i_2, \ j_1 \neq j_2, \ k_1 \neq k_2\}.$

The question posted by Erdös asks the lower bound of $|A|$ for an arbitrary N. Using Cauchy–Schwartz Inequality, we may show that $|A|$ is no larger than $O(N^{\frac{11}{4}})$. But Erdös conjectured that $O(N^{\frac{11}{4}})$ is the lower bound of $|A|$. i.e. for any $\varepsilon > 0$, there is a set A satisfying the box condition such that $|A| = O(N^{\frac{11}{4}-\varepsilon})$. To provide evidence for Erdös' conjecture, efforts have been made to look for examples with exponents of N as close to $\frac{11}{4}$ as possible. The currently knwon exponent closest to $\frac{11}{4}$ is $\frac{8}{3}$ discoved by N. Katz, E.Krop, and M. Maggioni.[2]

## I.2 Katz–Krop–Maggioni's Example

In Katz–Krop–Maggioni (KKM)'s paper, a finite field $\mathbb{F}_p$ and an extension field $\mathbb{F}_{p^3}$ are employed to find the desired set A. In the previous work done by K. Gunderson, V. Rodl, and A. Sidorenko [1], they used random planes over $\mathbb{F}_p$. However, the appearance of $\mathbb{F}_{p^3}$ in KKM's paper is a novelty. Let's look at the KKM example in detail.

Consider the finite field $\mathbb{F}_p$, then $\mathbb{F}_{p^3}$ is the extension of $\mathbb{F}_p$ by an irreducible cubic polynomial over $\mathbb{F}_p$. If we let $r$ be a root of this cubic polynomial, then the span of $\{1, r, r^2\}$ over $\mathbb{F}_p$ gives the extension field $\mathbb{F}_{p^3}$. i.e. any element $\alpha \in \mathbb{F}_{p^3}$ can be written as $a + br + cr^2$, where $a, b, c \in \mathbb{F}_p$.

let one corner of the grid be the origin and the three adjacent sides be the $x$- , $y$- , and $z$- axes. Then each vertex of the grid can be identified with a point in the 3–D Euclidean space shown below.

Given the $x$- and $y$- coordinates of a vertex, the KKM example uses a bilinear map to find out the $z$-coordinates such that $(x, y, z)$ is in the set A. Since each pair $(x, y)$ has $N$ possible choices for $z$, it turns out that each pair $(x, y)$ is mapped to a set of $z$-coordinates. The following steps illustrate the procedure.
First, index the $N$ points on $x$-axis with $\mathbb{F}_{p^3} \backslash \{0\}$.So $N = p^3 - 1$. Then index the $N$ points on $y$-axis with $\mathbb{F}_{p^3} \backslash \{0\}$.
Second, define the bilinear map M by the multiplication rule in $\mathbb{F}_{p^3}$:

$$M : \ \mathbb{F}_{p^3} \ \times \ \mathbb{F}_{p^3} \ \rightarrow \ \mathbb{F}_{p^3}$$

$$M(x, y) = x * y$$

where $*$ denotes multiplication in $\mathbb{F}_{p^3}$.
Third, assign each $x * y$ to a plane $P_{x*y}$ in $\mathbb{F}_p^3$, which is defined by the equation

$$ax_1 + bx_2 + cx_3 = 1,$$

when

$$x * y = a + br + cr^2.$$

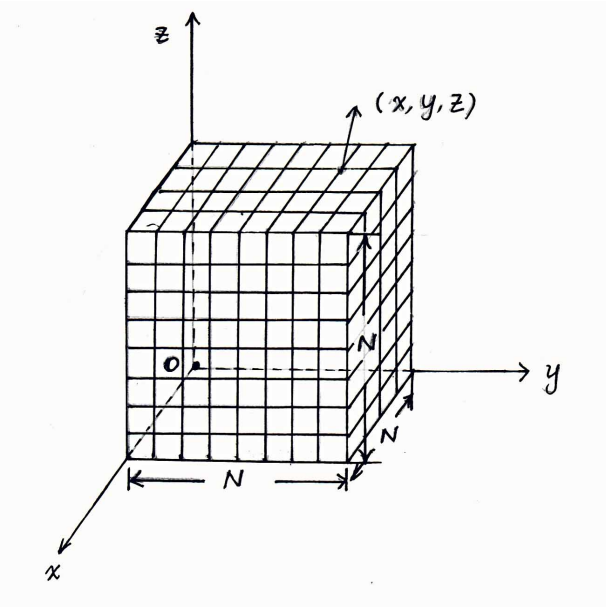



Figure I.1:

