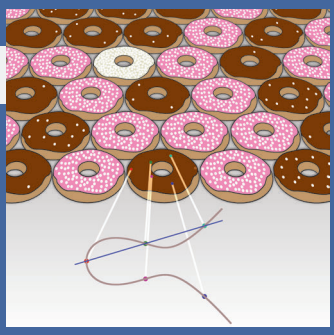
AMERICAN MATHEMATICAL SOCIETY

CURRENT EVENTS BULLETIN

Friday, January 11, 2013, 1:00 PM to 5:00 PM

Room 6F, Upper Level, San Diego Convention Center
Joint Mathematics Meetings, San Diego, CA



1:00 PM Wei Ho

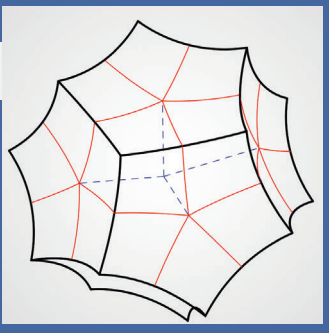
How many rational points does a random curve have?

Bhargava and his school have turned classical questions about cubic polynomial equations in two variables over the integers in a whole new direction.

2:00 PM Sam Payne

Topology of nonarchimedean analytic spaces

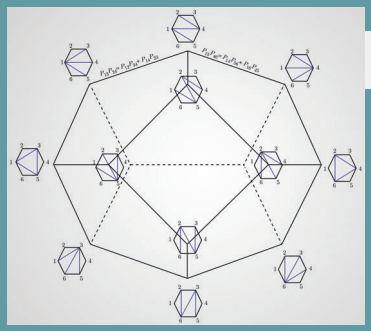
"Tropical" algebra and geometry is a burgeoning field, suggesting interesting paths into geometry over fields like the p-adic numbers.



3:00 PM Mladen Bestvina

Geometric group theory and 3-manifolds hand in hand: the fulfillment of Thurston's vision for three-manifolds

Another vindication of Thurston's fabulous insights into three-dimensional geometry comes with a wonderful group-theoretic construction.



4:00 PM Lauren Williams

Cluster algebras

An exotic but simple structure that depends on a finite directed graph was brought to light in representation theory by Fomin and Zelevinsky. It has now been shown to play an important role in subjects as diverse as Poisson geometry and the theory of triangulations of surfaces as well.

Organized by David Eisenbud, University of California, Berkeley

125th Anniversary
AMERICAN MATHEMATICAL SOCIETY
AMS
Advancing Mathematics Since 1888

Introduction to the Current Events Bulletin

Will the Riemann Hypothesis be proved this week? What is the Geometric Langlands Conjecture about? How could you best exploit a stream of data flowing by too fast to capture? I think we mathematicians are provoked to ask such questions by our sense that underneath the vastness of mathematics is a fundamental unity allowing us to look into many different corners -- though we couldn't possibly work in all of them. I love the idea of having an expert explain such things to me in a brief, accessible way. And I, like most of us, love common-room gossip.

The Current Events Bulletin Session at the Joint Mathematics Meetings, begun in 2003, is an event where the speakers do not report on their own work, but survey some of the most interesting current developments in mathematics, pure and applied. The wonderful tradition of the Bourbaki Seminar is an inspiration, but we aim for more accessible treatments and a wider range of subjects. I've been the organizer of these sessions since they started, but a varying, broadly constituted advisory committee helps select the topics and speakers. Excellence in exposition is a prime consideration.

A written exposition greatly increases the number of people who can enjoy the product of the sessions, so speakers are asked to do the hard work of producing such articles. These are made into a booklet distributed at the meeting. Speakers are then invited to submit papers based on them to the *Bulletin of the AMS*, and this has led to many fine publications.

I hope you'll enjoy the papers produced from these sessions, but there's nothing like being at the talks -- don't miss them!

David Eisenbud, Organizer
University of California, Berkeley
de@msri.org

For PDF files of talks given in prior years, see
http://www.ams.org/ams/current-events-bulletin.html.
The list of speakers/titles from prior years may be found at the end of this booklet.

HOW MANY RATIONAL POINTS DOES A RANDOM CURVE HAVE?

WEI HO

Abstract. A large part of modern arithmetic geometry is dedicated to or motivated by the study of rational points on varieties. For an elliptic curve over $\mathbb{Q}$, the set of rational points forms a finitely generated abelian group. The ranks of these groups, when ranging over all elliptic curves, are conjectured to be evenly distributed between rank 0 and rank 1, with higher ranks being negligible. We will describe these conjectures and discuss some results on bounds for average rank, highlighting the recent work of Bhargava and Shankar.

Contents

1. Rational points on varieties

Finding solutions to polynomial equations is one of the oldest problems in mathematics. Over the last few centuries, mathematicians have formalized the questions and established rigorous language to discuss this simple idea in different variations. While we have made tremendous strides in understanding the structure of these solutions in the last few decades, there remain many fundamental open questions, which lie at the forefront of modern arithmetic geometry.

2010 *Mathematics Subject Classification.* Primary 11G05, 14H52. Secondary 11G30, 14H25.
Partially supported by NSF grant DMS-0902853.

For the simplest case, let $f(x_1, \ldots, x_n)$ be a polynomial with coefficients in $\mathbb{Q}$. We may ask for rational solutions to $f = 0$, i.e., numbers $a_1, \ldots, a_n \in \mathbb{Q}$ such that $f(a_1, \ldots, a_n) = 0$. To phrase the question more geometrically, let X be the **variety** associated to f, which may be viewed geometrically as the *zero locus* of f, or solutions to $f = 0$, in $\mathbb{C}^n$. Then X will be $(n-1)$-dimensional, if f is not identically zero. Our problem may be restated as finding **rational points** on X, the set of which is denoted $X(\mathbb{Q})$. More specifically, we may ask questions such as the following:

- Does there exist a single rational point on X?
- Can we describe all rational points on X?
- If there are only finitely many rational points, can we enumerate them?
- If $X(\mathbb{Q})$ is an infinite set, what structure does $X(\mathbb{Q})$ have?

If we instead use any finite number of polynomials $f_1, \ldots, f_m \in \mathbb{Q}[x_1, \ldots, x_n]$, we define the analogous variety X to be the common zero locus of all of the polynomials f_i in $\mathbb{C}^n$. For general choices of f_i, the dimension of X will be $n-m$, if nonnegative, and 0 otherwise; the rule of thumb is that each polynomial condition imposed should reduce the dimension of X by 1.

Remark 1.1. While we restrict our attention to varieties defined over $\mathbb{Q}$, i.e., defined by polynomials with coefficients in $\mathbb{Q}$, many of the results that we will discuss have analogues over other number fields.

Even when X is 1-dimensional, mathematicians have not yet fully understood how many rational points are on X! In this note, we focus on this case, where X is a **curve**.

1.1. **Genus and the trichotomy.** The arithmetic and the geometry of algebraic curves rely heavily on an invariant called the **genus**. The genus of a curve may be defined in many ways, but the most intuitive definition is topological. A smooth curve X as defined above may be thought of as a **Riemann surface**[1] with finitely many punctures; after taking an appropriate compactification by filling these punctures, the resulting compact Riemann surface has a topological genus, which is essentially the number of "holes" or "handles." For example, a complex curve that is homeomorphic to a sphere (after compactification) has genus 0, while a genus 1 curve looks like the surface of a donut.

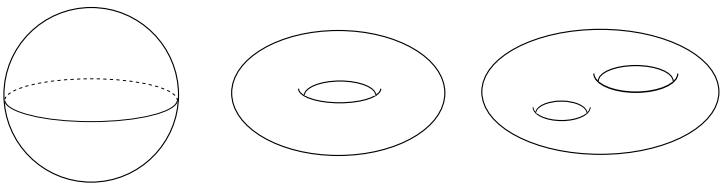


FIGURE 1. From left to right: curves of genus 0, 1, and 2.

For simplicity, we assume in the sequel that our curves X have been compactified[2], have no singularities[3], and are connected.

[1]The complex points of a smooth curve form a *two*-dimensional real manifold.

[2]In other words, we always implicitly work with projective curves.

[3]Intuitively, a *singularity* on a curve is a point where it is not smooth, like a node or a cusp. More precisely, it is a point with more than one tangent direction along the curve.

Geometric and arithmetic properties of curves are heavily influenced by their genera. For example, we have the following trichotomy:

	genus 0	genus 1	genus ≥ 2
canonical bundle	anti-ample	trivial	ample
curvature	positive	zero	negative
Kodaira dimension	$\kappa = -\infty$	$\kappa = 0$	$\kappa = 1$
automorphism group	3-dimensional	1-dimensional	finite
rational points	Hasse principle	finitely generated	finitely many

We now discuss the last row in more detail.

Curves of genus 0. For genus 0 curves, there are either no rational points at all or infinitely many, and it is fairly easy to determine which case applies to any given curve by the **Hasse principle**.

In particular, a genus 0 curve X has a rational point if and only if it has a point everywhere *locally*, which means that the equations defining X have a solution over the real numbers $\mathbb{R}$ and the p-adic numbers $\mathbb{Q}_p$ for all primes p. The non-existence of a $\mathbb{Q}_p$-point is always due to an *obstruction* modulo a power of the prime p.

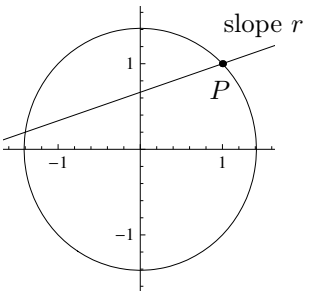
Example 1.2. Let X be the curve given by the vanishing of the polynomial $f = x^2 + y^2 - 3$. If there exists a rational solution to $f = 0$, by clearing denominators, there are relatively prime integers r, s, and t such that $r^2 + s^2 = 3t^2$. Because squares of integers are congruent to 0 or 1 modulo 4, reducing the equation modulo 4 shows that r^2 and s^2 are both congruent to 0 modulo 4. This in turn implies that all three integers are even, which is a contradiction. Therefore, the equation $f = 0$ has an obstruction modulo 4, implying that X has no point over $\mathbb{Q}_2$ and thus no rational point.

In fact, checking for local obstructions may be completed in a finite number of steps. Any curve of genus 0 over $\mathbb{Q}$ is isomorphic to a (compactified) plane conic defined by the vanishing of a polynomial of the form

$$ax^2 + by^2 + c, \tag{1}$$

where a, b, and c are squarefree, pairwise relatively prime integers. A theorem of Legendre implies that (1) has a rational solution if and only if a, b, and c do not all have the same sign, and $-ab$ is a square modulo c, $-bc$ is a square modulo a, and $-ac$ is a square modulo b.

If there is a single rational point P on a conic, then all other rational points come from intersecting the conic with a line of rational slope through P.



Example 1.3. If X is given by $x^2 + y^2 - 2 = 0$, then by inspection, the point $(x, y) = (1, 1)$ lies on X. All other points are parametrized by drawing lines of rational (or infinite) slope r through $(1, 1)$, and a simple computation shows that $X(\mathbb{Q})$ is the union of the point $(1, -1)$ and the points

$$\left(\frac{r^2 - 2r - 1}{r^2 + 1}, \frac{-r^2 - 2r + 1}{r^2 + 1}\right)$$

for all $r \in \mathbb{Q}$.

Curves of genus at least 2. Mordell's 1922 conjecture [Mor22] predicted that there could not be very many rational points on a curve of genus ≥ 2; it was proved by Faltings [Fal83] (as a corollary to an even more powerful theorem):

Theorem 1.4 (Faltings 1983)**.** *Let X be a curve of genus at least* 2 *over $\mathbb{Q}$. Then the set $X(\mathbb{Q})$ of rational points is finite.*

The original proof uses deep ideas from p-adic Hodge theory, Arakelov theory, and moduli theory, and later proofs and improvements by Vojta [Voj91], Faltings [Fal91], Bombieri [Bom90], and others use Diophantine approximation methods. None of the proofs, however, are *effective* in the sense of giving a list of the points in $X(\mathbb{Q})$. In practice, a combination of techniques — including Chabauty's method, Brauer-Manin obstructions, and descent — often are enough to produce the points in $X(\mathbb{Q})$. In §4.3, we will outline recent progress on bounding the number of rational points on curves of genus ≥ 2.

Curves of genus 1. The case of genus 1 curves is the richest arithmetically, the most complicated, and the most mysterious to this day. A genus one curve defined over $\mathbb{Q}$ may have no rational points at all, finitely many, or infinitely many — and it is generally difficult to determine which! Techniques for other genera, like the Hasse principle, no longer apply, e.g., there are plenty of genus one curves which have points everywhere locally but no global rational point.

Genus one curves over $\mathbb{Q}$ with a given rational point are known as **elliptic curves**. Section 1.2 will describe the structure of rational points on elliptic curves in more detail.

1.2. **Rational points on elliptic curves.** An elliptic curve over $\mathbb{Q}$ is isomorphic to the projective closure of the zero locus of a Weierstrass equation

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6 \tag{2}$$

with all $a_i \in \mathbb{Q}$. When defining a nonsingular curve, the equation (2) may be transformed (over $\mathbb{Q}$) into **short Weierstrass form**

$$y^2 = x^3 + Ax + B \tag{3}$$

for $A, B \in \mathbb{Q}$ with nonzero discriminant $\Delta = -16(4A^3 + 27B^2)$; the nonvanishing of the discriminant ensures that the curve is nonsingular. There is a marked rational point "at infinity," which is denoted O. We say that the elliptic curve given by (3) has **height**

$$\mathrm{ht}(E) := \max(4|A|^3, 27B^2).$$

The coefficients of 4 and 27 are for convenience; only the exponents matter for most purposes.

Many such equations define isomorphic elliptic curves. In particular, for $t \in \mathbb{Q}^\times$, scaling x and y by t^{-2} and t^{-3}, respectively, in equation (3) gives the new equation

$$y^2 = x^3 + t^4A + t^6B.$$

In other words, the group $\mathbb{Q}^\times$ acts on the space of all equations of the form (3) with nonzero discriminant. To choose one representative equation from each $\mathbb{Q}^\times$-orbit, we define **minimal Weierstrass** equations to be those of the form (3), with $A, B \in \mathbb{Z}$ and the condition that there is no prime p such that p^4 divides A and p^6 divides B. Each elliptic curve over $\mathbb{Q}$ has a unique minimal Weierstrass model, and we will call its discriminant the **minimal** discriminant of the elliptic curve.

FIGURE 2. The real points of the elliptic curves $y^2 = x^3 - x + 1$ (left) and $y^2 = x^3 - x$ (right).

The solutions of a Weierstrass equation lying in any field have a rich structure. The complex points of an elliptic curve make up a one-holed torus, as discussed earlier. The real points are smooth curves in $\mathbb{R}^2$ with one or two components; see Figure 2.

Group Law. A beautiful and incredibly useful fact is that the set of solutions with values in any given field forms a group! An even more powerful statement for rational points is given by a theorem of Mordell [Mor22]:

Theorem 1.5 (Mordell 1922)**.** *The set $E(\mathbb{Q})$ of rational points of an elliptic curve E defined over $\mathbb{Q}$ forms a finitely generated abelian group, i.e.,*

$$E(\mathbb{Q}) = \mathbb{Z}^r \oplus E(\mathbb{Q})_{\text{tors}} \tag{4}$$

for some nonnegative integer r and finite abelian group $E(\mathbb{Q})_{\text{tors}}$.

The group structure on the points of an elliptic curve uses the point O at infinity as the identity element, and it is most easily described geometrically. For the graph of an elliptic curve in short Weierstrass form, as seen in Figure 3, the line L through any two points P_1 and P_2 will intersect a third point P_3 by Bezout's theorem. The vertical line through P_3 intersects another point on the elliptic curve, which is the composition $P_1 + P_2$ of P_1 and P_2.

In other words, the three (not necessarily distinct) intersection points P_1, P_2, and P_3 of any line L with the elliptic curve sum to the identity in the group law. The identity point O may be one of these points, e.g., a vertical line intersects O, a point P, and its negative. Moreover, if P_1 and P_2 are rational points, then the line L has rational slope, so $P_1 + P_2$ is also a rational point.

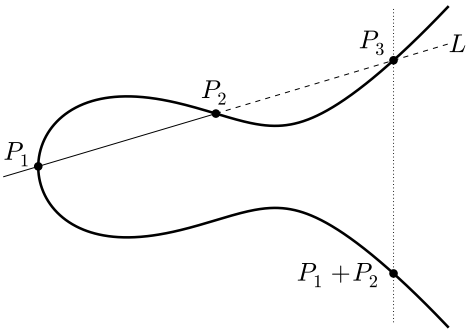


FIGURE 3. The group law on an elliptic curve.

For an elliptic curve E over $\mathbb{Q}$, the torsion subgroup $E(\mathbb{Q})_{\text{tors}}$ of $E(\mathbb{Q})$ is fairly well understood, by a deep theorem of Mazur [Maz77]:

Theorem 1.6 (Mazur 1977)**.** *For an elliptic curve E defined over $\mathbb{Q}$, the torsion subgroup $E(\mathbb{Q})_{\text{tors}}$ is one of the following groups:*

$$\begin{aligned} &\mathbb{Z}/d\mathbb{Z} && \text{for } 1 \le d \le 10 \text{ or } d = 12 \\ &\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2d\mathbb{Z} && \text{for } 1 \le d \le 4. \end{aligned}$$

By Theorem 1.6 and Hilbert's irreducibility theorem, almost all[4] elliptic curves have a trivial torsion subgroup. Moreover, there are explicit methods to quickly compute the torsion subgroup for any given elliptic curve. For example, the rational 2-torsion points of an elliptic curve in short Weierstrass form are determined by factoring the right hand side cubic polynomial of (3) over $\mathbb{Q}$.