Fourth, index the $N$ points on $z$-axis with $\mathbb{F}_p^3 \backslash \{(0,0,0)\}$.Then the set of $z$-coordinates correspoinding to the pair $(x, y)$ are points lying on the plane $P_{x*y}$. Thus,

$$z = \{(x_1, x_2, x_3) | ax_1 + bx_2 + cx_3 = 1\}.$$

Since there are $p^2$ points lying on the plane $P_{x*y}$, each pair $(x, y)$ is mapped to $p^2$ $z$-coordinates. And there are $(p^3 - 1)^2$ pairs of $(x, y)$ in the $xy$-plane. Then $|A|$ can be calculated as follows:

$$\begin{aligned} |A| &= p^2(p^3-1)^2 \\ &= (p^3)^{\frac{2}{3}}(p^3-1)^2 \\ &\approx (p^3-1)^{\frac{2}{3}}(p^3-1)^2 \\ &= (p^3-1)^{\frac{8}{3}} \\ &= N^{\frac{8}{3}} \end{aligned}$$

Thus A is the set that satisfies the Erdös box condition and has $O(N^{\frac{8}{3}})$ elements. The significance of using multiplication in $\mathbb{F}_{p^3}$ is that give any nontrivial rectangle in the $xy$-plane, the four planes determined by the four corners of the rectangle intersect at most once. i.e.

$$|P_{x_1*y_1} \cap P_{x_1*y_2} \cap P_{x_2*y_1} \cap P_{x_2*y_2}| = 0 \text{ or } 1 \tag{I.1}$$

for any $x_1 \neq x_2, y_1 \neq y_2$.
This asserts that there does not exist $z_1 \neq z_2$ such that

$$z_1 \ , \ z_2 \in P_{x_1 * y_1} \cap P_{x_1 * y_2} \cap P_{x_2 * y_1} \cap P_{x_2 * y_2}.$$

If this condition fails, $A$ will contain the eight corners of a nontrivial box, whose vertices are $(x_1, y_1, z_1), (x_1, y_1, z_2), \ldots, (x_2,$ as shown below:

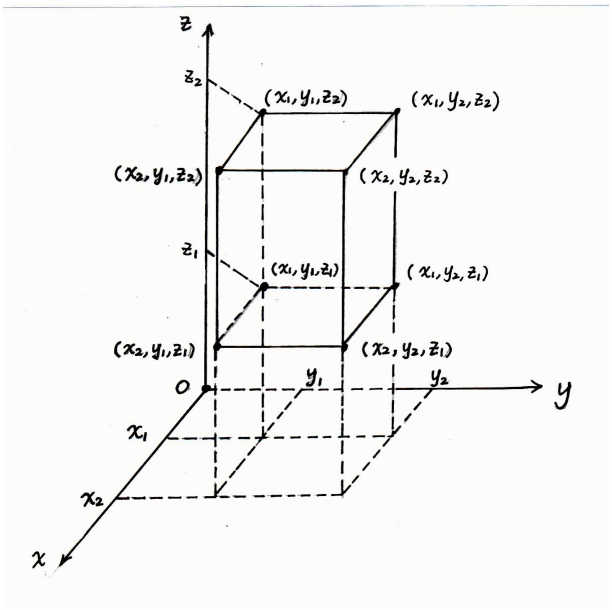



Figure I.2:

The KKM paper proves (I.1) by using the fact that linear operator $M_{y_1 * y_2{}^{-1}}$ does not preserve any line not containing the origin, where $M_{y_1 * y_2{}^{-1}}$ is multiplication in $\mathbb{F}_{p^3}$ by $y_1 * y_2{}^{-1}$ for any $y_1 \neq y_2$.

As a continuation of the KKM example, my work focuses on the question: Is the appearance of $\mathbb{F}_{p^3}$ a necessity or just a matter of convenience when the map is restricted to be bilinear? The conjecture is: If M is bilinear, then M is uniquely defined by the KKM example up to bordering.