Example 1.7. The elliptic curve $y^2 = x(x-1)(x-2)$ has rational 2-torsion subgroup $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$, consisting of the points $(0,0)$, $(1,0)$, $(2,0)$, and O (as the identity). The elliptic curve $y^2 = x(x^2+x+1)$ has only the rational 2-torsion points $(0,0)$ and O.

In contrast, the **rank** of $E(\mathbb{Q})$, denoted r in (4), is much more mysterious! It is not even known if the rank can be arbitrarily large; the current record — due to Elkies — is an elliptic curve of rank at least 28. It is quite difficult in general to rigorously prove that a given elliptic curve has a certain rank, though L-function computations often give conjectural answers, as we describe very briefly below.

Analytic rank. The **L-function** $L(E,s)$ of an elliptic curve E over $\mathbb{Q}$ is a Dirichlet series given by an Euler product formula involving the number N_p of $\mathbb{F}_p$-points on the reduction of E over $\mathbb{F}_p$, for all primes p:

$$L(E,s) := \prod_{\text{good } p} \frac{1}{1 - a_p p^{-s} + p^{1-2s}} \prod_{\text{bad } p} \frac{1}{1 - a_p p^{-s}} = \sum_{n \ge 1} a_n n^{-s}$$

where $a_p = 1 + p - N_p$ for "good" primes p and $a_p = -1$, 0, or 1 for a finite set of "bad" primes p. By the work of Wiles and others [Wil95, TW95, BCDT01], the L-function extends to an entire function on the complex plane, and $\Lambda(E,s) := \text{cond}(E)^{s/2}(2\pi)^{-s}\Gamma(s)L(E,s)$ satisfies a functional equation

$$\Lambda(E,s) = u_E \Lambda(E, 2-s),$$

where $\text{cond}(E)$ is an invariant of E called the **conductor**. The **root number** u_E is either $+1$ or -1.

The order of vanishing of this L-function $L(E,s)$ at $s = 1$ is the **analytic rank** of E. The Birch and Swinnerton-Dyer (BSD) Conjecture [BSD63, BSD65] claims that the analytic rank of E is equal to the rank of E defined earlier. A more refined version in fact gives a formula for the coefficient of the leading term in the Taylor expansion at $s = 1$ in terms of arithmetic invariants of E, including the Tate–Shafarevich group $Ш(E)$, which will be discussed in §2.3.

Thus, assuming the BSD conjecture, one may study the rank of an elliptic curve by understanding its analytic rank, e.g., by computing the apparent order of vanishing of $L(E,s)$ at $s = 1$.

[4]Here, "almost all" means that the density of curves, as defined in §2.1, with trivial torsion subgroup is 1, when ordered by height.

2. Ranks of elliptic curves

Most of the remainder of this note is devoted to conjectures and results on the *distribution* of ranks for elliptic curves. In other words, if we choose a "random" elliptic curve, what do we expect its rank to be?

2.1. **Densities and averages.** In order to formulate our question rigorously, we need to specify what is meant by "random" in this setting. We first need an ordering for the (infinite) set of elliptic curves, usually by some sort of invariant; possibilities include ordering elliptic curves up to isomorphism by conductor or by minimal discriminant, or ordering short Weierstrass equations by height or discriminant.

In all these cases, there are only finitely many objects (elliptic curves up to isomorphism or short Weierstrass equations with integral coefficients, for example) with the absolute value of that invariant bounded by any positive number X. When the invariant is the discriminant or the conductor, this finiteness is due to Siegel's classical theorem on the finiteness of S-integral points on elliptic curves [Sie66]. We may thus define quotients like the following:

$$\mathrm{P}(\text{invariant}, X, \mathrm{rk} = i) := \frac{\#\{\text{elliptic curves } E \text{ with invariant } \leq X \text{ and rank } i\}}{\#\{\text{elliptic curves } E \text{ with invariant } \leq X\}}$$

where the invariant could be the conductor, absolute value of the discriminant, or the height, for example. Then we might ask whether this quantity converges as X tends to infinity, and if so, we may consider the limit

$$\mathrm{P}(\text{invariant}, \mathrm{rk} = i) := \lim_{X\to\infty} \mathrm{P}(\text{invariant}, X, \mathrm{rk}(E) = i)$$

as the **density** of elliptic curves, ordered by that invariant, with rank i (or equivalently, the probability an elliptic curve has rank i). We may define a lower density or upper density by instead using liminf or limsup, respectively.

We then define the **average rank** for elliptic curves, ordered by an invariant, as

$$\lim_{X\to\infty} \sum_{i\geq 0} i \cdot \mathrm{P}(\text{invariant}, X, \mathrm{rk} = i) = \lim_{X\to\infty} \frac{\sum_{\text{invariant}(E)\leq X} \mathrm{rk}(E)}{\sum_{\text{invariant}(E)\leq X} 1}$$

if this limit exists. Again, a lower average or an upper average is defined using liminf or limsup, respectively; we will sometimes call these the limsup and the liminf of the average rank.

We may also define averages or higher moments for distributions of other quantities associated to elliptic curves in an analogous way.

2.2. **The Minimalist Conjecture.** The basic conjecture for the distribution of ranks of elliptic curves is based on the philosophy that elliptic curves should not have any more points than they must.

The widely believed Parity Conjecture, which is a consequence of a weak form of the BSD conjecture, asserts that the parity of the rank of an elliptic curve is equal to the parity of the analytic rank, which is even exactly when $u_E = +1$. It is also strongly expected that the root numbers u_E have probability $1/2$ of being $+1$, and probability $1/2$ of being -1. We are therefore led to the following:

Minimalist Conjecture. *The densities of elliptic curves having rank* 0 *and having rank* 1 *are both exactly* $1/2$.

Although no ordering is specified in the statement above, it is conjectured for any reasonable ordering, such as the examples given in §2.1. Note that the Minimalist Conjecture implies that the density of rank i elliptic curves, for $i \geq 2$, is 0.

The first version of the conjecture was stated in the 1979 work of Goldfeld [Gol79] for quadratic twist families of elliptic curves. Given an elliptic curve E in short Weierstrass form (3), define its quadratic twist E_d by a nonzero squarefree integer d as the elliptic curve $y^2 = x^3 + d^2Ax + d^3B$. Then Goldfeld's conjecture [Gol79] asserts that for a fixed elliptic curve E, the average rank of the elliptic curves E_d is $1/2$, when ordered by $|d|$, that is,

$$\lim_{X\to\infty} \frac{\sum_{|d|\leq X} \mathrm{rk}(E)}{\sum_{|d|\leq X} 1} = \frac{1}{2},$$

where the sums are over nonzero squarefree integers d. Much work has been done towards this conjecture for quadratic twist families; see Silverberg's survey [Sil07].

The Minimalist Conjecture for all elliptic curves and for quadratic twist families is also supported by the philosophy of Katz–Sarnak [KS99] and later random matrix theory computations and heuristics of Keating–Snaith [KS00], Conrey–Keating–Rubinstein–Snaith [CKRS02], Watkins [Wat08], and others. See [BMSW07, Poo12] for excellent surveys of many aspects of this conjecture.

At various points since Goldfeld's work, the conjecture has been disbelieved, mostly because computations have not seemed to support it. For example, in [KS99], the data of Kramarz–Zagier [ZK87] (extended later by Watkins [Wat07]) for a special family of elliptic curves is noted to have a large number of higher rank elliptic curves, with the suggestion that computational capabilities were not yet powerful enough to reflect the true distribution.

The more general computations for the family of all elliptic curves ordered by conductor, by Brumer–McGuinness [BM90], Stein–Watkins [SW02], Cremona [Cre06], and Bektemirov–Mazur–Stein–Watkins [BMSW07], also display a surprisingly large percentage of higher rank elliptic curves. As a result, their data imply asymptotics for average ranks that appear significantly higher than $1/2$; see Figure 4. It has

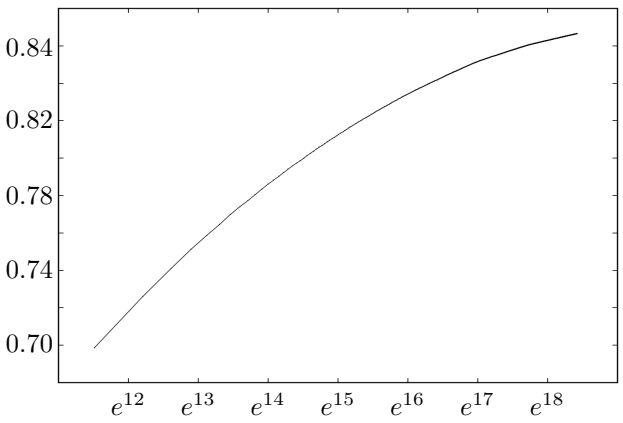


FIGURE 4. Average rank of elliptic curves in the Stein-Watkins database, up to conductor 10^8. Data and graph by Bektemirov–Mazur–Stein–Watkins [BMSW07].

been suggested that when restricted to these computational ranges, the data shows the strong effects of secondary terms for the asymptotic number of curves of given rank up to conductor X.

Theoretical work on this conjecture is perhaps more optimistic. Brumer [Bru92] showed that, assuming the Generalized Riemann Hypothesis (GRH), the limsup of the average analytic rank of all elliptic curves ordered by height is bounded above by 2.3. Thus, also assuming the BSD conjecture, Brumer's result implies that the limsup of the average rank is bounded above by 2.3. This bound was improved by Heath-Brown [HB04] to 2, and then by Young [You06] to 25/14, still assuming BSD and GRH. Young's bound was the first theoretical result implying that a positive proportion of elliptic curves have rank 0 or 1, assuming BSD and GRH.

Finally, the recent work of Bhargava–Shankar [BS10a] gives an unconditional upper bound:

Theorem 2.1 (Bhargava–Shankar 2010)**.** *The limsup of the average rank of elliptic curves over $\mathbb{Q}$, ordered by height, is bounded above by* 3/2.

They consider elliptic curves in short Weierstrass form with integral coefficients, and the theorem holds for both all such Weierstrass equations or only minimal ones. Much of the remainder of this note will focus on this theorem, as well as generalizations and corollaries; see also Poonen's Bourbaki exposé [Poo12] for an excellent detailed exposition of [BS10a].

2.3. **Selmer groups.** Studying the Selmer group for an elliptic curve is one of the only currently known methods to establish an upper bound on its rank. This method is used both for computations for individual curves (e.g., Cremona's `mwrank` program [Cre12]) and results for all curves together (as in Theorem 2.1).

The utility of Selmer groups comes from two facts: first, they are finite and often computable, and second, the size of the Selmer group of an elliptic curve gives an upper bound for its rank. More precisely, for a prime p, the p-Selmer group $\mathrm{Sel}_p(E)$ of an elliptic curve E is an elementary abelian p-group, i.e., isomorphic to the product of a nonnegative number of $\mathbb{Z}/p\mathbb{Z}$'s, and its p-rank $\mathrm{rk}_p(\mathrm{Sel}_p(E))$ is the number of factors of $\mathbb{Z}/p\mathbb{Z}$. Then

$$\mathrm{rk}_p(\mathrm{Sel}_p(E)) \geq \mathrm{rk}(E). \tag{5}$$

Bhargava–Shankar [BS10a] prove Theorem 2.1 by combining (5) for $p = 2$ and the following stronger result:

Theorem 2.2 (Bhargava–Shankar 2010)**.** *The average size of the* 2*-Selmer group for elliptic curves over $\mathbb{Q}$, ordered by height, is* 3.

Definitions. This subsection may be safely skipped at a first reading; it is more important to understand how to actually access elements of the p-Selmer group, as described in the next subsection.

We define the p-Selmer group for an elliptic curve E over $\mathbb{Q}$ and a prime p. One of the motivating ideas behind the definition is that local computations are often much more feasible than global ones; we also saw this idea in action in §1.1 during the discussion on genus zero curves.

Because the points of E over any field form a group, there is a multiplication-by-p map $E(\bar{\mathbb{Q}}) \xrightarrow{p} E(\bar{\mathbb{Q}})$, which is surjective and whose kernel is the p-torsion subgroup

$E(\bar{\mathbb{Q}})[p]$ of $E(\bar{\mathbb{Q}})$. The Galois group $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ acts on each of these groups, so the short exact sequence

$$0 \to E(\bar{\mathbb{Q}})[p] \to E(\bar{\mathbb{Q}}) \to E(\bar{\mathbb{Q}}) \to 0$$

of $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$-modules induces a long exact sequence of Galois cohomology, from which we extract the first row of the commutative diagram (6) below.

$$\begin{array}{ccccccccc} (6) \qquad 0 & \longrightarrow & E(\mathbb{Q})/pE(\mathbb{Q}) & \longrightarrow & H^1(\mathbb{Q}, E[p]) & \xrightarrow{\alpha} & H^1(\mathbb{Q}, E)[p] & \longrightarrow & 0 \\ & & \downarrow & & \downarrow{\scriptstyle \prod \mathrm{Res}_\nu} & \searrow^{\beta} & \downarrow{\scriptstyle \prod \mathrm{Res}_\nu} & & \\ 0 & \to & \prod_\nu E(\mathbb{Q}_\nu)/pE(\mathbb{Q}_\nu) & \to & \prod_\nu H^1(\mathbb{Q}_\nu, E[p]) & \to & \prod_\nu H^1(\mathbb{Q}_\nu, E)[p] & \to & 0 \end{array}$$

The analogous procedure over each local completion $\mathbb{Q}_\nu$ for primes ν (including $\mathbb{Q}_\nu := \mathbb{R}$ for $\nu = \infty$) gives the second row of (6). The leftmost vertical map is given by the inclusions of $E(\mathbb{Q})$ into each $E(\mathbb{Q}_\nu)$, and the latter two vertical maps are products over all primes ν of the usual restriction maps $\mathrm{Res}_\nu : H^1(\mathbb{Q}, A) \to H^1(\mathbb{Q}_\nu, A)$ for $A = E[p]$ and $A = E$. Then we make the following definitions:

(i) The **p-Selmer group** $\mathrm{Sel}_p(E)$ of E is the kernel of the map β in (6).
(ii) The **Tate–Shafarevich group** $Ш(E)$ is the kernel of the map

$$\prod_\nu \mathrm{Res}_\nu : H^1(\mathbb{Q}, E) \to \prod_\nu H^1(\mathbb{Q}_\nu, E).$$

Applying the Snake Lemma to a variant of (6) gives the key exact sequence

$$0 \to E(\mathbb{Q})/pE(\mathbb{Q}) \to \mathrm{Sel}_p(E) \to Ш(E)[p] \to 0, \tag{7}$$

which leads to the inequality (5).

Visualizing elements of Selmer groups. Elements of the Tate–Shafarevich group $Ш(E)$ and the p-Selmer group $\mathrm{Sel}_p(E)$ of an elliptic curve E over $\mathbb{Q}$ may be concretely realized using genus one curves and their Jacobians.

The **Jacobian** $\mathrm{Jac}(C)$ of a genus one curve C is the connected component of its automorphism group[5]. It is also a curve of genus one, and if C has a rational point, then the Jacobian of C is in fact isomorphic to C over $\mathbb{Q}$. As any automorphism group of course has a group structure, including an identity element, the Jacobian of C is an elliptic curve! We call a genus one curve C a **torsor**, or principal homogeneous space, for $\mathrm{Jac}(C)$. In other words, a torsor for an elliptic curve E is a genus one curve with an action of E that is simply transitive over $\bar{\mathbb{Q}}$.

The elements of $H^1(\mathbb{Q}, E)$ may be thought of as isomorphism classes of torsors for E, that is, genus one curves C over $\mathbb{Q}$ whose Jacobians are isomorphic to E. A trivial torsor is isomorphic to E itself, meaning that it has a rational point, and similarly, a torsor C that maps to 0 under Res_ν has a point in $\mathbb{Q}_\nu$.

Therefore, the elements of the Tate–Shafarevich group $Ш(E)$ are exactly those torsors, up to isomorphism, that have points over every local completion $\mathbb{Q}_\nu$, also known as **locally soluble**. The nonzero elements of $Ш(E)$ are locally soluble torsors without a global rational point, implying that they fail the Hasse principle.

[5]This definition only works for curves of genus one. More generally, the Jacobian is defined to be the dual of the moduli space of degree 0 line bundles.

Elements of the p-Selmer group may be represented as locally soluble torsors C for E, along with a degree p line bundle on C (or equivalently, a rational degree p divisor[6] on C). This degree p line bundle on C is equivalent to remembering an algebraic map from C to $(p-1)$-dimensional projective space. As we will describe in more detail in later sections, this description of p-Selmer elements may be made yet more explicit for small values of p.

Here is a table summarizing these interpretations of the groups, for an elliptic curve E over $\mathbb{Q}$:

Group	Elements (up to isomorphism)
$H^1(\mathbb{Q}, E)$	torsors C for E
$Ш(E)$	locally soluble torsors C for E
$\mathrm{Sel}_p(E)$	pairs (C, L): locally soluble torsors C for E with degree p line bundles L on C
$E(\mathbb{Q})/pE(\mathbb{Q})$	pairs (C, L): trivial(ized) torsors C for E with degree p line bundles L on C

Heuristics and other work. In the last several years, several heuristics have been developed for the distributions for the Tate–Shafarevich group and p-Selmer groups for elliptic curves over $\mathbb{Q}$ (and over other number fields).