## I.3 Main Results

Observe that if the function $M_y : \mathbb{F}_{p^3} \to \mathbb{F}_{p^3}$ is defined as: $M_y(x) = M(x, y)$ for all $x \in \mathbb{F}_{p^3} \backslash \{0\}$, then $M_y$ is linear due to the bilinearity of $M$. In fact, $\mathbb{F}_{p^3}$ is identical to $\mathbb{F}_p^3$, because if $\alpha = a + br + cr^2 \in \mathbb{F}_{p^3}$, then $\alpha \in \mathbb{F}_p^3 = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$, where $a, b, c \in \mathbb{F}_p$. Thus $M_y$ can be transformed to be a linear operator acting on the vector space $\mathbb{F}_p^3$ as $M_y : \mathbb{F}_p^3 \to \mathbb{F}_p^3$. And $M_y$ becomes a $3 \times 3$ matrix over $\mathbb{F}_p$.

For each $y$ in $\mathbb{F}_{p^3} \backslash \{0\}$, there is an associated linear operator $M_y$. Let $S$ be a set of these linear operator $M_y$'s, and the goal is to completely characterize the set $S$.

**Theorem I.3.1.** *S satisfies the Erdös box condition if and only if* $(M_{y_1}{M_{y_2}}^{-1})^t$ *does not preserve a line not containing the origin for any* $M_{y_1} \neq M_{y_2} \in S$.

*Proof.* We will use the same strategy of the KKM example.
If $S$ satisfies the Erdös box condition, then

$$|P_{M_{y_1(x_1)}} \cap P_{M_{y_1(x_2)}} \cap P_{M_{y_2(x_1)}} \cap P_{M_{y_2(x_2)}}| = 0 \text{ or } 1 \tag{I.2}$$

for any $x_1 \neq x_2, y_1 \neq y_2$. Since

$$P_{M_{y_1(x_1)}} \cap P_{M_{y_1(x_2)}} = L_{y_1} \tag{I.3}$$

is either empty or a line, we may assume it is a line. Similarly, we may assume

$$P_{M_{y_2(x_1)}} \cap P_{M_{y_2(x_2)}} = L_{y_2} \tag{I.4}$$

is a line. Then (I.2) becomes

$$|L_{y_1} \cap L_{y_2}| = 0 \text{ or } 1.$$

In fact, the intersection of two different lines is either a point or empty, we could prove (I.2) by showing that $L_{y_1}$ and $L_{y_2}$ are not the same line.
From (I.3),

$$L_{y_1} : \begin{cases} (X\ Y\ Z)M_{y_1(x_1)} = 1 \\ (X\ Y\ Z)M_{y_1(x_2)} = 1 \end{cases} \qquad (I)$$

where $X, Y, Z$ are variables in $\mathbb{F}_p$ .
From (I.4),

$$L_{y_2} : \begin{cases} (X'\ Y'\ Z')M_{y_2(x_1)} = 1 \\ (X'\ Y'\ Z')M_{y_2(x_2)} = 1 \end{cases} \qquad (II)$$

where $X', Y', Z'$ are variables in $\mathbb{F}_p$.
If the linear operator $(M_{y_1}{M_{y_2}}^{-1})^t$ preserves a line, $L_{y_1}$ can be mapped to $L_{y_2}$ through the following steps:

1. Left multiplication by $(M_{y_1}{M_{y_2}}^{-1})^t : \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \rightarrow (M_{y_1}{M_{y_2}}^{-1})^t \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$;
2. Let $\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = (M_{y_1}{M_{y_2}}^{-1})^t \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$;
3. Take the transpose of both sides: $(X'\ Y'\ Z') = (X\ Y\ Z)M_{y_1}{M_{y_2}}^{-1}$;

4. Right multiplication by ${M_{y_2}{M_{y_1}}^{-1}}$: $(X'\ Y'\ Z')M_{y_2}{M_{y_1}}^{-1} = (X\ Y\ Z)$;

5. Substitute $(X'\ Y'\ Z')M_{y_2}{M_{y_1}}^{-1}$into (I), we get (II).

Thus $L_{y_1} = L_{y_2}$. However, Erdös box condition implies that $L_{y_1} \neq L_{y_2}$. Hence, $(M_{y_1}{M_{y_2}}^{-1})^t$ cannot preserve any line not containing the origin.
To prove the converse, note that if $(M_{y_1}{M_{y_2}}^{-1})^t$ doesn't preserve a line for any $M_{y_1} \neq M_{y_2}$ , we get $L_{y_1} \neq L_{y_2}$. Then $|L_{y_1} \cap L_{y_2}| = 0$ or 1 and (I.2) is true. Q.E.D.