Delaunay's heuristics [Del01, Del07] for the distribution of Tate–Shafarevich groups generalizes the ideas behind the Cohen–Lenstra–Martinet heuristics for class groups of number fields [CL84, CM87]. The main idea is that Tate–Shafarevich groups appear as random finite abelian groups with nondegenerate alternating bilinear pairings, weighted by the inverse of the size of the automorphism group.

The work of Poonen–Rains [PR12] models p-Selmer groups as intersections of random maximal isotropic subspaces in an infinite-dimensional quadratic space over $\mathbb{F}_p$. They obtain conjectural distributions for p-Selmer groups of elliptic curves (and abelian varieties). All of the currently known theoretical results on average sizes of Selmer groups, such as Theorem 2.2, agree with the Poonen–Rains heuristics.

Even more recently, Bhargava, Kane, Lenstra, Poonen, and Rains [BKL$^+$] have extended these heuristics to model both Selmer groups and Tate-Shafarevich groups simultaneously, by studying the distribution of the exact sequence that is the direct limit over n of (7) with p replaced by p^n.

There has also been recent progress in studying distributions of 2-Selmer groups for quadratic twist families, including work of Heath-Brown, Swinnerton-Dyer, Kane, Yu, Mazur, Rubin, and Klagsbrun, among many others [HB93, HB94, SD08, Kan12, Yu06, Yu05, MR10, KMR11].

Finally, a common theme in arithmetic geometry is to replace a number field by a function field, since geometry often helps in the latter case. Here, if $\mathbb{Q}$ is replaced by the function field $\mathbb{F}_q(t)$, de Jong [dJ02] gives an upper bound for the average size of 3-Selmer groups of elliptic curves over $\mathbb{F}_q(t)$. His methods for parametrizing elements of the Selmer group are similar to those described in §3.1 for Theorem 2.2 (and analogous results for Theorem 4.2).

[6]A rational degree p divisor on C is equivalent to a formal sum of p points of $C(\bar{\mathbb{Q}})$ which are together defined over $\mathbb{Q}$.

3. The average size of 2-Selmer groups

We now explain the main ideas behind Theorem 2.2 and the following stronger statement from [BS10a]:

Theorem 3.1 (Bhargava–Shankar 2010)**.** *Let $\mathscr{F}$ be any family of elliptic curves $E : y^2 = x^3 + Ax + B$ defined by finitely many congruence conditions on the integral coefficients A and B. Then the average size of $\mathrm{Sel}_2(E)$ for elliptic curves E in $\mathscr{F}$, ordered by height, is* 3*.*

The key observations are that binary quartic forms are closely related to elements of 2-Selmer groups of elliptic curves, and that it is possible to "count" integral binary quartic forms using techniques from the geometry of numbers.

More precisely, we will see in §3.1 that binary quartic forms with rational coefficients, up to standard transformations, with certain local properties correspond exactly to 2-Selmer elements of elliptic curves. The classical invariant theory of binary quartic forms plays a crucial role in this relationship; in particular, it gives the vertical map in diagram (8) below.

$$\begin{array}{ccc} \left\{\begin{array}{l}\text{2-Selmer elements}\\ \text{of elliptic curves } E\end{array}\right\} & \xrightarrow[\text{conditions}]{\text{local}} & \left\{\begin{array}{l}\text{binary quartic forms}\\ \text{up to equivalence}\end{array}\right\} \\ & \searrow^{\text{fiber over } E = \mathrm{Sel}_2(E)} & \Big\downarrow \text{invariant theory} \\ & & \{\text{elliptic curves } E\} \end{array} \tag{8}$$

In §3.2, we explain how suitably enhanced techniques from the geometry of numbers are used to count the number of binary quartic forms with bounded height[7]. Incorporating the local conditions by using sieve methods (see §3.3) produces a count of 2-Selmer elements for elliptic curves up to a given height. Because the fiber of the squiggly arrow in diagram (8) above an elliptic curve E is exactly $\mathrm{Sel}_2(E)$, dividing this count by the number of elliptic curves up to the same height, and then taking the limit of that quotient as the height tends to infinity, gives the average we seek.

This method is not known to work if the elliptic curves are ordered by discriminant or by conductor, instead of by height; the asymptotic number of elliptic curves with discriminant or conductor less than X, as X tends to infinity, is not even known.

3.1. Binary quartic forms and elliptic curves.

In the classical work [BSD63] of Birch and Swinnerton-Dyer that inspired the BSD conjecture, they study and use the relationship between binary quartic forms and 2-Selmer elements of elliptic curves.

A **binary quartic form** over $\mathbb{Q}$ is a homogeneous polynomial of degree 4 in two variables with rational coefficients, e.g.,

$$f(x_1, x_2) := ax_1^4 + bx_1^3x_2 + cx_1^2x_2^2 + dx_1x_2^3 + ex_2^4 \tag{9}$$

[7]The height of a binary quartic form is the same height, up to a constant, as for its associated elliptic curve.

with $a, b, c, d, e \in \mathbb{Q}$. The set of all binary quartic forms over $\mathbb{Q}$ is a 5-dimensional $\mathbb{Q}$-vector space V, with coordinates given by the coefficients a, b, c, d, and e. The group $\mathrm{GL}_2(\mathbb{Q})$ acts on the elements of V via

$$g \cdot f(x_1, x_2) = (\det g)^{-2} f((x_1, x_2) \cdot g) \tag{10}$$

for all $g \in \mathrm{GL}_2(\mathbb{Q})$; since scalar matrices act trivially, this action induces an action of $\mathrm{PGL}_2(\mathbb{Q})$. We call two binary quartic forms f and f' **equivalent** if there exists $g \in \mathrm{PGL}_2(\mathbb{Q})$ and $\lambda \in \mathrm{GL}_1(\mathbb{Q}) = \mathbb{Q}^\times$ such that $f' = \lambda^2(g \cdot f)$. In other words, the space V is a certain representation of the group $\mathrm{PGL}_2(\mathbb{Q}) \times \mathrm{GL}_1(\mathbb{Q})$, and two binary quartic forms are equivalent if they are in the same orbit of the group.

Under the action (10) of $\mathrm{GL}_2(\mathbb{Q})$, or equivalently, under the induced action of $\mathrm{PGL}_2(\mathbb{Q})$, the invariants of a binary quartic form (9) form a polynomial ring generated by two invariants:

$$I(f) := 12ae - 3bd + c^2$$
$$J(f) := 72ace + 9bcd - 27ad^2 - 27b^2e - 2c^3.$$

The **discriminant** $\Delta(f) := 4I(f)^3 - J(f)^2$ is nonzero exactly if f has four distinct roots over $\bar{\mathbb{Q}}$. The **height** of f is $\mathrm{ht}(f) := \max(|I(f)^3|, J(f)^2/4)$.

Genus one curves from binary quartic forms. Given a binary quartic form $f(x_1, x_2)$ with nonzero discriminant, one may construct a genus one curve $C(f)$ explicitly as the smooth compactification of the affine curve

$$y^2 = f(x_1, x_2).$$

This genus one curve is the double cover of the projective line ramified at exactly the roots of f (which may not be individually defined over $\mathbb{Q}$). It therefore comes equipped with a degree 2 line bundle $L(f)$, namely the pullback of the line bundle $\mathcal{O}(1)$ from $\mathbb{P}^1$; equivalently, a rational degree 2 divisor on $C(f)$ is given by the formal sum of the two points in the preimage of any rational point on $\mathbb{P}^1$ under this double cover.

If f' is an equivalent binary quartic form, then $C(f')$ and $C(f)$ are isomorphic, and the line bundles for each also correspond to one another under this isomorphism. In fact, binary quartic forms with nonzero discriminant up to equivalence are exactly in one-to-one correspondence with isomorphism classes of genus one curves with degree 2 line bundles!

Moreover, the Jacobian $E(f)$ of $C(f)$ depends only on the two invariants $I(f)$ and $J(f)$; it may be written in short Weierstrass form as

$$E(f) : y^2 = x^3 - \frac{I(f)}{3}x - \frac{J(f)}{27}. \tag{11}$$

Therefore, from our visualization of 2-Selmer elements described in §2.3, we see that for a binary quartic form f, if $C(f)$ is locally soluble, then the pair $(C(f), L(f))$ corresponds to an element of $\mathrm{Sel}_2(E(f))$. More precisely, let $V(\mathbb{Q})^{\mathrm{ls}}$ be the subset of locally soluble binary quartic forms $f(x_1, x_2)$ over $\mathbb{Q}$ with $\Delta(f) \neq 0$, i.e., those for which $y^2 = f(x_1, x_2)$ has a $\mathbb{Q}_\nu$-solution for all primes ν (including $\mathbb{Q}_\infty = \mathbb{R}$). Note that $V(\mathbb{Q})^{\mathrm{ls}}$ is preserved under the action of $\mathrm{GL}_2(\mathbb{Q}) \times \mathbb{Q}^\times$.

The equivalence classes of $V(\mathbb{Q})^{\mathrm{ls}}$ are in correspondence with 2-Selmer elements of elliptic curves; that is, we have the bijection

$$\mathrm{PGL}_2(\mathbb{Q}) \times \mathbb{Q}^\times \backslash V(\mathbb{Q})^{\mathrm{ls}} \overset{1-1}{\longleftrightarrow} \left\{ (E, \zeta) : \begin{matrix} E \text{ elliptic curve} \\ \zeta \in \mathrm{Sel}_2(E) \end{matrix} \right\} / \cong .$$

For any specific elliptic curve $E_{AB} : y^2 = x^3 + Ax + B$, we may specialize to the correspondence

$$\mathrm{PGL}_2(\mathbb{Q}) \backslash V_{AB}(\mathbb{Q})^{\mathrm{ls}} \overset{1-1}{\longleftrightarrow} \mathrm{Sel}_2(E_{AB}),$$

where $V_{AB}(\mathbb{Q})^{\mathrm{ls}}$ consists of binary quartic forms f with invariants $I(f) = -3A$ and $J(f) = -27B$.

Finding binary quartic forms with specified invariants is the best known way to explicitly compute the 2-Selmer group (and often, the rank) for a given elliptic curve; see, e.g., Cremona's `mwrank` program [Cre12].

Example 3.2. The only rational binary quartic forms, up to the action of $\mathrm{PGL}_2(\mathbb{Q})$, with invariants $I = 48$ and $J = -432$ are $f_0 = x_1^4 - 6x_1^2x_2^2 + 4x_1x_2^3 + x_2^4$ and $f_1 = x_1^4 + 4x_1x_2^3 + 4x_2^4$. They each have Jacobian isomorphic to the elliptic curve E given by $y^2 = x^3 - 16x + 16$. Thus,

$$\mathrm{Sel}_2(E) \cong \mathbb{Z}/2\mathbb{Z},$$

with f_0 representing the identity element. In this case, because E has at least one rational point $(x, y) = (0, 4)$ and $E(\mathbb{Q})_{\mathrm{tors}}$ is trivial, the sequence (7) implies that $\mathrm{rk}(E) = 1$ and $Ш(E)[2] = 0$.

In order to find the average size of the 2-Selmer group, the goal is therefore to count the number of equivalence classes in $V(\mathbb{Q})^{\mathrm{ls}}$ up to bounded height. The first step is to simply count the number of $\mathrm{PGL}_2(\mathbb{Z})$-equivalence classes of binary quartic forms with *integral* coefficients.

3.2. **Counting binary quartic forms using the geometry of numbers.** Methods from the geometry of numbers have been previously successful in similar counting questions, such as determining the number of equivalence classes of binary quadratic and binary cubic forms [Mer74, Sie44, Dav51b, Dav51c]. Bhargava–Shankar give an asymptotic count of the number of irreducible integral binary quartic forms, up to equivalence, of bounded height:

Theorem 3.3 ([BS10a])**.** *For $0 \le i \le 2$, let $N^{(i)}(X)$ be the number of $\mathrm{PGL}_2(\mathbb{Z})$-equivalence classes of irreducible integral binary quartic forms having $4 - 2i$ real roots and height less than X. Then*

$$N^{(0)}(X) = \frac{4}{135}\zeta(2)X^{5/6} + O(X^{3/4+\epsilon})$$
$$N^{(1)}(X) = \frac{32}{135}\zeta(2)X^{5/6} + O(X^{3/4+\epsilon})$$
and $$N^{(2)}(X) = \frac{8}{135}\zeta(2)X^{5/6} + O(X^{3/4+\epsilon}).$$

One may also impose finitely many congruence conditions on the coefficients a, b, c, d, and e of the binary quartic forms, e.g., requiring a to be 0 modulo p for a prime p. Then the number of equivalence classes of such integral binary quartic forms with height bounded by X is the total number of equivalence classes (the appropriate $N^{(i)}(X)$ from Theorem 3.3) multiplied by the p-adic density of each congruence condition imposed, with the same error term of $O(X^{3/4+\epsilon})$. This p-adic density is an easily computable fraction depending on p.

Remark 3.4. Although the exact statement of Theorem 3.3 is not strictly necessary for the proof of Theorem 3.1, the ideas and results used in the proof of Theorem 3.3 are essentially a subset of those needed for the average Selmer result.

The main idea in proving Theorem 3.3 is to reduce the question to counting lattice points in a nicely shaped domain, in which case the number of lattice points is approximately the volume of the domain. The major complication arises when the domain has cusps, which may be visualized as thin regions going off to infinity. A priori, these cuspidal regions may contain many or few integral points; see Figure 5. A clever "averaging" technique — first introduced by Bhargava in [Bha05, Bha10] for asymptotic counts of quartic and quintic rings — helps control exactly which points lie in the cusps.

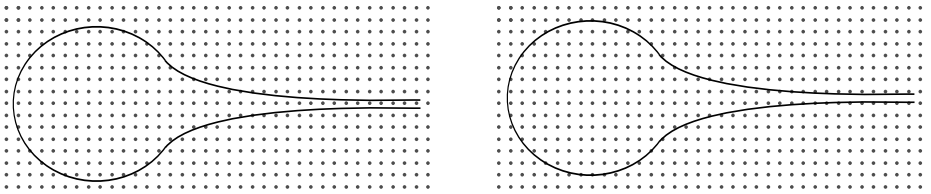


FIGURE 5. A domain with many lattice points in the cusp (left), and a domain with few lattice points in the cusp (right).

Let $V^{(i)}$ denote the subset of V corresponding to binary quartic forms with $4-2i$ roots. For the remainder of this section, we will focus on the case where the binary quartic forms have 4 real roots (that is, when $i = 0$); the other two cases are similar.

Reduction theory and fundamental domains. A **fundamental domain** or **set** for a group acting on a space is a set of elements in the space containing exactly one representative for each orbit. To count $\mathrm{PGL}_2(\mathbb{Z})$-orbits of the space $V(\mathbb{Z})$ of integral binary quartic forms, we may try to count lattice (integral) points in a fundamental domain for the action of $\mathrm{PGL}_2(\mathbb{Z})$ on $V(\mathbb{R})$. It is easier to break up this latter action into two intermediate ones, splitting the problem into two steps:

(i) find a fundamental set for the action of $\mathrm{PGL}_2(\mathbb{R}) \times \mathbb{R}^\times$ on $V(\mathbb{R})$, and
(ii) find a fundamental domain for the action of $\mathrm{PGL}_2(\mathbb{Z})$ on $\mathrm{PGL}_2(\mathbb{R}) \times \mathbb{R}^\times$.

For (i), such a fundamental set is easy to explicitly construct. One checks that a binary quartic form with all real roots and invariants I and J defines a unique $\mathrm{PGL}_2(\mathbb{R})$-orbit in $V(\mathbb{R})$ with those invariants. Thus, a fundamental set L consists of real binary quartic forms whose invariants range over all I and J, up to scaling (because of the action of $\mathbb{R}^\times$). Note that for any $h \in \mathrm{PGL}_2(\mathbb{R}) \times \mathbb{R}^\times$, the set hL is also a fundamental set. It is crucial that we may choose L such that hL is always a compact set.

Example 3.5. In fact, we may choose representatives for a fundamental set with height 1. One fundamental set for $V^{(0)}(\mathbb{R})$ is

$$L = \left\{ f_t(x_1, x_2) = x_1^3 x_2 - \frac{1}{3} x_1 x_2^3 - \frac{t}{27} x_2^4 : -2 \le t \le 2 \right\},$$

where $I(f_t) = 1$, $J(f_t) = t$, and discriminant $\Delta(f_t) = 4 - t^2 > 0$.

For (ii), there is a standard decomposition, due to Gauss, of a fundamental domain $\mathcal{F}$ for $\mathrm{PGL}_2(\mathbb{Z})\backslash\mathrm{PGL}_2(\mathbb{R})\times\mathbb{R}^\times$. This description also gives explicit coordinates for $\mathcal{F}$.

Combining (i) and (ii) shows that the set $\mathcal{F}hL$, for any $h\in\mathrm{PGL}_2(\mathbb{R})\times\mathbb{R}^\times$, contains a representative from each $\mathrm{PGL}_2(\mathbb{Z})$-orbit of $V^{(0)}(\mathrm{R})$. In fact, when viewed as a multiset, $\mathcal{F}hL$ overcounts each orbit — by the size of the stabilizer in $\mathrm{PGL}_2(\mathbb{R})$ of the binary quartic divided by the size of its stabilizer in $\mathrm{PGL}_2(\mathbb{Z})$. For binary quartics in $V^{(0)}(\mathbb{R})$, this quotient is $4/1=4$ almost always (in a sense that may be made precise), so it suffices to assume that each orbit is counted four times. In other words, the set $\mathcal{F}hL$ is (almost) a union of four fundamental domains for the action of $\mathrm{PGL}_2(\mathbb{Z})$ on $V^{(0)}(\mathbb{R})$.