**Corollary 1.** *S satisfies the Erdös box condition if and only if ${M_{y_1}{M_{y_2}}^{-1}}$ has irreducible characteristic polynomial over $\mathbb{F}_p$.*

*Proof.* By Theorem (I.3.1), we know that $(M_{y_1}{M_{y_2}}^{-1})^t$ can't preserve any line not containing the origin, which means $(M_{y_1}{M_{y_2}}^{-1})^t$ can't preserve the two dimensional subspace spanned by the line. Thus $(M_{y_1}{M_{y_2}}^{-1})^t$ doesn't preserve any two dimensional subspace in $\mathbb{F}_p^3$. From Caley-Hamilton Theorem [3], we see that $(M_{y_1}{M_{y_2}}^{-1})^t$ does not have a quadratic factor in its characteristic polynomial. Hence, $(M_{y_1}{M_{y_2}}^{-1})^t$ has an irreducible characteristic polynomial over $\mathbb{F}_p$. Since ${M_{y_1}{M_{y_2}}^{-1}}$ is the transpose of $(M_{y_1}{M_{y_2}}^{-1})^t$, it follows that ${M_{y_1}{M_{y_2}}^{-1}}$ has an irreducible characteristic polynomial over $\mathbb{F}_p$. Q.E.D.

**Corollary 2.** *S satisfies the Erdös box condition if and only if QSR satisfies the Erdös box condition for any invertible matrices Q and R.*

*Proof.* The notation $QSR$ represents multiplying every matrix in $S$ by $Q$ and $R$. Thus $QSR = \{QM_yR\,|M_y \in S\}$. By corollary 1, we see that S satisfies the box condition if and only if ${M_{y_1}{M_{y_2}}^{-1}}$ has an irreducible characteristic polynomial. Consider two matrices in $S$, say $QM_{y_1}R$ and $QM_{y_2}R$, $(QM_{y_1}R)(QM_{y_2}R)^{-1} = Q(M_{y_1}{M_{y_2}}^{-1})Q^{-1}$ is a similar matrix to ${M_{y_1}{M_{y_2}}^{-1}}$. Thus $(QM_{y_1}R)(QM_{y_2}R)^{-1}$ also has an irreducible characteristic polynomial. Hence, $QSR$ also satisfies the box condition. It is easy to verify that the converse is also true. Q.E.D.

Using corollary 2, we are able to modify $S$ by an appropriate choice of $Q$ and $R$, so that it contains an identity matrix $I$ . Then we will reach corollary 3 as follows.

**Corollary 3.** *If S contains an identity matrix I and S satisfies the Erdös box condition, then any matrix in S is either a multiple of I or has an irreducible characteristic polynomial over $\mathbb{F}_p$ .*

*Proof.* If $M_y \in S$ is a multiple of $I$, say $M_y = dI$, where $0, 1 \neq d \in \mathbb{F}_p$, then $M_y$ takes the plane $ax_1 + bx_2 + cx_3 = 1$ into $ax_1 + bx_2 + cx_3 = d$. Clearly these two planes have empty intersection, because they are parallel to each other.
If a matrix $M_y \in S$ is not a multiple of $I$, by corollary 1, $M_y(I)^{-1} = M_y$ has an irreducible characteristic polynomial over $\mathbb{F}_p$ . Q.E.D.

Corollary 3 establishes the most significant property of $S$.

## I.4 More on the Conjecture

Previously, the bilinear form

$$M: \mathbb{F}_{p^3} \times \mathbb{F}_{p^3} \to \mathbb{F}_{p^3}$$

is a function over $\mathbb{F}_{p^3}$. After identifying $\mathbb{F}_{p^3}$ with $\mathbb{F}_p^3$, the bilinear form becomes

$$M: \mathbb{F}_p^3 \times \mathbb{F}_p^3 \to \mathbb{F}_p^3.$$

We introduce three bilinear forms $A_1, B_1, C_1$, each of which corresponds to one coordinate of $M(x,y)$ in the following way:

$A_1 : \mathbb{F}_{p^3} \times \mathbb{F}_{p^3} \to \mathbb{F}_p$ and $A_1(x,y) =$ 1st coordinate of $M(x,y)$

$B_1 : \mathbb{F}_{p^3} \times \mathbb{F}_{p^3} \to \mathbb{F}_p$ and $B_1(x,y) =$ 2nd coordinate of $M(x,y)$

$C_1 : \mathbb{F}_{p^3} \times \mathbb{F}_{p^3} \to \mathbb{F}_p$ and $C_1(x,y) =$ 3rd coordinate of $M(x,y)$.

Then

$$M(x,y) = \begin{pmatrix} A_1(x,y) \\ B_1(x,y) \\ C_1(x,y) \end{pmatrix}.$$

Using the matrix representation of bilinear form, we have

$$\begin{aligned} M(x,y) &= \begin{pmatrix} x^t A_1 y \\ x^t B_1 y \\ x^t C_1 y \end{pmatrix} \\ &= (x^t A_1 y \;\; x^t B_1 y \;\; x^t C_1 y)^t \\ &= [x^t (A_1 y \; B_1 y \; C_1 y)]^t \\ &= (A_1 y \; B_1 y \; C_1 y)^t x \end{aligned}$$

Therefore each matrix $M_y \in S$ can be expressed as $M_y = (A_1 y \; B_1 y \; C_1 y)^t$. Let $y = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$. Let $A_1 =$ $(A_{1(1)} \; A_{1(2)} \; A_{1(3)})$, $B_1 = (B_{1(1)} \; B_{1(2)} \; B_{1(3)})$, $C_1 = (C_{1(1)} \; C_{1(2)} \; C_{1(3)})$, where $A_{1(i)}, B_{1(i)}, C_{1(i)}$ $(1 \le i \le 3)$ are columns of $A_1, B_1, C_1$, respectively. Then

$$M_y = (A_1 y \; B_1 y \; C_1 y)^t$$

$$= \left(aA_{1(1)} + bA_{1(2)} + cA_{1(3)} \;\; aB_{1(1)} + bB_{1(2)} + cB_{1(3)} \;\; aC_{1(1)} + bC_{1(2)} + cC_{1(3)}\right)^t$$

$$= a\left(A_{1(1)} \; B_{1(1)} \; C_{1(1)}\right)^t + b\left(A_{1(2)} \; B_{1(2)} \; C_{1(2)}\right)^t + c\left(A_{1(3)} \; B_{1(3)} \; C_{1(3)}\right)^t$$

$= aA_2 + bB_2 + cC_2,$

where $A_2 = (A_{1(1)}\ B_{1(1)}\ C_{1(1)})^t$, $B_2 = (A_{1(2)}\ B_{1(2)}\ C_{1(2)})^t$, $C_2 = (A_{1(3)}\ B_{1(3)}\ C_{1(3)})^t$. Thus $\{A_2\ B_2\ C_2\}$ is a basis of $S$.

In view of corollary 2, we may choose appropriate matrices $Q$ and $R$ to make one of $A_2, B_2, C_2$ to be the identity matrix $I$. If $QA_2R = I, QB_2R = A,$ and $QC_2R = B$, then $\{I,\ A,\ B\}$ is a basis of $S$.

Now we are ready to prove the conjecture.

## Case One: A and B are commutative

**Lemma 1.** *Let $T$ be the set of matrices in $M_{3\times3}(\mathbb{F}_p)$ that commute with $A$, then $T = span\{I, A, A^2\}$.*

*Proof.* Suppose $K \in$ span $\{I, A, A^2\}$, then $K = aI + bA + cA^2$ for some $a, b, c \in \mathbb{F}_p$.

$$\because KA = (aI + bA + cA^2)A = aA + bA^2 + cA^3$$

$$AK = (aI + bA + cA^2)A = aA + bA^2 + cA^3$$

$$\therefore KA = AK, \text{and K commutes with A.}$$

Thus span $\{I, A, A^2\} \subseteq T$.