We are now interested in counting the number of integer points in $\mathcal{F}hL$ of bounded height (and dividing by 4).

Averaging and volumes. As alluded to earlier, the number of lattice points in a domain like $\mathcal{F}hL$ is essentially the volume of the domain, by ideas of Minkowski and refinements by Davenport [Dav51a, Dav64], but one needs to control the points in the cusp of this domain.

In order to thicken the cusp for better control, we take not just a single domain $\mathcal{F}hL$, but a small ball's worth of such domains by letting the element h vary in a compact set. To obtain the final answer, the number of lattice points in the union of these domains $\mathcal{F}hL$, counted with multiplicity, must be divided by the volume of this compact set.

This larger domain with a thicker cusp may be split into two parts, the main body and the cusp; a clever choice of where to exactly separate the two will give the desired estimates. In particular, let the cusp be the part of the fundamental domain containing the binary quartic forms $f(x_1,x_2)$ from (9) for which the absolute value of the coefficient a of x_1^4 is strictly less than 1. Then any integral binary quartic form in the cusp has a equal to 0 and hence is reducible!

The volume of the main body then approximates the number of lattice points in it, and one may show that it contains a negligible number of reducible binary quartic forms.

Remark 3.6. Theorem 3.3 only concerns irreducible binary quartic forms, but when we return in §3.3 to counting binary quartic forms corresponding to 2-Selmer elements, we will include the reducible binary quartic forms found in the cusp.

The final step is to compute the volume of this main body, which may be done explicitly. A critical lemma in this computation involves changing from the standard Euclidean measure on the space $V(\mathbb{R})$ to the product of the Haar measure on the group $\mathrm{PGL}_2(\mathbb{R})$ and the measures given by the invariants I and J. This Jacobian computation mirrors the intuitive idea that $V(\mathbb{R})$ is roughly a product of $\mathrm{PGL}_2(\mathbb{R})$ and the quotient $\mathrm{PGL}_2(\mathbb{R})\backslash V(\mathbb{R})$.

3.3. **Sieves and uniformity estimates.** For Theorems 2.2 and 3.1, the relevant count is for *rational* equivalence classes of binary quartic forms corresponding to *locally soluble* genus one curves. Thus, we need to add several steps to the ideas from §3.2:

(a) As mentioned in Remark 3.6, the reducible binary quartic forms in the cusp must be incorporated.

(b) Find an integral representative for each $\mathrm{PGL}_2(\mathbb{Q}) \times \mathbb{Q}^\times$-orbit of $V(\mathbb{Q})^{\mathrm{ls}}$ with integral[8] invariants (and determine exactly how many each rational orbit contains).

(c) Impose the necessary local conditions — via sieve methods — to restrict to the space $V(\mathbb{Q})^{\mathrm{ls}}$ of locally soluble binary quartic forms.

Part (a) is important but straightforward. As mentioned earlier, the cusp region contains binary quartic forms that have a linear factor; these exactly correspond to the identity elements in the 2-Selmer groups! As these and other reducible forms do not appear often in the main body, the main body counts only irreducible binary quartic forms, corresponding to non-identity elements of the Selmer groups.

The first part of (b) is a standard fact in this case [BSD63, CFS10]; it is essentially a local computation. That is, given a rational binary quartic form f in $V(\mathbb{Q})^{\mathrm{ls}}$ with integral invariants, then for all primes p, there exists an element $g_p \in \mathrm{PGL}_2(\mathbb{Q}_p)$ such that the binary quartic form $g_p \cdot f$ has coefficients in $\mathbb{Z}_p$. Then we may use the idea of weak approximation to "glue" together all of these g_p's into an element $g \in \mathrm{PGL}_2(\mathbb{Q})$, as PGL_2 has class number one; the binary quartic form $g \cdot f$ then has integral coefficients.

In general, however, the orbit of such an $f \in V(\mathbb{Q})^{\mathrm{ls}}$ may contain many $\mathrm{PGL}_2(\mathbb{Z})$-orbits, so we need to weight each integral orbit by $1/n$, where n is the number of integral orbits for that rational orbit. This weighting may in fact be incorporated into the sieve for part (c). Again using the fact that the group PGL_2 has class number one, this last step is a local computation; the global weight is a product of local weights, which are related to the size of the stabilizers of the binary quartic forms in $\mathrm{PGL}_2(\mathbb{Q}_p)$.

This "geometric sieve," originating in work of Ekedahl [Eke91] and extended by Poonen [Poo03, Poo04] and Bhargava [Bha11], is the final step. Imposing *finitely* many congruence conditions on the binary quartic forms translates into multiplying the original count by the local densities for each condition, as mentioned after Theorem 3.3. However, we now need to impose a condition for every prime p, so to obtain an actual limit (as opposed to only a limsup), a certain *uniformity estimate* is needed.[9] In particular, one shows that the binary quartic forms that are "bad" at a prime p are rare as p approaches infinity, so they may be safely ignored.

In the end, the product of all these local factors simplifies[10] to be an invariant of the group PGL_2, called the **Tamagawa number** $\tau(\mathrm{PGL}_2)$. In other words, the limit as $X \to \infty$ of the weighted number of irreducible integral binary quartic forms in $V(\mathbb{Z})^{\mathrm{ls}}$ with height $< X$, divided by the number of elliptic curves of height $< X$, is $\tau(\mathrm{PGL}_2) = 2$. For the average for a family $\mathscr{F}$ of elliptic curves defined by finitely many congruence conditions, the local factors would affect the numerator and denominator equally, so the (limit of the) quotient would not change.

Finally, adding in the cusp contribution (for the identity elements in the 2-Selmer groups) implies the average size of the 2-Selmer group is

$$2 + 1 = 3.$$

[8]For simplicity, we are ignoring some factors of 2 and 3 throughout the discussion of this part.

[9]In the original paper [BS10a], obtaining this uniformity estimate is the most difficult and technical part, but the refined geometric sieve in [Bha11] significantly simplifies the computation needed here.

[10]See also [Poo12] for an explanation of this fact by computing an adelic volume instead.

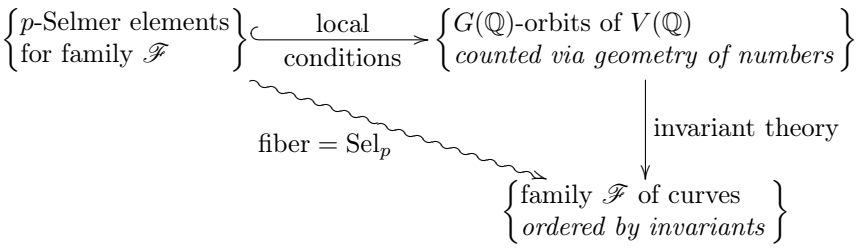
4. Generalizations and corollaries

We now outline generalizations of Theorem 2.2 and the methods discussed in §3 to other p-Selmer groups, other families of elliptic curves, and even families of higher genus curves. In §4.2, we also explain some corollaries for densities of low rank elliptic curves.

The strategy for proving Theorem 2.2 presented in §3 relies heavily on a description of 2-Selmer elements as equivalence classes of binary quartic forms with certain local properties. The geometry-of-numbers techniques apply to the situation after reducing the question to counting lattice points in a fundamental domain for the action of a group on a vector space.

Generalizing these methods thus depends on relating elements of Selmer groups to the orbits of a vector space V under the action of a group G; these orbits may then be counted as before. We modify diagram (8) to reflect the more general goal:

(12)



The family $\mathscr{F}$ for Theorem 2.2 consists of elliptic curves in short Weierstrass form. More generally, one may choose $\mathscr{F}$ to be other families of elliptic curves or even higher genus curves, whose Jacobians have analogously defined p-Selmer groups.

Finding appropriate groups G and vector spaces V related to the Selmer elements is still a relatively ad hoc process. For elliptic curves, we generally use the geometric description of elements of p-Selmer groups — as locally soluble torsors with degree p line bundles — to find such G and V.

Remark 4.1. The method summarized in diagram (12) was also previously used by Davenport–Heilbronn [DH69] and Bhargava [Bha05] to prove two of the only known cases of the Cohen-Lenstra-Martinet heuristics on distributions of ideal class groups of number fields. In those cases, the family $\mathscr{F}$ is replaced by a family of number fields (quadratic fields and cubic fields, respectively, ordered by discriminant), and the p-torsion of the ideal class group is the analogue of the p-Selmer group.

4.1. **Other Selmer groups for elliptic curves.** We survey recent results on average sizes of Selmer groups for elliptic curves; the methods behind these theorems all arise from the ideas highlighted in diagram (12).

In [BS10b], Bhargava–Shankar extend their methods from [BS10a] to 3-Selmer groups, by using the classical description of 3-Selmer elements as locally soluble curves cut out by ternary cubic forms, up to equivalence.

Theorem 4.2 (Bhargava–Shankar 2010)**.** *The average size of the* 3*-Selmer group for elliptic curves over* $\mathbb{Q}$*, ordered by height, is* 4*.*

The average for 3-Selmer groups gives an improved upper bound of 7/6 for the limsup of the average rank of elliptic curves.

In fact, Bhargava–Shankar have work in progress, using similar methods, to show that the average size of the 4- and 5-Selmer groups for elliptic curves, ordered by height, is 7 and 6, respectively. With some additional work, they are able to use these averages to show that the limsup of the average rank is in fact bounded above by 0.89.

In joint work with Bhargava [BH12b], we find the average sizes of 2- and/or 3-Selmer groups for various families of elliptic curves, such as the family

$$\mathscr{F}_1 := \left\{y^2 + a_3 y = x^3 + a_2 x^2 + a_4 x : a_2, a_3, a_4 \in \mathbb{Z}, \Delta \neq 0\right\} \tag{13}$$

of elliptic curves with one marked point, ordered by analogous notions of height. These averages rely on explicit descriptions of Selmer elements for these families as orbits of certain representations [BH12a]. Upper bounds on average ranks of elliptic curves in these families are also obtained in the same way.

For all the families considered in [BH12b], we find that the marked points on the elliptic curves essentially act independently. For example, for the family $\mathscr{F}_1$, independence would imply that the single marked point should increase the p-rank of the p-Selmer group by 1, and indeed, the 2- and 3-Selmer groups have average sizes $3 \cdot 2 = 6$ and $4 \cdot 3 = 12$, respectively.

4.2. **Lots of rank 0 and rank 1 curves.** Using the average size of 3-Selmer groups, one may deduce the existence of many elliptic curves with rank 0 and as a result, many for which the BSD conjecture is true!

Dokchitser–Dokchitser [DD10] prove the **p-parity conjecture** over $\mathbb{Q}$, which states that the root number of an elliptic curve over $\mathbb{Q}$ is determined by the parity of its p-Selmer rank. By using congruence conditions to construct a positive-density family of elliptic curves with equidistributed root number, Bhargava–Shankar combine the p-parity conjecture with Theorem 4.2 to show that a positive density[11] of all elliptic curves, when ordered by height, have rank exactly 0.

In addition, applying Skinner–Urban's results [SU06, SU10] on the main conjecture of Iwasawa theory for GL_2 shows that a positive proportion of elliptic curves have *analytic* rank exactly 0. Since the BSD conjecture is known for curves of analytic rank 0 by Kolyvagin's work [Kol88], Bhargava–Shankar conclude that a positive proportion of elliptic curves over $\mathbb{Q}$ satisfy the BSD conjecture.

Moreover, with the assumption that $Ш(E)$ for any elliptic curve E is finite (or the weaker assumption that the 3-torsion subgroup $Ш(E)[3]$ is always a square), they find a positive density of elliptic curves with rank 1.

Similar results hold for the family $\mathscr{F}_1$ from (13): a positive density of elliptic curves in $\mathscr{F}_1$ have rank 1, and conditional on the finiteness of Ш, a positive density have rank 2.

4.3. **Higher genus curves.** As mentioned in §1.1, while curves of genus at least 2 have finitely many rational points, determining the number of rational points is still quite difficult. Given an ordering of curves in a particular family, one may ask similar questions as for elliptic curves, e.g., for any finite number N, what is the density of curves with N points? What is the average number of rational points, if finite?

[11] All of these statements on having a positive density or proportion of curves with a given property are, more precisely, about the lower density of such curves being positive.

The techniques discussed in §3, surprisingly, give some results towards these questions, at least for hyperelliptic curves with rational Weierstrass points. Bhargava–Gross [BG12a] first find a description of the 2-Selmer elements here using rational orbits of a certain representation of odd orthogonal groups (see also work of Thorne [Tho12] for more such parametrizations for higher genus curves using Lie theory). They then compute the average size of the 2-Selmer group:

Theorem 4.3 (Bhargava–Gross 2012)**.** *Fix $g \geq 1$. Then the average size of the 2-Selmer group for Jacobians of genus g hyperelliptic curves over $\mathbb{Q}$ with a rational Weierstrass point, ordered by height, is* 3*.*

Not only does this give an upper bound of 3/2 for the limsup of the average Mordell-Weil rank of the Jacobians of such curves, but it also may be used — along with the method of Chabauty and Coleman [Cha41, Col85] —- to show that there are many curves with very few points. Poonen and Stoll [PS] have recently improved upon the results from [BG12a] of this type:

Corollary 4.4 (Poonen–Stoll 2012)**.** *Fix $g \geq 3$. Then a positive proportion of genus g hyperelliptic curves over $\mathbb{Q}$ with a rational Weierstrass point have no other rational points, and a majority of such curves have at most* 7 *rational points.*

In fact, Poonen–Stoll show that as the genus g tends to infinity, the lower density of these curves for which the given Weierstrass point is the only rational point tends to 1.

Finally, Bhargava and Gross [BG12b] have very recently shown, using methods analogous to [BG12a], that most hyperelliptic curves over $\mathbb{Q}$ have zero rational points!

* ∼ * ∼ *

Acknowledgments. We thank Manjul Bhargava, Bhargav Bhatt, Arul Shankar, and Brian Street for reading and commenting on previous drafts. We thank Simon Spicer for his help with collecting data for ranks and Selmer groups, and we thank Nick Katz for suggesting relevant references. All computations were done with SAGE [S$^+$12].

References

[BCDT01] Christophe Breuil, Brian Conrad, Fred Diamond, and Richard Taylor, *On the modularity of elliptic curves over* **Q***: wild 3-adic exercises*, J. Amer. Math. Soc. **14** (2001), no. 4, 843–939 (electronic).

[BG12a] Manjul Bhargava and Benedict H. Gross, *The average size of the 2-Selmer group of jacobians of hyperelliptic curves having a rational Weierstrass point*, 2012, `http://arxiv.org/abs/1208.1007`.

[BG12b] ———, *Most hyperelliptic curves over $\mathbb{Q}$ have no rational points*, in preparation, 2012.

[BH12a] Manjul Bhargava and Wei Ho, *Coregular spaces and genus one curves*, preprint, 2012.

[BH12b] ———, *On the average sizes of Selmer groups in families of elliptic curves*, preprint, 2012.

[Bha05] Manjul Bhargava, *The density of discriminants of quartic rings and fields*, Ann. of Math. (2) **162** (2005), no. 2, 1031–1063.

[Bha10] ———, *The density of discriminants of quintic rings and fields*, Ann. of Math. (2) **172** (2010), no. 3, 1559–1591.

[Bha11] ———, *The geometric squarefree sieve and unramified nonabelian extensions of quadratic fields*, preprint, 2011.

[BKL+] Manjul Bhargava, Daniel Kane, Hendrik Lenstra, Bjorn Poonen, and Eric Rains, *Modeling the distribution of Selmer groups, Shafarevich–Tate groups, and ranks of elliptic curves*, in preparation. Extended abstract available in pp. 45–48 of `http://www.mfo.de/document/1232/OWR_2012_38.pdf`.

[BM90] Armand Brumer and Oisín McGuinness, *The behavior of the Mordell-Weil group of elliptic curves*, Bull. Amer. Math. Soc. (N.S.) **23** (1990), no. 2, 375–382.

[BMSW07] Baur Bektemirov, Barry Mazur, William Stein, and Mark Watkins, *Average ranks of elliptic curves: tension between data and conjecture*, Bull. Amer. Math. Soc. (N.S.) **44** (2007), no. 2, 233–254.

[Bom90] Enrico Bombieri, *The Mordell conjecture revisited*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4) **17** (1990), no. 4, 615–640.

[Bru92] Armand Brumer, *The average rank of elliptic curves. I*, Invent. Math. **109** (1992), no. 3, 445–472.

[BS10a] Manjul Bhargava and Arul Shankar, *Binary quartic forms having bounded invariants, and the boundedness of the average rank of elliptic curves*, 2010, `http://arxiv.org/abs/1006.1002`.

[BS10b] ———, *Ternary cubic forms having bounded invariants, and the existence of a positive proportion of elliptic curves having rank 0*, 2010, `http://arxiv.org/abs/1007.0052`.

[BSD63] B. J. Birch and H. P. F. Swinnerton-Dyer, *Notes on elliptic curves. I*, J. Reine Angew. Math. **212** (1963), 7–25.

[BSD65] ———, *Notes on elliptic curves. II*, J. Reine Angew. Math. **218** (1965), 79–108.

[CFS10] John E. Cremona, Tom A. Fisher, and Michael Stoll, *Minimisation and reduction of 2-, 3- and 4-coverings of elliptic curves*, Algebra Number Theory **4** (2010), no. 6, 763–820.

[Cha41] Claude Chabauty, *Sur les points rationnels des courbes algébriques de genre supérieur à l'unité*, C. R. Acad. Sci. Paris **212** (1941), 882–885.

[CKRS02] J. B. Conrey, J. P. Keating, M. O. Rubinstein, and N. C. Snaith, *On the frequency of vanishing of quadratic twists of modular L-functions*, Number theory for the millennium, I (Urbana, IL, 2000), A K Peters, Natick, MA, 2002, pp. 301–315.

[CL84] H. Cohen and H. W. Lenstra, Jr., *Heuristics on class groups of number fields*, Number theory, Noordwijkerhout 1983 (Noordwijkerhout, 1983), Lecture Notes in Math., vol. 1068, Springer, Berlin, 1984, pp. 33–62.

[CM87] H. Cohen and J. Martinet, *Class groups of number fields: numerical heuristics*, Math. Comp. **48** (1987), no. 177, 123–137.

[Col85] Robert F. Coleman, *Effective Chabauty*, Duke Math. J. **52** (1985), no. 3, 765–770.

[Cre06] John Cremona, *The elliptic curve database for conductors to 130000*, Algorithmic number theory, Lecture Notes in Comput. Sci., vol. 4076, Springer, Berlin, 2006, pp. 11–29.

[Cre12] John Cremona, ***mwrank*** *program*, 2012, `http://homepages.warwick.ac.uk/~masgaj/mwrank/`.

[Dav51a] H. Davenport, *On a principle of Lipschitz*, J. London Math. Soc. **26** (1951), 179–183.

[Dav51b] ———, *On the class-number of binary cubic forms. I*, J. London Math. Soc. **26** (1951), 183–192.

[Dav51c] ———, *On the class-number of binary cubic forms. II*, J. London Math. Soc. **26** (1951), 192–198.

[Dav64] ———, *Corrigendum: "On a principle of Lipschitz"*, J. London Math. Soc. **39** (1964), 580.

[DD10] Tim Dokchitser and Vladimir Dokchitser, *On the Birch-Swinnerton-Dyer quotients modulo squares*, Ann. of Math. (2) **172** (2010), no. 1, 567–596.

[Del01] Christophe Delaunay, *Heuristics on Tate-Shafarevitch groups of elliptic curves defined over* $\mathbb{Q}$, Experiment. Math. **10** (2001), no. 2, 191–196.

[Del07] ———, *Heuristics on class groups and on Tate-Shafarevich groups: the magic of the Cohen-Lenstra heuristics*, Ranks of elliptic curves and random matrix theory, London Math. Soc. Lecture Note Ser., vol. 341, Cambridge Univ. Press, Cambridge, 2007, pp. 323–340.

[DH69] H. Davenport and H. Heilbronn, *On the density of discriminants of cubic fields*, Bull. London Math. Soc. **1** (1969), 345–348.

[dJ02] A. J. de Jong, *Counting elliptic surfaces over finite fields*, Mosc. Math. J. **2** (2002), no. 2, 281–311, Dedicated to Yuri I. Manin on the occasion of his 65th birthday.

[Eke91] Torsten Ekedahl, *An infinite version of the Chinese remainder theorem*, Comment. Math. Univ. St. Paul. **40** (1991), no. 1, 53–59.

[Fal83] G. Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern*, Invent. Math. **73** (1983), no. 3, 349–366.

[Fal91] Gerd Faltings, *Diophantine approximation on abelian varieties*, Ann. of Math. (2) **133** (1991), no. 3, 549–576.

[Gol79] Dorian Goldfeld, *Conjectures on elliptic curves over quadratic fields*, Number theory, Carbondale 1979 (Proc. Southern Illinois Conf., Southern Illinois Univ., Carbondale, Ill., 1979), Lecture Notes in Math., vol. 751, Springer, Berlin, 1979, pp. 108–118.

[HB93] D. R. Heath-Brown, *The size of Selmer groups for the congruent number problem*, Invent. Math. **111** (1993), no. 1, 171–195.

[HB94] ———, *The size of Selmer groups for the congruent number problem. II*, Invent. Math. **118** (1994), no. 2, 331–370, With an appendix by P. Monsky.

[HB04] ———, *The average analytic rank of elliptic curves*, Duke Math. J. **122** (2004), no. 3, 591–623.

[Kan12] Daniel M. Kane, *On the ranks of the 2-Selmer groups of twists of a given elliptic curve*, 2012, `http://arxiv.org/abs/1009.1365`.

[KMR11] Zev Klagsbrun, Barry Mazur, and Karl Rubin, *Selmer ranks of quadratic twists of elliptic curves*, 2011, `http://arxiv.org/abs/1111.2321`.

[Kol88] V. A. Kolyvagin, *Finiteness of $E(\mathbf{Q})$ and Ш$(E, \mathbf{Q})$ for a subclass of Weil curves*, Izv. Akad. Nauk SSSR Ser. Mat. **52** (1988), no. 3, 522–540, 670–671.

[KS99] Nicholas M. Katz and Peter Sarnak, *Random matrices, Frobenius eigenvalues, and monodromy*, American Mathematical Society Colloquium Publications, vol. 45, American Mathematical Society, Providence, RI, 1999.

[KS00] J. P. Keating and N. C. Snaith, *Random matrix theory and $\zeta(1/2+it)$*, Comm. Math. Phys. **214** (2000), no. 1, 57–89.

[Maz77] B. Mazur, *Modular curves and the Eisenstein ideal*, Inst. Hautes Études Sci. Publ. Math. (1977), no. 47, 33–186.

[Mer74] F. Mertens, *Ueber einige asymptotische Gesetze der Zahlentheorie*, J. Reine Angew. Math. **77** (1874), 289–338.

[Mor22] Louis J. Mordell, *On the rational solutions of the indeterminate equation of the third and fourth degrees*, Proc. Cambridge Philos. Soc. **21** (1922), 179–192.

[MR10] Barry Mazur and Karl Rubin, *Ranks of twists of elliptic curves and Hilbert's tenth problem*, Invent. Math. **181** (2010), no. 3, 541–575.

[Poo03] Bjorn Poonen, *Squarefree values of multivariable polynomials*, Duke Math. J. **118** (2003), no. 2, 353–373.

[Poo04] ———, *Bertini theorems over finite fields*, Ann. of Math. (2) **160** (2004), no. 3, 1099–1127.

[Poo12] ———, *Average rank of elliptic curves*, Séminaire Bourbaki, 2011-2012, 64ème année no. 1049.

[PR12] Bjorn Poonen and Eric Rains, *Random maximal isotropic subspaces and Selmer groups*, J. Amer. Math. Soc. **25** (2012), no. 1, 245–269.

[PS] Bjorn Poonen and Michael Stoll, *Chabauty's method proves that most odd degree hyperelliptic curves have only one rational point*, in preparation.

[S$^+$12] W. A. Stein et al., *Sage Mathematics Software (Version 5.3)*, The Sage Development Team, 2012, `http://www.sagemath.org`.

[SD08] Peter Swinnerton-Dyer, *The effect of twisting on the 2-Selmer group*, Math. Proc. Cambridge Philos. Soc. **145** (2008), no. 3, 513–526.

[Sie44] Carl Ludwig Siegel, *The average measure of quadratic forms with given determinant and signature*, Ann. of Math. (2) **45** (1944), 667–685.

[Sie66] ———, *Über einige Anwendungen diophantischer Approximationen (1929)*, Gesammelte Abhandlungen. Bände I, II, III, Springer-Verlag, 1966, pp. 209–266.

[Sil07] A. Silverberg, *The distribution of ranks in families of quadratic twists of elliptic curves*, Ranks of elliptic curves and random matrix theory, London Math. Soc. Lecture Note Ser., vol. 341, Cambridge Univ. Press, Cambridge, 2007, pp. 171–176.

[SU06] Christopher Skinner and Eric Urban, *Vanishing of L-functions and ranks of Selmer groups*, International Congress of Mathematicians. Vol. II, Eur. Math. Soc., Zürich, 2006, pp. 473–500.

[SU10] Christopher Skinner and Eric Urban, *The Iwasawa main conjectures for GL(2)*, preprint. Available at `http://www.math.columbia.edu/~urban/eurp/MC.pdf`, 2010.

[SW02] William A. Stein and Mark Watkins, *A database of elliptic curves—first report*, Algorithmic number theory (Sydney, 2002), Lecture Notes in Comput. Sci., vol. 2369, Springer, Berlin, 2002, pp. 267–275.

[Tho12] Jack Thorne, *The arithmetic of simple singularities*, Ph.D. thesis, Harvard University, 2012.

[TW95] Richard Taylor and Andrew Wiles, *Ring-theoretic properties of certain Hecke algebras*, Ann. of Math. (2) **141** (1995), no. 3, 553–572.

[Voj91] Paul Vojta, *Siegel's theorem in the compact case*, Ann. of Math. (2) **133** (1991), no. 3, 509–548.

[Wat07] Mark Watkins, *Rank distribution in a family of cubic twists*, Ranks of elliptic curves and random matrix theory, London Math. Soc. Lecture Note Ser., vol. 341, Cambridge Univ. Press, Cambridge, 2007, pp. 237–246.

[Wat08] ———, *Some heuristics about elliptic curves*, Experiment. Math. **17** (2008), no. 1, 105–125.

[Wil95] Andrew Wiles, *Modular elliptic curves and Fermat's last theorem*, Ann. of Math. (2) **141** (1995), no. 3, 443–551.

[You06] Matthew P. Young, *Low-lying zeros of families of elliptic curves*, J. Amer. Math. Soc. **19** (2006), no. 1, 205–250.

[Yu05] Gang Yu, *Average size of 2-Selmer groups of elliptic curves. II*, Acta Arith. **117** (2005), no. 1, 1–33.

[Yu06] ———, *Average size of 2-Selmer groups of elliptic curves. I*, Trans. Amer. Math. Soc. **358** (2006), no. 4, 1563–1584 (electronic).

[ZK87] D. Zagier and G. Kramarz, *Numerical investigations related to the L-series of certain elliptic curves*, J. Indian Math. Soc. (N.S.) **52** (1987), 51–69 (1988).

Department of Mathematics, Columbia University, New York, NY 10027
E-mail address: `who@math.columbia.edu`

TOPOLOGY OF NONARCHIMEDEAN ANALYTIC SPACES

SAM PAYNE

Abstract. This note surveys basic topological properties of nonarchimedean analytic spaces, in the sense of Berkovich, including the recent tameness results of Hrushovski and Loeser. We also discuss interactions between the topology of nonarchimedean analytic spaces and classical algebraic geometry.

Contents

1. Introduction

1.1. **Complex algebraic geometry.** At its most basic, classical complex algebraic geometry studies the common zeros in $\mathbb{C}^n$ of a collection of polynomials in $\mathbb{C}[x_1, \ldots, x_n]$. Such an algebraic set may have interesting topology, but is not pathological. It can be triangulated and admits a strong deformation retract onto a finite simplicial complex. Furthermore, it contains an everywhere dense open subset that is a complex manifold, and whose complement is an algebraic set of smaller dimension. Proceeding inductively, every algebraic set in $\mathbb{C}^n$ decomposes as a finite union of complex manifolds, and many of the deepest and most fundamental results in complex algebraic geometry are proved using holomorphic functions and differential forms, Hodge theory, and Morse theory on these manifolds.

Partially supported by NSF DMS-1068689 and NSF CAREER DMS-1149054.

1.2. **Beyond the complex numbers.** Modern algebraic geometers are equally interested in the common zeros in K^n of a collection of polynomials in $K[x_1, \ldots, x_n]$, for fields K other than the complex numbers. For instance, the field of rational numbers $\mathbb{Q}$ is interesting for arithmetic purposes, while the field of formal Laurent series $\mathbb{C}((t))$ is used to study deformations of complex varieties. Like $\mathbb{C}$, such fields have natural norms. The p-adic norm $|\ |_p$ is given by writing a rational number uniquely as $\frac{p^a r}{s}$, with p, r, and s relatively prime, and then setting

$$\left|\frac{p^a r}{s}\right|_p = p^{-a}.$$

The t-adic norm $|\ |_t$ is given similarly, by writing a formal Laurent series uniquely as t^a times a formal power series with nonzero constant term, and then setting

$$\left|t^a \sum a_i t^i\right|_t = e^{-a}.$$

One can make sense of convergent power series with respect to these norms, and it is tempting to work naively analytic functions, given locally by convergent power series, in this context. Difficulties arise immediately, however, for essentially topological reasons, unless the field happens to be $\mathbb{C}$. The pleasant properties of analytic functions in complex geometry depend essentially on $\mathbb{C}$ being an archimedean field.

1.3. **What is an archimedean field?** The *archimedean axiom* says that, for any $x \in K^*$, there is a positive integer n such that $|nx| > 1$. An archimedean field is one in which this axiom holds, such as the real numbers and the complex numbers. However, there are essentially no other examples. The archimedean axiom is satisfied only by $\mathbb{C}$, with powers of the usual norm, and restrictions of these norms to subfields. In particular, the only complete archimedean fields are $\mathbb{R}$ and $\mathbb{C}$.

1.4. **A nonarchimedean field is any other complete normed field.** We are not talking about the snake house or a rare collection of exotic creatures. Nonarchimedean fields are basically the whole zoo. Examples include the completion $\mathbb{Q}_p$ of $\mathbb{Q}$ with respect to the p-adic norm, and the field of formal Laurent series $\mathbb{C}((t))$. Also, every field is complete and hence nonarchimedean with respect to the trivial norm, given by $|x| = 1$ for $x \in K^*$.

The norm on a nonarchimedean field extends uniquely to its algebraic closure. The algebraic closure may not be complete*, but the completion of a normed algebraically closed field is again algebraically closed

*This is not difficult to see in examples. For instance, the algebraic closure of $\mathbb{C}((t))$ is the field of Puiseux series $\mathbb{C}\{\{t\}\} = \bigcup_n \mathbb{C}((t^{1/n}))$. The exponents appearing in a Puiseux series have denominators bounded below, but these bounds need not

[BGR84, Proposition 3.4.1.3]. So the completion of the algebraic closure of a nonarchimedean field is both nonarchimedean and algebraically closed. A typical example is $\mathbb{C}_p$, the completion of the algebraic closure of $\mathbb{Q}_p$.

1.5. **The ultrametric inequality.** In a nonarchimedean field, a much stronger version of the triangle inequality holds. The *ultrametric inequality* says that

$$|x+y| \leq \max\{|x|,|y|\}, \text{ with equality if } |x| \neq |y|.$$

This property deserves a few moments of contemplation. It implies that, if y is a point in the open ball

$$B(x,r) = \{y \in |K| \mid |y-x| < r\},$$

then $B(x,r) = B(y,r)$. In other words, every point in a nonarchimedean ball is a center of the ball.

1.6. **Nonarchimedean fields are totally disconnected.** Because of the ultrametric inequality, the open ball $B(x,r)$ in a nonarchimdean field is closed in the metric topology. Since these sets form a basis for the topology, the field K is totally disconnected. Doing naive analysis on such a totally disconnected set is totally unreasonable. For instance, if f and g are any two polynomials, then the piecewise defined function

$$\Phi(x) = \begin{cases} f(x) & \text{if } x \in B(0,1); \\ g(x) & \text{otherwise,} \end{cases}$$

is continuous. Even worse, this function Φ is "analytic" in the naive sense that it is given by a convergent power series in a neighborhood of every point.

1.7. **Grothendieck topologies.** Let K be a nonarchimedean field. Since K^n is totally disconnected in its metric topology, a purely naive approach to analytic geometry over K is doomed to failure. One kludge is to discard the naive notion of topology.

Let us return for a moment to the space of rational numbers $\mathbb{Q}$, which is totally disconnected in its metric topology. The interval $[0,1]$ in $\mathbb{Q}$ is not compact, but any open cover of $[0,1]$ by segments with rational centers and rational radii has a finite subcover. Similarly, this totally disconnected set cannot be decomposed into a disjoint union of two open segments with rational centers and rational radii. This suggests that $[0,1] \cap \mathbb{Q}$ is in some sense compact and connected with respect to covers by open intervals with rational centers and rational radii, and one should restrict to considering such covers in order to do analysis on $\mathbb{Q}$. An

be uniform on a Cauchy sequence. For instance, the sequence $x_n = \sum_{j=1}^{n} t^{j+1/j}$ is Cauchy, but has no limit in $\mathbb{C}\{\{t\}\}$.

approach like this can work, once one gives up the idea that an arbitrary union of open sets should be open. Naive topology involving open sets and covers by open sets is then replaced by a *Grothendieck topology*, consisting of a collection of covers satisfying certain axioms that are satisfied by usual open covers in topology.

1.8. **Rigid analytic geometry.** John Tate developed a satisfying and powerful theory of nonarchimedean analytic geometry, based on sheaves of analytic functions in a Grothendieck topology on K^n. His theory is called *rigid analytic geometry*, and the fundamental algebraic object in the theory, the ring of convergent power series on the unit disc, is called the *Tate algebra*. Algebraic properties of the Tate algebra, including the fact that it is noetherian, play an essential role in all forms of nonarchimedean analytic geometry, whether one works in the rigid setting or follows the approach of Berkovich. See [BGR84] for a comprehensive treatment of the foundations of rigid analytic geometry.

1.9. **Filling in gaps between points.** As mentioned above, one can develop a version of analysis on $\mathbb{Q}$ by replacing the metric topology with a suitable Grothendieck topology. Nevertheless, most mathematicians prefer to add in new points that "fill in the gaps" between the rational numbers and do analysis on the real numbers instead. Note the fundamental absurdity of this construction. Although $\mathbb{Q}$ is dense in $\mathbb{R}$, it has measure zero. Once we have filled in the gaps, we can more or less ignore the points in $\mathbb{Q}$ when we do analysis. What is added is so much larger than what we started with.