To show they are actually equal, we need to prove span $\{I, A, A^2\}$ and $T$ have the same dimention over $\mathbb{F}_p$. Since $f(t)$ splits in $\mathbb{F}_{p^3}$, $A$ is diagonalizable in $\mathbb{F}_{p^3}$, and there exists an invertible matrix $Q \in M_{3\times3}(\mathbb{F}_{p^3})$ such that $Q^{-1}AQ$ is a diagonal matrix. Let $T'$ be the set of matrices in $M_{3\times3}(\mathbb{F}_{p^3})$ that commute with A. Using the fact that $A, B$ are simultaneously diagonalizable if $A, B \in M_{3\times3}(\mathbb{F}_{p^3})$ commute, we can write $T'$ as

$$T' = \{B \mid Q^{-1}BQ \text{ is a diagonal matrix in } M_{3\times3}(\mathbb{F}_{p^3})\}.$$

Since the set of diagonal matrices in $M_{3\times3}(\mathbb{F}_{p^3})$ is

$$\text{span } \{ \left(\begin{smallmatrix}1&0&0\\0&0&0\\0&0&0\end{smallmatrix}\right)\ \left(\begin{smallmatrix}0&0&0\\0&1&0\\0&0&0\end{smallmatrix}\right)\ \left(\begin{smallmatrix}0&0&0\\0&0&0\\0&0&1\end{smallmatrix}\right) \} \text{ over } \mathbb{F}_{p^3},$$

it follows that

$$T' = \text{span } \{ Q^{-1}\left(\begin{smallmatrix}1&0&0\\0&0&0\\0&0&0\end{smallmatrix}\right)Q\ \ Q^{-1}\left(\begin{smallmatrix}0&0&0\\0&1&0\\0&0&0\end{smallmatrix}\right)Q\ \ Q^{-1}\left(\begin{smallmatrix}0&0&0\\0&0&0\\0&0&1\end{smallmatrix}\right)Q \} \text{ over } \mathbb{F}_{p^3}.$$

Hence, $T'$ has dimension 3 over $\mathbb{F}_{p^3}$. This indicates that $T' =$ span $\{I, A, A^2\}$ over $\mathbb{F}_{p^3}$. Recall $\mathbb{F}_p$ is a subfield of $\mathbb{F}_{p^3}$, so $T =$ span $\{I, A, A^2\}$ over $\mathbb{F}_p$, where $T$ is a subspace of $T'$. Q.E.D.

**Theorem I.4.1.** *Let $\{I,\ A,\ B\}$ be a basis of $S$, and let $\mathbb{F}_{pI}$ be the field $\{dI \mid d \in \mathbb{F}_p$, $I$ is the identity matrix$\}$. If $S$ satisfies the Erdös box condition and $A$, $B$ commute, then $S$ is a cubic extension of the field $\mathbb{F}_{pI}$, excluding the zero matrix.*

*Proof.* Since $\{I, A, B\}$ is a basis of S, matrices $I, A, B$ are linearly independent. So $A$ is not a multiple of $I$. Then corollary 3 tells us that $A$ has an irreducible characteristic polynomial over $\mathbb{F}_p$. Let the characteristic polynomial of $A$ be

$$f(t) = -t^3 + \alpha t^2 + \beta t + \gamma,$$

where $\alpha$, $\beta$, $\gamma \in \mathbb{F}_p$. By Caley-Hamilton Theorem [3], $A$ satisfies $f(t)$, i.e.

$$f(A) = -A^3 + \alpha A^2 + \beta A + \gamma I = 0.$$

Therefore span $\{I, A, A^2\}$ is a cubic extention of $\mathbb{F}_{pI}$ generated by the element $A$.

Since $A$ and $B$ commute, we know that $B \in$ span $\{I, A, A^2\}$ by lemma 1. Thus span $\{I, A, B\} \subseteq$ span $\{I, A, A^2\}$. Because the two vector spaces have the same dimension over $\mathbb{F}_p$, we conclude that

$$\text{span } \{I, A, B\} = \text{span}\{I, A, A^2\}.$$

So S is a cubic extension of the field $\mathbb{F}_{pI}$, excluding the zero matrix.

Q.E.D.

Theorem I.4.1 asserts that when $A$ and $B$ commute, $S$ is the same as that in the KKM example. That is to say, $S = \{\ M_y \mid\ M_y$ is the linear operator on $\mathbb{F}_p^3$ induced by mulitiplication in $\mathbb{F}_{p^3}$ by $y$, where $y \in \mathbb{F}_{p^3} \setminus \{0\}\}$.

## Case Two: A and B are not commutative

Before we discuss the general case, let's look at a special case first.