In the late 1980s and early 1990s, Vladimir Berkovich developed a new version of analytic geometry over nonarchimedean fields. At the heart of his construction is a topological space that fills in the gaps between the points of K^n, producing a path connected, locally compact Hausdorff space that contains K^n, with its metric topology, as an everywhere dense subset. The underlying algebra and analysis in Berkovich's theory are essentially the same as in rigid analytic geometry, but one now has an honest, naive topological space to work with as well. The subject of this note is the topology of the spaces appearing in Berkovich's theory, recent results on the tameness of these spaces, and relations between topological invariants of these spaces and more classical notions in complex algebraic geometry.

2. Nonarchimedean analytification.

Nonarchimedean analytification is a functor from algebraic varieties (or, more generally, separated schemes of finite type) over a nonarchimedean field to analytic spaces in the sense of Berkovich. For simplicity, we focus on the case of an affine variety. Analytifications of arbitrary schemes are obtained from analytifications of affine schemes by a natural gluing procedure.

2.1. **The analytification of an affine scheme.** Let K be a nonarchimedean field. Consider polynomials $f_1, \ldots, f_r \in K[x_1, \ldots, x_n]$, and let X be the space of solutions to this system of equations.[†] If $x = (x_1, \ldots, x_n)$ is a point in $X(K)$, then there is an associated valuation on the quotient ring

$$K[X] = K[x_1, \ldots, x_n]/(f_1, \ldots, f_r).$$

Here, a ring valuation is simply a function from $K[X]$ to $\mathbb{R} \cup \infty$ that satisfies the usual axioms of a valuation on a field, specifically that

$$\mathrm{val}(fg) = \mathrm{val}(f) + \mathrm{val}(g)$$

and

$$\mathrm{val}(f+g) \geq \min\{\mathrm{val}(f), \mathrm{val}(g)\}, \text{ with equality if } \mathrm{val}(f) \neq \mathrm{val}(g).$$

We will only consider ring valuations with the additional property that the restriction to K is the given valuation. The valuation associated to a point x in $X(K)$ is given simply by

$$f \mapsto \mathrm{val}(f(x)).$$

Here we follow the usual convention that $\mathrm{val}(0)$ is ∞. The one significant difference is that a ring valuation may send nonzero elements of the ring to ∞.

Definition 1. *The analytification X^{an} is the space of all ring valuations on $K[X]$ that extend the given valuation on K.*[‡]

[†]In other words, X is the Zariski spectrum $\mathrm{Spec}\, K[x_1, \ldots, x_n]/(f_1, \ldots, f_r)$, a locally ringed space whose underlying topological space is the set of prime ideals $\mathfrak{p}$ in this quotient ring, with the Zariski topology. Yoneda's Lemma identifies this space with the functor that associates to a K-algebra S the set of solutions to $f_1, \ldots, f_r$ in S^n. In particular, if $L|K$ is an extension field then $X(L)$ is the set of points $y \in L^n$ such that $f_i(y) = 0$, for $1 \leq i \leq r$.

[‡]A word of caution is in order. The system of polynomials $f_1, \ldots, f_r$ may not have any solutions defined over K. Nevertheless, if $K[X]$ is nonzero then X^{an} is not empty. One way to see this is to observe that the system $f_1, \ldots, f_r$ has solutions over the algebraic closure $\overline{K}$. Since K is complete, its valuation extends uniquely to $\overline{K}$, and hence a solution with coordinates in $\overline{K}$ also determines a point of X^{an}. More generally, if $L|K$ is an extension field with a valuation that extends the given one on K, then any solution to $f_1, \ldots, f_r$ with coordinates in L determines a point of X^{an}.

We write x for a point of X^{an}, when thinking geometrically, and val_x for the corresponding valuation on $K[X]$. The topology on X^{an} is the subspace topology for the natural inclusion

$$X^{\mathrm{an}} \subset (\mathbb{R} \cup \infty)^{K[X]}.$$

This is the coarsest topology such that, for each $f \in K[X]$, the function on X^{an} given by $x \mapsto \mathrm{val}_x(f)$ is continuous.

Theorem 1 ([Ber90])**.** *The topological space X^{an} is Hausdorff, locally compact, and locally path connected. The induced topology on the subset $X(K)$ of points with coordinates in K is the metric topology, and if K is algebraically closed then this subset is everywhere dense.*

In this sense, X^{an} is a reasonably topological space on which to do analysis that "fills in the gaps" between the points in the totally disconnected set $X(K)$.

2.2. **Structure sheaf and morphisms.** As an analytic space, X^{an} comes with much more structure than just a topology. It has a sheaf of analytic functions given locally by convergent power series, and analytic maps to and from other analytic spaces. The pull back of an analytic function under an analytic map is analytic, and there are well-behaved notions of open and closed embeddings, as well as flat, smooth, proper, finite, and étale morphisms in the category of analytic spaces. One example of an étale morphism is given by pulling back the structure sheaf to a topological covering space, so the fundamental group of the underlying topological space gives essential information regarding the étale homotopy type of the analytic space. The details may be found in [Ber93]. Here, we are content to simply study the topological space underlying X^{an}.

2.3. **Projection to the scheme.** Let x be a point in X^{an}. The *kernel* of val_x, by which we mean $\mathrm{val}_x^{-1}(\infty)$ is a prime ideal $\mathfrak{p}$, and val_x factors through a valuation on the residue field $\kappa_{\mathfrak{p}}$, the fraction field of the quotient $K[X]/\mathfrak{p}$, whose restriction to K is the given one.

Definition 2. *For any extension field $L|K$, let $\mathcal{V}_{\mathbb{R}}(L)$ be the space of all valuations on L that extend the given valuation on K.*

The map taking a point in the analytification to the kernel of the corresponding valuation gives a natural surjection

$$X^{\mathrm{an}} \to X$$

whose fiber over a point $\mathfrak{p}$ is $\mathcal{V}_{\mathbb{R}}(\kappa_{\mathfrak{p}})$. In particular, the analytification decomposes as a disjoint union

$$X^{\mathrm{an}} = \bigsqcup_{\mathfrak{p}\in X} \mathcal{V}_{\mathbb{R}}(\kappa_{\mathfrak{p}}). \tag{1}$$

Note that the topology on the scheme X is never Hausdorff unless X has dimension zero. Its nonclosed points are the nonmaximal prime ideals $\mathfrak{p} \subset K[X]$, and the closure of $\mathfrak{p}$ is the irreducible variety $\operatorname{Spec} K[X]/\mathfrak{p}$. The process of analytification produces a Hausdorff space by replacing each nonclosed point $\mathfrak{p}$ with the space of valuations $\mathcal{V}_{\mathbb{R}}(\kappa_{\mathfrak{p}})$. Each closed point of X is a maximal ideal $\mathfrak{m} \subset K[X]$. The residue field $\kappa_{\mathfrak{m}}$ at a maximal ideal is algebraic over K [AM69, Proposition 7.9], so the valuation on K extends uniquely. Therefore, there is a single point of X^{an} over each closed point of X.

2.4. **A quotient description of X^{an}.** The decomposition above shows that each point of X^{an} is associated to a point of X together with a valuation on its residue field that extends the given one on K. For some purposes, rather than keeping tack of all of these residue fields, it makes more sense to consider points defined over arbitrary valued extensions $L|K$. One can still recover the analytification X^{an} by taking a quotient by an appropriate equivalence relation, as follows.

Here, and throughout, a valued extension $L|K$ is a field L together with a valuation $\mathrm{val} : L^* \to \mathbb{R}$ that extends the given valuation on K. We consider triples consisting of an extension field of K, a valuation that extends the given one, and a point of X over this valued extension, and the equivalence relation generated by setting

$$(L, \mathrm{val}, x) \sim (L', \mathrm{val}', x')$$

whenever there is an embedding $L \subset L'$ such that $\mathrm{val}'|_L = \mathrm{val}$ and $x \mapsto x'$ under the induced inclusion $X(L) \subset X(L')$.

Proposition 1. *The analytification X^{an} is the space of equivalence classes of points of X over valued extensions of K:*

$$X^{\mathrm{an}} = \{(L, \mathrm{val}, x)\}/\sim .$$

Much of the recent progress in understanding the topology of nonarchimedean analytic spaces has come through logic and model theory, and this description of X^{an} in terms of equivalence classes of points over valued extensions is closest in spirit to the *spaces of stably dominated types* that appear prominently in this context. Note, however, that the model theorists typically consider valuations into ordered groups of arbitrary rank, not just the real numbers.

3. Examples: The affine line, curves, and the affine plane

If X has dimension 0 then X^{an} is equal to X, and both have the discrete topology. We now consider the first nontrivial cases of analytifications.

3.1. **Analytification of the line: trivial valuation.** The simplest example to consider is the affine line

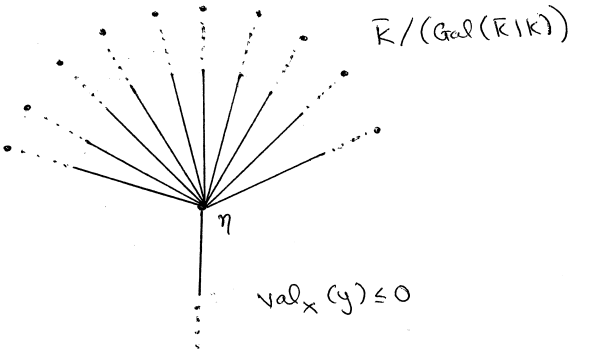
$$\mathbb{A}^1 = \operatorname{Spec} K[y],$$

in the case where the valuation on K is trivial. Let x be a point in $\mathbb{A}^1$. If $\mathrm{val}_x(y)$ is negative and

$$f = a_0 + a_1 y + \cdots + a_d y^d$$

is a polynomial of degree d, then $\mathrm{val}_x(f) = d\mathrm{val}_x(y)$. Therefore, val_x is uniquely determined by $\mathrm{val}_x(y)$, and the limit, as $\mathrm{val}_x(y)$ goes to 0, is the trivial valuation η on the function field $K(y)$. This gives an embedded copy of $\mathbb{R}_{\leq 0}$ in $(\mathbb{A}^1)^{\mathrm{an}}$.

Now, suppose val_x is not trivial, and $\mathrm{val}_x(y) \geq 0$. Then val_x is nonnegative on all of $K[y]$, and the set of f such that $\mathrm{val}_x(f) > 0$ is a nonzero prime ideal. Each such ideal is generated by a unique irreducible monic polynomial $g \in K[y]$. Given such a g and a positive real number t, there is a unique valuation on $K[y]$ such that $\mathrm{val}(g) = t$; it takes $g^a \cdot h$ to at, for h relatively prime to g. The limit of these valuations as t goes to 0 is again the trivial valuation η, and the limit as t goes to ∞ is the closed point corresponding to the maximal ideal $\mathfrak{m}_g$ generated by g. This gives a rough picture of $(\mathbb{A}^1)^{\mathrm{an}}$ as a sort of tree, with an infinite stem consisting of valuations on $K(y)$ that are negative on y, and infinitely many leaves corresponding to the irreducible polynomials in $K[y]$. Equivalently, the leaves correspond to closed points in the scheme $\mathbb{A}^1$ over K, or elements $\overline{K}/\mathrm{Gal}(\overline{K}|K)$.



Some discussion of the topology on this tree is in order. The topology on the subset where $\mathrm{val}(y) \geq 0$ is *not* the cone over the discrete set

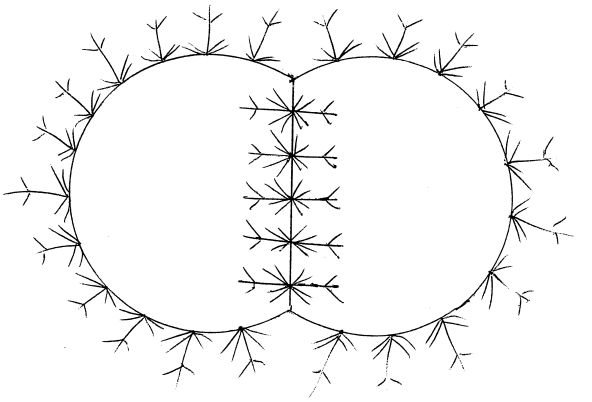
$\overline{K}/\mathrm{Gal}(\overline{K}|K)$. Rather, it is an inverse limit of cones over finite subsets of $\overline{K}/\mathrm{Gal}(\overline{K}|K)$, so any neighborhood of η contains all but finitely many of the copies of the branches. To see this, note that for any $f \in K[y]$, the induced function $(\mathbb{A}^1)^{\mathrm{an}} \to \mathbb{R} \cup \infty$ taking x to $\mathrm{val}_x(f)$ is zero on all but finitely many of these rays, those corresponding to irreducible factors of f. By construction, the preimages of neighborhoods of 0 in $\mathbb{R} \cup \infty$, each of which contains all but finitely many of the branches, form a basis for the neighborhoods of η.

In this way, not only does $\mathcal{V}_{\mathbb{R}}(K(y))$ fill in the gaps to connect the set of closed points in $\mathbb{A}^1_K$ with the discrete (metric) topology, it also interpolates between the metric topology and the cofinite (Zariski) topology in a subtle and beautiful way.

3.2. **Analytification of the line: nontrivial valuation.** The analytification of $\mathbb{A}^1$ in the case of a nontrivial valuation is again a tree, but now the set of branch points is dense. At each branch point, the local topology is like the topology at η in the analytification of the line with respect to the trivial valuation. It is described beautifully, and in detail, in Section 1 of [Bak08a].

3.3. **Analytification of curves.** The analytification of any smooth curve X looks locally similar to that of the line. If the valuation is trivial, then X^{an} has finitely many open branches, corresponding to the points of the smooth projective model that are not in X, and the rest is an inverse limit of cones over finite subsets of $X(\overline{K})/\mathrm{Gal}(\overline{K}|K)$.

If the valuation on K is nontrivial, then X^{an} is locally homeomomorphic to $(\mathbb{A}^1)^{\mathrm{an}}$, but may have nontrivial global topology, as in the following example.



The dual graph of the special fiber of a semistable formal model embeds in X^{an} as a strong deformation retract. For instance, an elliptic curve of bad reduction has a semistable formal model whose special fiber is a

loop of copies of $\mathbb{P}^1$, and its analytification deformation retracts onto a circle. Every finite graph occurs in this way, as the dual graph of the special fiber of a formal model, and hence as a deformation retraction of an analytic curve.

See [BPR11, Section 5] for further details on the structure theory of nonarchimedean analytic curves in the case where K is algebraically closed. The general case is similar; if K is not algebraically closed then X^{an} is the analytification of the base change to the completion of the algebraic closure, modulo the action of $\mathrm{Gal}(\overline{K}|K)$.

In the case where K has a countable dense subset, as is the case for $\mathbb{Q}_p$, $\mathbb{C}_p$, and $\mathbb{C}((t))$, the topology of analytic curves over K is closely related to that of the "universal dendrite," and each such curve admits a closed embedding in the euclidean space $\mathbb{R}^2$ [HLP12].

3.4. **Toward the analytification of the affine plane.** Let us try to form a mental image of the analytification of the affine plane, using the discussion of curves, above, and the decomposition (1), in the case where the valuation is trivial. To start, note that $(\mathbb{A}^2)^{\mathrm{an}}$ contains the analytification of any plane curve, and the complement of the union of these analytic curves is $\mathcal{V}_{\mathbb{R}}(K(y,z))$, the space of real valuations on the function field $K(y,z)$ that are trivial on K. Just as the closed points of a curve X lie at the ends of infinite branches of $\mathcal{V}_{\mathbb{R}}(K(X))$, the analytifications of curves in $\mathbb{A}^2$ lie in some sense at infinity, as limits of two-dimensional membranes in $\mathcal{V}_{\mathbb{R}}(K(y,z))$. Of course, the situation is somewhat more complicated, since the analytifications of curves X and X' are joined at their points of intersection.

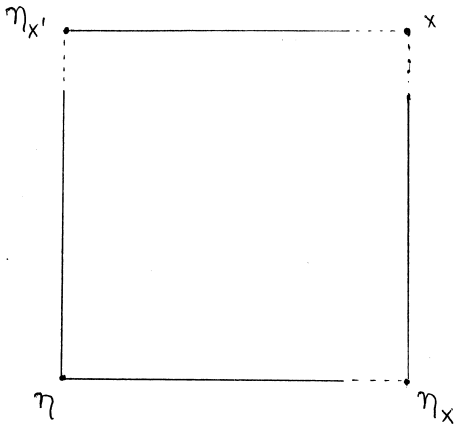
So, imagine a network of analytic curves at infinity, one for each curve in the plane, glued along leaves of the infinite branches corresponding to their points of intersection. We now begin to describe how $\mathcal{V}_{\mathbb{R}}(K(y,z))$ fills in the interior of this network. Suppose X and X' meet transversely at a point x in $\mathbb{A}^2$. Then any function $f \in K[y,z]$ can be expanded locally near x as a power series in local coordinates given by defining equations for X and X'. There is then a cone of "monomial valuations" in these local coordinates, defined as follows. Let $v = (v_1, v_2)$ be a point in the cone $\mathbb{R}^2_{\geq 0}$, and let f whose expansion in these local coordinates is

$$f = \sum a_{ij} s^i t^j.$$

Then the valuation corresponding to v takes f to the "v-weight of the smallest monomial" in this expansion,

$$\mathrm{val}_v(f) = \min\{iv_1 + jv_2 \mid a_{ij} \neq 0\}.$$

The closure of this cone in $(\mathbb{A}^2)^{\mathrm{an}}$ is a copy of $(\mathbb{R}_{\geq 0} \cup \infty)^2$, joining the trivial valuation η on $K(y,z)$ to the trivial valuations η_X and $\eta_{X'}$ on $K(X)$ and $K(X')$, respectively, and the point x.