**Proposition.(Haile's Special Case).** *Suppose $P \equiv 1$ (mod 3). Let $\alpha$ and $\beta$ be two elements in $\mathbb{F}_p$ that are not cubes. Let $A$ and $B$ be $A^3 = \alpha I$ and $B^3 = \beta I$, respectively. Moreover, $AB = wBA$, where $w$ is a cubic root of 1 and $w \neq 1$. Then $S$ does not satisfy the Erdös box condition.*

*Proof.* To show S does not satisfy the box condition, we need to prove there exist $a$ and $b \in \mathbb{F}_{p^3}$ such that $aA+bB$ has a reducible chracteristic polynomial over $\mathbb{F}_p$. Using the condition $AB = wBA$ and $w^2+w+1 = 0$, we find that

$$\begin{aligned}
(aA + bB)^3 &= a^3A^3 + a^2b[(w^2 + w + 1)BA^2] + \\
&\quad ab^2[(w^2 + w + 1)B^2A) + B^2A] + \\
&\quad b^3B^3 \\
&= a^3A^3 + b^3B^3 \\
&= a^3(\alpha I) + b^3(\beta I) \\
&= (a^3\alpha + b^3\beta)I.
\end{aligned}$$

If we could prove $a^3\alpha + b^3\beta$ is a cube in $\mathbb{F}_p$, then $aA + bB$ has a reducible characteristic polynomial.

Suppose $g$ is a cyclic generator of the field $\mathbb{F}_p$ such that $\mathbb{F}_p = \{g, g^2, \ldots, g^{p-1}\}$. Let

$Q_1 = \{g^1, g^4, \ldots, g^{k_1}, \ldots, g^{p-3}\}$, where $k_1 = 3n - 2, 1 \leq n \leq \frac{p-1}{3}$;

$Q_2 = \{g^2, g^5, \ldots, g^{k_2}, \ldots, g^{p-2}\}$, where $k_2 = 3n - 1, 1 \leq n \leq \frac{p-1}{3}$;

$Q_3 = \{g^3, g^6, \ldots, g^{k_3}, \ldots, g^{p-1}\}$, where $k_3 = 3n, \qquad 1 \leq n \leq \frac{p-1}{3}$.

We see that $Q_3$ is the set of cubes in $\mathbb{F}_p$, and $Q_1 = g^2Q_3$, $Q_2 = gQ_3$. Since $\alpha$,$\beta \in Q_1 \cup Q_2$, and $a^3$, $b^3 \in Q_3$, we have $a^3\alpha$ , $b^3\beta \in Q_1 \cup Q_2$. This is because multiplying any element in $Q_1$ (or $Q_2$) by a cube yields another element in $Q_1$ (or $Q_2$).Let $\alpha' = a^3\alpha$ and $\beta' = b^3\beta$. If we can find $\alpha'$, $\beta' \in Q_1 \cup Q_2$ such that $\alpha' + \beta' \in Q_3$, then $aA + bB$ has a reducible characteristic polynomial. There are two cases.

**1.** $P \equiv 1 \pmod 3$ but $P \not\equiv 1 \pmod 9$

Since $p-1$ is divisible by 3 and $1 = g^{p-1}$, the cubic roots of 1 are $g^{\frac{p-1}{3}}$, $g^{\frac{2(p-1)}{3}}$, and 1. However, $g^{\frac{p-1}{3}}$ and $g^{\frac{2(p-1)}{3}}$ are not cubes, because $p-1$ is not divisible by 9. Thus $g^{\frac{p-1}{3}}$ , $g^{\frac{2(p-1)}{3}} \in Q_1 \cup Q_2$. Choose $\alpha' = g^{\frac{p-1}{3}}$ and $\beta' = g^{\frac{2(p-1)}{3}}$, we get

$$t^3 - 1 = (t - \alpha')(t - \beta')(t - 1)$$

Thus $\alpha' + \beta' + 1 = 0$ or $\alpha' + \beta' = -1 \in Q_3$.

**2.** $P \equiv 1 \pmod 3$ and $P \equiv 1 \pmod 9$

Now $p-1$ is divisible by 9. So $g^{\frac{p-1}{3}}$ and $g^{\frac{2(p-1)}{3}}$ are cubes. Then $(Q_3+Q_3)\cap Q_3 \neq \phi$. Proof by contradiction. Assume there does not exist $\alpha', \beta' \in Q_1 \cup Q_2$ such that $\alpha' + \beta' \in Q_3$. Then consider $Q_1 \times Q_2 = \{(\alpha', \beta') \mid \alpha' \in Q_1 \ , \ \beta' \in Q_2\}$. Divide $Q_1 \times Q_2$ into two subsets $J_1$ and $J_2$ such that

$$J_1 = \{(\alpha', \beta') \mid \alpha' + \beta' \in Q_1\}$$
$$J_2 = \{(\alpha', \beta') \mid \alpha' + \beta' \in Q_2\}$$

In subset $J_1$, we have

$$\alpha' + \beta' = \gamma' \in Q_1 \tag{I.5}$$

for any $(\alpha', \beta') \in J_1$. Multiply both sides of (I.5) by $g^2$, we get

$$g^2\alpha' + g^2\beta' = g^2\gamma'$$
$$g^2\alpha' - g^2\gamma' = -g^2\beta'$$

where $g^2\alpha', -g^2\gamma' \in Q_3$, and $-g^2\beta' \in Q_1$. Thus we get a pair $(g^2\alpha', -g^2\gamma') \in Q_3 \times Q_3$, whose sum is in $Q_1$. Because the number of pairs $(\alpha', -\gamma') \in J_1$ is $|J_1|$, there are totally $|J_1|$ pairs of $(g^2\alpha', -g^2\gamma') \in Q_3 \times Q_3$, whose sum is in $Q_1$.

Similarly, in subset $J_2$, we have

$$\alpha' + \beta' = \gamma' \in Q_2 \tag{I.6}$$

for any $(\alpha', \beta') \in J_2$. Multiply both sides of (I.6) by $g$, we get

$$g\alpha' + g\beta' = g\gamma'$$
$$g\beta' - g\gamma' = -g\alpha',$$

where $g\beta', -g\gamma' \in Q_3$, and $-g\alpha' \in Q_2$. Thus we get a pair $(g\beta', -g\gamma') \in Q_3 \times Q_3$, whose sum is in $Q_2$. Because the number of pairs $(\beta', -\gamma') \in J_2$ is $|J_2|$, there are totally $|J_2|$ pairs of $(g\beta', -g\gamma') \in Q_3 \times Q_3$, whose sum is in $Q_2$.

Consequently, $|J_1| + |J_2| = |Q_1 \times Q_2| = |Q_3 \times Q_3|$ implies that for each pair of elements in $Q_3 \times Q_3$, their sum is either in $|Q_1|$ or $|Q_2|$. Thus $Q_3 + Q_3 \subseteq Q_1 \cup Q_2$. This contradicts with $Q_3 + Q_3 \cap Q_3 \neq \phi$.

Q.E.D.

## I.5 Conclusion

In Haile's special case, we've proved that if $A$ and $B$ do not commute, $S$ does not satisfy the Erdös box condition. In the general case, we are trying to prove that this is also true. Using Dickson's Theorem [4], we are able to characterize $A$ and $B$ such that $aA + bB$ has an irreducible characteristic polynomial for any $a, b \in \mathbb{F}_p$. However, we feel that $A$ and $B$ will not satisfy the condition that $(M_{y_1} {M_{y_2}}^{-1})^t$ for any $M_{y_1} \neq M_{y_2} \in$ span $\{I, A, B\}$. If this feeling is correct, the conjecture that M is uniquely defined by the KKM example is true.

## I.6 Acknowledgement

Many thanks to Dr. Katz for his patient teaching and support.Also thank Kevin Pilgrim and Mandie McCarty for organizing everything.

## Bibliography

1. D.Gunderson, V. Rodl, and A. Sidorenko *Extremal problems for sets forming Boolean algebras and complete partite hypergraphs*, J.Combin.Theory Ser. A 88 (1999), 342-367

2. N.Katz, E.Krop, and M. Maggioni *Remarks on the Box Problem*, Mathematical Research Letters 9 (2002), 515-519

3. S.Friedberg, A. Insel, and L. Spence *Linear Algebra*, Pearson Education, New Jersey,2003

4. L. Carlitz *A Theorem of Dickson on Nonvanishing Cubic Forms in a Finite Field*, Proceedings of American Mathematical Society, Vol.8, No.5, 1957, 975-977