In the geometry of this cone, the limit of any ray with positive finite slope is x, so it is perhaps best imagined as a curved membrane stretching an infinite distance toward x, from the frame formed by the rays joining η_X and $\eta_{X'}$.

Understanding how all of these membranes fit together in $(\mathbb{A}^2)^{\mathrm{an}}$ is challenging, especially as one must also keep track of the topology in a neighborhood of η. Moreover, we are still far from a full description of the underlying set of $(\mathbb{A}^2)^{\mathrm{an}}$. All of the valuations on $K(y,z)$ that are monomial in some system of local coordinates are of the simplest flavor in the sense of pure valuation theory; they satisfy Abhyankar's inequality [Abh56], which says that transcendence degree of the residue field extension plus the rank of the extension of the value group is less than or equal to the transcendence degree of the total extension, with equality. There are many valuations on $K(y,z)$ that do not appear in this way. For instance, a pair of formal power series h and h' in $K[[t]]$ define a formal germ of a curve in $\mathbb{A}^2$, and there is a valuation on $K[y,z]$ obtained by pulling functions back to this germ and computing order of vanishing at $t=0$. If these power series are algebraically independent over K, then the image of this germ is Zariski dense in $\mathbb{A}^2$, and Abhyankar's inequality is strict. These points are outside of the infinite union of membranes described above.

Each valuation corresponding to a point in $(\mathbb{A}^2)^{\mathrm{an}}$, including those where Abhyankar's inequality is strict, may be obtained as a limit of monomial valuations in various systems of local coordinates (or even as a limit of valuations corresponding to closed points), and the same is true

in higher dimensions and on singular spaces, but the precise way that all of the pieces fit together becomes more and more difficult to describe. It was not known until quite recently, through the groundbreaking work of Hrushovski and Loeser [HL10], that the analytification of a smooth variety over a trivially valued field is locally contractible.

4. Tameness of analytifications

Beyond the case of curves, which can be treated more or less by hand, it is not obvious that analytifications of algebraic varieties are not pathological topological spaces. For instance, while it is straightforward to show that the analytification of a smooth variety over a trivially valued field is contractible, the only known proof that they are locally contractible passes through model theory and spaces of stably dominated types. This is just one of the fundamental consequences of the tameness theorem of Hrushovski and Loeser [HL10].

4.1. **Semialgebraic sets.** To state the tameness theorem, it is most helpful to talk about analytic spaces that are more general than analytifications of algebraic varieties. Such spaces appear also in applications to complex algebraic geometry, including for the study of Milnor fibers, links of singularities, and mixed Hodge structures of open varieties. For simplicity, we restrict to constructions within a single affine variety. The interested reader may consult the references for generalizations to quasiprojective varieties and varieties moving in families.

Definition 3. *Let X be an affine algebraic variety over K. A semialgebraic subset $U \subset X^{\mathrm{an}}$ is a finite boolean combination of subsets of the form*

$$\{x \in X^{\mathrm{an}} \mid \mathrm{val}_x(f) \bowtie \lambda \mathrm{val}_x(g)\},$$

with $f, g \in K[X]$, $\lambda \in \mathbb{R}$, and $\bowtie \in \{\leq, \geq, <, >\}$.

By construction, every point in X^{an} has a basis of neighborhoods that are semialgebraic sets. Every semialgebraic subset of X^{an} is an *analytic domain*, in the sense of [Ber90], so it inherits a canonical analytic structure from X, including a sheaf of analytic functions, coherent sheaves of modules, an étale topology, and so on.

4.2. **Statement of the tameness theorem.** We now state a basic version of the main result from [HL10].

Theorem 2. *Let $U \subset X^{\mathrm{an}}$ be a semialgebraic subset. Then there is a finite simplicial complex $\Delta \subset U$, of dimension less than or equal to $\dim(X)$, and a strong deformation retraction $U \times [0,1] \to \Delta$.*

Since Δ is locally contractible, and the topology on X^{an} has a semialgebraic basis, it follows that X^{an} is locally contractible. The homotopy type of the complex Δ is a fundamental invariant of a semialgebraic space U, and many applications to complex geometry involve understanding the cohomology and fundamental groups of these complexes.

The approach of Hrushovski and Loeser does not involve the construction of nice models or toroidal compactifications. It thereby avoids resolution of singularities, and is insensitive to the residue characteristic. As mentioned above, the proof involves a detailed study of spaces of stably dominated types, a notion coming from model theory [HHM08]. The case of curves is treated by hand, and the general case is a subtle induction on dimension, which involves birationally fibering an n-dimensional variety by curves over a base of dimension $n-1$. In particular, the proof of the tameness theorem for a single variety in dimension n requires a tameness statement for families of varieties in lower dimensions. See the Bourbaki notes of Ducros [Duc12] for an excellent introduction to this work, and the original paper [HL10] for further details.

4.3. **History.** In the case where the valuation on K is nontrivial, Berkovich constructed finite simplicial complexes Δ associated to semistable and plursitable formal models of the analytic space U, which he called *skeletons*, and used them to prove local contractibility of smooth analytic spaces in arbitrary characteristic, when the valuation is nontrivial [Ber99, Ber04]. These complexes are geometric realizations of the dual complex of the special fiber of the model, and embed in U as strong deformation retracts. The tameness theorem follows from the theory of skeletons of semistable and pluristable formal models, provided the valuation is nontrivial and such models exist, as is the case in residue characteristic zero. Thuillier has given a beautiful and closely related construction of skeletons for toroidal embeddings in the case of a trivial valuation [Thu07].

5. Applications to complex algebraic geometry

The link between algebraic and analytic geometry over nonarchimedean fields is as close as the link between complex algebraic and complex analytic geometry. Coherent algebraic sheaves have coherent analytifications, analytifications of étale algebraic morphisms are étale, and there are comparison theorems for ℓ-adic étale cohomology [Ber90, Ber93]. It is not at all surprising, then, that nonarchimedean analytic techniques are powerful for studying algebraic varieties over nonarchimedean fields.

However, nonarchimedean analytic techniques are equally powerful for studying algebraic varieties over the complex numbers. The reason is simple: nonarchimedean fields such as $\mathbb{C}_p$ and the completion of $\mathbb{C}\{\{t\}\}$ are isomorphic to $\mathbb{C}$ as abstract fields. The isomorphism is not explicit

or geometric, but elimination of quantifiers for algebraically closed fields implies that any two uncountable algebraically closed fields of the same cardinality and characteristic are isomorphic [Mar02, Proposition 2.2.5]. In particular, whenever one can use nonarchimedean analytic techniques to produce a variety over $\mathbb{C}_p$ with a certain collection of algebraic properties, it follows that there exists a variety over $\mathbb{C}$ with the same collection of properties.

Perhaps more surprisingly, one can also get significant mileage by studying analytifications of open and singular complex varieties with respect to the trivial valuation, as in the discussions of singular cohomology and Milnor fibers, below.

5.1. **Tropical geometry.** Many applications of nonarchimedean analytic spaces in complex geometry involve less information than the full structure sheaf, but more than the mere topological space. Tropical geometry resides firmly in this intermediate realm. For instance, if X is a curve then $\mathcal{V}_{\mathbb{R}}(K(X))$, the complement in X^{an} of the set of closed points of X, inherits a natural metric. Through the tropical Riemann-Roch theorem [BN07, GK08, MZ08], Baker's specialization lemma and its generalizations [Bak08b, AC12, AB12], Poincaré Lelong formula [Thu05, BPR11], and the theory of harmonic morphisms of metric graphs [BN09, BR10], this metric is a powerful tool in the study of linear series on algebraic curves. It has been used to characterize dual graphs of special fibers of regular semistable models of curves of a given gonality [Cap12], to compute the gonality of curves that are generic with respect to their Newton polygon [CC12], to bound the gonality of Drinfeld modular curves [CKK12], and to give a new proof of the Brill-Noether theorem [CDPR12].

In the remaining sections, we survey some of the applications of nonarchimedean analytic geometry to classical complex varieties that involve only the toplogy of analytifications.

5.2. **Singular cohomology.** Let X be an algebraic variety over the complex numbers, and let $X(\mathbb{C})$ be the associated complex analytic space. Recall that Deligne defined a canonical mixed Hodge structure on the rational cohomology $H^*(X, \mathbb{Q})$, and one part of this structure is the weight filtration

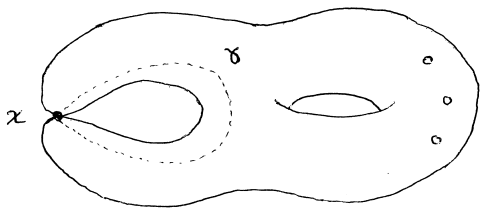
$$W_0 H^k(X(\mathbb{C}), \mathbb{Q}) \subset \cdots \subset W_{2k} H^k(X(\mathbb{C}), \mathbb{Q}) = H^k(X(\mathbb{C}), \mathbb{Q}),$$

which is strictly functorial for algebraic morphisms. This means that if $f : X' \to X$ is a morphism then

$$f^*(H^k(X(\mathbb{C}), \mathbb{Q}) \cap W_j H^k(X'(\mathbb{C}), \mathbb{Q}) = f^* W_j H^k(X(\mathbb{C}), \mathbb{Q}).$$

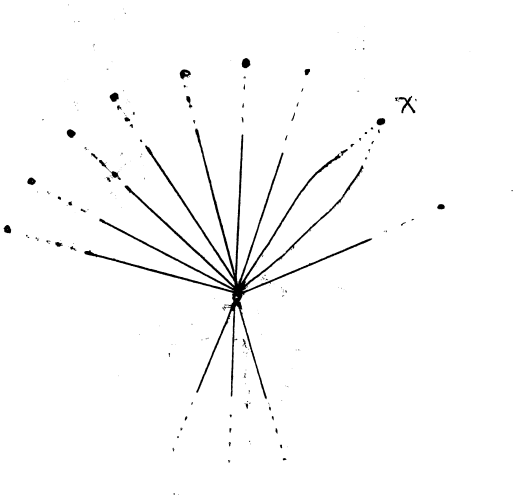
If X is smooth and compact then $W_{k-1}H^k(X,\mathbb{Q}) = 0$ and $W_kH^k(X,\mathbb{Q}) = H^k(X,\mathbb{Q})$. In other words, $H^k(X(\mathbb{C}),\mathbb{Q})$ is of pure weight k. In the general case, where X may be singular and noncompact, the graded pieces of H^k of weight less than k encode information on the singularities of X, while the pieces of weight greater than k encode information about the link of the boundary in a compactification. This is perhaps best illustrated by an example.

Example 1. Consider a curve X of geometric genus 1, with three punctures and a single node x, as shown.



Its normalization $\widetilde{X}$ is obtained by resolving the node. The homology $H_1(X(\mathbb{C}),\mathbb{Q})$ is generated by a loop γ through the node, two loops around the genus, and two loops around the punctures. In the corresponding dual basis for $H^1(X(\mathbb{C}),\mathbb{Q})$, the class γ^* generates $W_0H^1(X(\mathbb{C}),\mathbb{Q})$.

We now consider the nonarchimedean analytification of X with respect to the trivial valuation on $\mathbb{C}$. As in the examples from Section 3, the analytification of the normalization $\widetilde{X}$ of X is an infinite tree, with three unbounded branches corresponding to the punctures. The remaining branches end in leaves, corresponding to the closed points of $\widetilde{X}$ and, in X^{an}, two of these closed points are identified at the node x. The analytification is therefore as shown here.



Note that the identification of two closed points in $\widetilde{X}$ creates an extra loop in both $X(\mathbb{C})$ and X^{an} and that, on $X(\mathbb{C})$, the new class in H^1 has weight zero.

Theorem 3 ([Ber00])**.** *Let X be a complex algebraic variety, and let X^{an} be its analytification with respect to the trivial valuation on $\mathbb{C}$. Then there is a natural isomorphism*

$$H^*(X^{\mathrm{an}}, \mathbb{Q}) \cong W_0 H^*(X(\mathbb{C}), \mathbb{Q}).$$

A similar result holds for varieties defined over a local field, such as $\mathbb{Q}_p$. In these cases, $H^k(X^{\mathrm{an}}, \mathbb{Q}_\ell)$ is canonically identified with the weight zero part of $H^k_{et}(X, \mathbb{Q}_\ell)$.

Singular cohomology of skeletons and their relations to weight filtrations have also appeared in the tropical geometry literature, for instance in [Hac07, HK08, KS10]. The key fact is that tropicalizations are skeletons in the case where all initial degenerations are smooth and irreducible, and there are natural parametrizing complexes for tropicalizations that are skeletons more generally, in the "schön" case, where all initial degenerations are smooth, but possibly reducible. See [Gub12] for details on the relation between tropicalizations, initial degenerations, and formal models.

5.3. **Beyond singular cohomology.** There is far more information in the topology of X^{an} than in its rational homology. For instance, the analytification of a totally degenerate Enriques surface has no ℓ-adic homology of weight zero in degree above zero, but its analytification is not contractible. It has the homotopy type of $\mathbb{RP}^2$. In this case, the fundamental group of X^{an} agrees with the étale fundamental group of X. The exact relationship between the fundamental group of an analytification and the algebraic invariants of the variety is not known in general.

5.4. **The analytic Milnor fiber.** Some constructions on complex varieties that seem more topological than algebraic have semialgebraic analogues in nonarchimedean analytic geometry. A prime example is the *analytic Milnor fiber* of Nicaise and Sebag. Although their construction works more generally, for the Milnor fiber of a point in the special fiber of a one-parameter family of varieties, we consider, for simplicity, just the Milnor of a point on a hypersurface. Here, the implicit one parameter family is the space of varieties defined by $f(x) = t$, where $f(x) = 0$ is the hypersurface. See [Mil68] for details on the classical Milnor fiber, for isolated singularities in a complex hypersurface.

Let X be the hypersurface in $\mathbb{C}^n$ with defining equation $f \in \mathbb{C}[y_1, \ldots, y_n]$, and let x be a point in X. Again, consider the nonarchimedean analytification of X with respect to the trivial valuation. Say that a point of X^{an} specializes to x if it is defined over a valued extension of $\mathbb{C}$ and extends to a point over the valuation ring with special fiber x. The condition of specializing to x is semialgebraic; if $x = (x_1, \ldots, x_n)$ then this condition is equivalent to $\mathrm{val}(y_i - x_i) > 0$ for $1 \leq i \leq n$.

Definition 4. *The analytic Milnor fiber is the semialgebraic set in* X^{an} *consisting of points* x' *over valued extensions of* $\mathbb{C}$ *that specialize to* x *and satisfy* $\mathrm{val}(f(x')) = 1$.

The analytic Milnor fiber is, in an appropriate sense,defined over $\mathbb{C}((t))$, so it carries an action of the absolute Galois group of $\mathbb{C}((t))$. Nicaise and Sebag show that the ℓ-adic étale cohomology of $\mathcal{F}_x$, with the action of the procyclic generator of this Galois group, is canonically identified with the ℓ-adic singular cohomology of the Milnor fiber, with its monodromy action [NS07].

There are multiple advantages to the approach of Nicaise and Sebag. First, their definition makes sense in much greater generality, for varieties over an arbitrary field with the trivial valuation, such as $\overline{\mathbb{F}}_p$, where one does not have the complex topology to work with. Furthermore, over $\mathbb{C}$, their construction puts the Milnor fiber and its monodromy action in a context (semialgebraic sets) where motivic integration makes sense [HK06]. This opens the possibility of a conceptual approach to the monodromy conjectures of Igusa [Igu75] and of Denef and Loeser [DL98].

References

[AB12] O. Amini and M. Baker, *Linear series on metrized complexes of algebraic curves*, preprint arXiv:1204.3508, 2012.

[Abh56] S. Abhyankar, *On the valuations centered in a local domain*, Amer. J. Math. **78** (1956), 321–348.

[AC12] O. Amini and L. Caporaso, *Riemann-Roch theory for weighted graphs and tropical curves*, preprint arXiv:1112.5134v2, 2012.

[AM69] M. Atiyah and I. Macdonald, *Introduction to commutative algebra*, Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1969.

[Bak08a] M. Baker, *An introduction to Berkovich analytic spaces and non-Archimedean potential theory on curves*, p-adic geometry, Univ. Lecture Ser., vol. 45, Amer. Math. Soc., Providence, RI, 2008, pp. 123–174.

[Bak08b] ———, *Specialization of linear systems from curves to graphs*, Algebra Number Theory **2** (2008), no. 6, 613–653.

[Ber90] V. Berkovich, *Spectral theory and analytic geometry over non-Archimedean fields*, Mathematical Surveys and Monographs, vol. 33, American Mathematical Society, Providence, RI, 1990.

[Ber93] ———, *Étale cohomology for non-Archimedean analytic spaces*, Inst. Hautes Études Sci. Publ. Math. (1993), no. 78, 5–161 (1994).

[Ber99] ———, *Smooth p-adic analytic spaces are locally contractible*, Invent. Math. **137** (1999), no. 1, 1–84.

[Ber00] ———, *An analog of Tate's conjecture over local and finitely generated fields*, Internat. Math. Res. Notices (2000), no. 13, 665–680.

[Ber04] ———, *Smooth p-adic analytic spaces are locally contractible. II*, Geometric aspects of Dwork theory. Vol. I, II, Walter de Gruyter GmbH & Co. KG, Berlin, 2004, pp. 293–370.

[BGR84] S. Bosch, U. Güntzer, and R. Remmert, *Non-Archimedean analysis*, Grundlehren der Mathematischen Wissenschaften, vol. 261, Springer-Verlag, Berlin, 1984.

[BN07] M. Baker and S. Norine, *Riemann-Roch and Abel-Jacobi theory on a finite graph*, Adv. Math. **215** (2007), no. 2, 766–788.

[BN09] ———, *Harmonic morphisms and hyperelliptic graphs*, Int. Math. Res. Not. (2009), no. 15, 2914–2955.

[BPR11] M. Baker, S. Payne, and J. Rabinoff, *Nonarchimedean geometry, tropicalization, and metrics on curves*, preprint, arXiv:1104.0320v1, 2011.

[BR10] Matthew Baker and Robert Rumely, *Potential theory and dynamics on the Berkovich projective line*, Mathematical Surveys and Monographs, vol. 159, American Mathematical Society, Providence, RI, 2010.

[Cap12] L. Caporaso, *Gonality of algebraic curves and graphs*, preprint arXiv:1201.6246v3, 2012.

[CC12] W. Castryck and F. Cools, *Newton polygons and curve gonalities*, J. Algebraic Combin. **35** (2012), no. 3, 345–366.

[CDPR12] F. Cools, J. Draisma, S. Payne, and E. Robeva, *A tropical proof of the Brill-Noether theorem*, Adv. Math. **230** (2012), no. 2, 759–776.

[CKK12] G. Cornelissen, F. Kato, and J. Kool, *A combinatorial Li-Yau inequality and rational points on curves*, preprint arXiv:1211.2681, 2012.

[DL98] J. Denef and F. Loeser, *Motivic Igusa zeta functions*, J. Algebraic Geom. **7** (1998), no. 3, 505–537.

[Duc12] A. Ducros, *Les espaces de Berkovich sont modérés, d'après E. Hrushovski et F. Loeser*, preprint arXiv:1210.4336, 2012.

[GK08] A. Gathmann and M. Kerber, *A Riemann-Roch theorem in tropical geometry*, Math. Z. **259** (2008), no. 1, 217–230.

[Gub12] W. Gubler, *A guide to tropicalizations*, preprint, arXiv:1108.6126v2, 2012.

[Hac07] P. Hacking, *Homology of tropical varieties*, preprint, arXiv:0711.1847v2, 2007.

[HHM08] D. Haskell, E. Hrushovski, and D. Macpherson, *Stable domination and independence in algebraically closed valued fields*, Lecture Notes in Logic, vol. 30, Association for Symbolic Logic, Chicago, IL, 2008.

[HK06] E. Hrushovski and D. Kazhdan, *Integration in valued fields*, Algebraic geometry and number theory, Progr. Math., vol. 253, Birkhäuser Boston, Boston, MA, 2006, pp. 261–405.

[HK08] D. Helm and E. Katz, *Monodromy filtrations and the topology of tropical varieties*, To appear in Canad. J. Math. arXiv:0804.3651v2, 2008.

[HL10] E. Hrushovski and F. Loeser, *Nonarchimedean topology and stably dominated types*, preprint, 2010.

[HLP12] E. Hrushovski, F. Loeser, and B. Poonen, *Berkovich spaces embed in euclidean spaces*, preprint, arXiv:1210.6485, 2012.

[Igu75] J.-I. Igusa, *Complex powers and asymptotic expansions. II. Asymptotic expansions*, J. Reine Angew. Math. **278/279** (1975), 307–321.

[KS10] E. Katz and A. Stapledon, *The tropical motivic nearby fiber*, To appear in Compositio Math. arxiv:1007.0511v1, 2010.

[Mar02] D. Marker, *Model theory*, Graduate Texts in Mathematics, vol. 217, Springer-Verlag, New York, 2002, An introduction.

[Mil68] John Milnor, *Singular points of complex hypersurfaces*, Annals of Mathematics Studies, No. 61, Princeton University Press, Princeton, N.J., 1968.

[MZ08] G. Mikhalkin and I. Zharkov, *Tropical curves, their Jacobians and theta functions*, Curves and abelian varieties, Contemp. Math., vol. 465, Amer. Math. Soc., Providence, RI, 2008, pp. 203–230.

[NS07] J. Nicaise and J. Sebag, *Motivic Serre invariants, ramification, and the analytic Milnor fiber*, Invent. Math. **168** (2007), no. 1, 133–173.

[Thu05] A. Thuillier, *Théorie du potentiel sur ler courbes en géométrie analytique non archimédienne. Applications à la theéorie d'Arakelov*, Ph.D. thesis, University of Rennes, 2005.

[Thu07] ———, *Géométrie toroïdale et géométrie analytique non archimédienne*, Manuscripta Math. **123** (2007), no. 4, 381–451.

Yale University Mathematics Department, 10 Hillhouse Ave, New Haven, CT 06511, U.S.A.

E-mail address: sam.payne@yale.edu

GEOMETRIC GROUP THEORY AND 3-MANIFOLDS HAND IN HAND: THE FULFILLMENT OF THURSTON'S VISION FOR THREE-MANIFOLDS

MLADEN BESTVINA

Abstract. In the late 70s, Thurston revolutionized our understanding of 3-manifolds. He stated a far reaching Geometrization Conjecture and proved it for a large class of manifolds, called Haken manifolds. He also posed 24 open problems, describing his vision of the structure of 3-manifolds.

Pieces of Thurston's Vision have been confirmed in the subsequent years. In the meantime, Dani Wise developed a sophisticated program to study cube complexes, and in particular to promote immersions to embeddings in a finite cover. Ian Agol completed Wise's program and as a result essentially all problems on Thurston's list are now solved. In these notes I will outline a proof that closed hyperbolic 3-manifolds are virtually Haken.

1. Introduction

One way to understand surfaces is to successively cut them along incompressible circles and arcs until a collection of disks is obtained. Figure 1 shows this process for the torus.

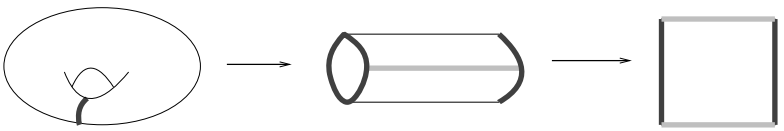


Figure 1. Cutting a torus. [**Please note that when colored lines are used in any of the figures or referenced in the text and/or figure captions, the colors are represented in the figures by the following gray scales: red=25% gray; green=50% gray; and blue=75% gray.**]

By *incompressible* we mean that circles are π_1-injective (or equivalently don't bound disks) and that arcs, whose boundaries are always in the boundary of the surface, do not cobound disks with arcs in the boundary of the surface. The collection of surfaces obtained by successive cuts is the *hierarchy* of the original surface. This process doesn't quite work for the 2-sphere (there are no incompressible circles!), but can be used, for example, to prove

that a homotopy equivalence between two *aspherical*[1] closed surfaces is homotopic to a homeomorphism, by inducting on the hierarchy.

In the 1960's Wolfgang Haken [**Hak62**] introduced the analogous notion of hierarchies for aspherical 3-manifolds. The cuts are required to be along aspherical incompressible surfaces. Haken established that if a closed aspherical 3-manifold admits the first cut, then it has a (Haken) hierarchy terminating in a collection of 3-balls[2]. Manifolds with a hierarchy are called Haken manifolds. Waldhausen [**Wal68**] proved that homotopy equivalent Haken manifolds are homeomorphic. In the same paper Waldhausen points out the possibility that non-Haken aspherical 3-manifolds might have a finite-sheeted covering space which is Haken. This weaker condition would suffice in many applications. This has become known as Waldhausen's virtual Haken conjecture, although I should point out that at the time all known aspherical non-Haken manifolds were of very simple type (small Seifert fibered spaces, i.e. admitting the structure of a circle fibration with 3 singular fibers).

In the late 1970's William P. Thurston completely reshaped 3-manifold theory. He stated his Geometrization Conjecture and proved it for Haken manifolds. He also showed that in some sense generic 3-manifolds are non-Haken, and in particular constructed the first examples of hyperbolic manifolds that were provably non-Haken.

In [**Thu82**] Thurston outlined his vision for 3-manifolds and posed 24 problems. The most famous of these, the Geometrization Conjecture, was settled by Perelman (see Morgan's Current Events lecture in 2004). Other major conjectures that followed were the Tameness Conjecture (Agol [**Agoa**] and Calegari-Gabai [**CG06**]), Ending Lamination Conjecture (Brock-Canary-Minsky [**BCM**]), and the Surface Subgroup Conjecture (Kahn-Marković [**KM**], see Brock's Current Events lecture in 2012).

The virtual Haken conjecture, using Perelman's work, reduces quickly to closed hyperbolic manifolds M. The Kahn-Marković theorem says that $\pi_1(M)$ contains many (quasi-convex) surface subgroups. Peter Scott observed in 1978 [**Sco78**] that in this situation one could prove that M is virtually Haken provided the surface subgroup H is *separable*: for every $g \in \pi_1(M) - H$ there is a finite index subgroup $G < \pi_1(M)$ such that $H < G$ but $g \notin G$. In one dimension lower, this is illustrated in Figure 2.

In the meantime, geometric group theorists have been considering groups and spaces in part motivated by 3-manifolds and hyperbolic geometry. Gromov [**Gro87**] introduced the concept of hyperbolic groups, a rich class of groups that contains fundamental groups of closed hyperbolic manifolds. Dani Wise, with coauthors, had been developing the theory of *special cube complexes*. The important thing is that the hyperbolic fundamental group

[1] A manifold is aspherical if its universal cover is contractible.

[2] One should really say *homotopy 3-balls*, as the Poincaré conjecture was not known at the time.

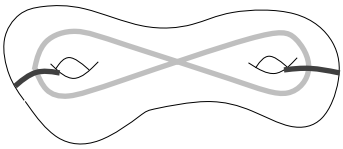


FIGURE 2. The red curve C with $H = \pi_1(C)$ can be lifted to an embedded curve in the double cover of the surface obtained by cutting along the blue curves, taking two copies, and regluing the boundaries of different copies with each other. The group element $g \notin H$ is represented by a half of C, forming a loop based at the intersection point.

of a compact special cube complex has the property that all of its quasi-convex subgroups are separable. On the other hand, using Kahn-Marković and an old construction of Sageev, Bergeron-Wise [**BW**] proved that for every closed hyperbolic 3-manifold there is a non-positively curved (NPC) compact cube complex with the same fundamental group (notice the absence of the adjective "special").

The last step was accomplished by Ian Agol in April 2012 [**Agob**]: *Every compact NPC cube complex with hyperbolic fundamental group is virtually special.* Putting it all together, the following four theorems essentially[3] finish off Thurston's list of problems. Let M be a hyperbolic 3-manifold, complete with finite volume.

(1) M is virtually Haken.
(2) M is *large*, i.e. it has a finite cover whose fundamental groups maps onto the free group F_2.
(3) $\pi_1(M)$ is LERF, i.e. it has the property that every finitely generated subgroup is separable.
(4) M has a finite cover that fibers over the circle.

When M is not closed, but has at least one cusp, it is automatically Haken and Cooper-Long-Reid [**CLR97**] proved that such M is large. It follows from the work of Wise [**Wis**] (which in turn uses the Tameness Conjecture, Agol's fibering criterion and [**CLR97**] that M virtually fibers and that $\pi_1(M)$ is LERF. We will focus on the case when M is closed.

I will discuss (special) cube complexes and Sageev's construction in detail, with a general mathematician in mind. I will then give an outline of Agol's argument, really more of a roadmap for those who wish to take a closer look at the proofs.

Casson had early ideas about the usefulness of cube complexes for 3-manifolds. Gromov's Link Condition (see below) put the theory on the firm footing. Aitchison and Rubinstein considered 3-manifolds that can be given the structure of an NPC cube complex [**AR90**]. Sageev's thesis [**Sag95**] , with his seminal construction of an NPC cube complex from a

[3]There is another problem left open, about volumes of hyperbolic manifolds.

codimension 1 subgroup, marks the beginning of a systematic study of NPC cube complexes.

I will not define the concepts of hyperbolic groups or CAT(0) spaces. They have become ubiquitous in modern mathematics since Gromov's groundbreaking paper [**Gro87**] readers unfamiliar with them are referred to [**BH99**] for a thorough introduction.

Acknowledgements. I would like to thank Michah Sageev for teaching me early on about cube complexes, to Dani Wise for explaining some of his ideas before they became theorems, to Ian Agol for beautiful lectures on which these notes are partially based, and to Jason Manning for patiently explaining even the most trivial points to me and for carefully reading a draft of these notes.

2. NPC cube complexes

An *n-cube* is an isometric copy of $[-1,1]^n$. It has *faces* obtained by fixing some of the coordinates to be ± 1. Each face is naturally a cube of appropriate dimension, e.g. a 2-cube (i.e. a square) has 4 faces of dimension 1 (edges) and 4 faces of dimension 0 (vertices). A *cube complex* is a space obtained from a collection of cubes by isometrically identifying some of the faces. For example, the 2-torus is a cube complex obtained from a square by identifying opposite faces. The reader is referred to Sageev's lecture notes [**Sag**] for a gentle introduction to NPC cube complexes.

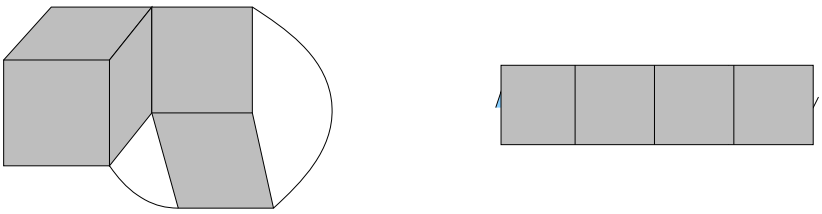


FIGURE 3. Examples of cube complexes. The second is the Möbius band.

We equip each cube with the natural Euclidean metric, and a cube complex with the path metric induced by this metric on the cubes.

The *link* of a vertex in an n-cube is the $(n-1)$-simplex of tangent vectors that point into the cube. The link of a vertex in a cube complex is naturally the union of simplices. The cube complex X is NPC (*non-positively curved*) if:

- the link of every vertex in X is a simplicial complex, which is also a flag complex.

So for example, the double of a square along its boundary is not NPC, since the link of a vertex is the double of an edge along its boundary, and this isn't a simplicial complex. A simplicial complex L is a *flag complex* if it is determined by its 1-skeleton, i.e. if $v_1, \cdots, v_k$ are distinct vertices such

that every pair bounds an edge, then L contains the simplex with vertices $v_1, \cdots, v_k$.[4] For example, the boundary of the 3-cube is not NPC since the links of vertices are hollow triangles. This definition is based on Gromov's *Link Condition* [**Gro87, BH99**] which says that a cube complex is NPC iff it is locally CAT(0). In particular, by (a version of) the Cartan-Hadamard theorem [**Gro87, BH99, Lea**] the universal cover of an NPC cube complex is $CAT(0)$.

2.1. **Hyperplanes.** Two parallel edges of a square in a cube complex are *square equivalent*, and we extend this to an equivalence relation on the set of all edges. In any cube $[-1,1]^n$, a *midcube* is the subset obtained by fixing one of the coordinates to be 0. Thus an n-cube has n midcubes, and each midcube intersects 2^{n-1} edges, all of which are square equivalent. The *hyperplane* dual to an equivalence class of oriented edges is the union of all midcubes that intersect only the edges contained in the equivalence class. A hyperplane is *embedded* if it intersects each cube in at most one midcube. Each midcube has a natural normal bundle, which is a trivial line bundle. These bundles glue together to form a normal bundle of a hyperplane. The hyperplane is 2-sided [1-sided] if the normal bundle is trivial [nontrivial].

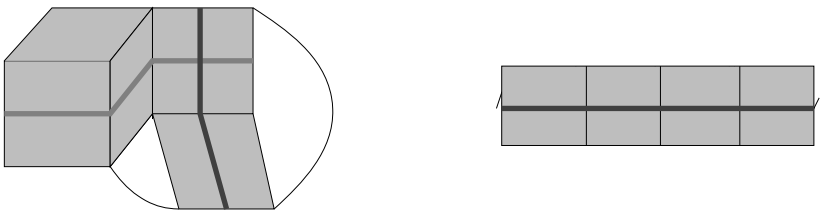


FIGURE 4. Examples of hyperplanes. In the first example the green hyperplane consists of a square and a segment. The hyperplane in the second example is 1-sided.

If X is CAT(0) then hyperplanes are embedded and they have the structure of CAT(0) cube complexes.

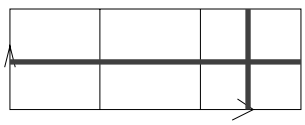


FIGURE 5. Self intersecting hyperplane.

[4]Warren Dicks succinctly phrases the condition like this: a non-simplex contains a non-edge.

3. Sageev's construction

The great thing about NPC cube complexes is that they keep track of intersections between hypersurfaces. First recall that if $S \subset M$ is an incompressible surface in a 3-manifold (or a circle in a surface) then $\pi_1(M)$ acts on a tree T encoding the pattern of the components of the preimage $\tilde{S}$ of S in the universal cover $\tilde{M}$ of M (these components are copies of the universal cover of S). The vertices of T are represented by the components of $\tilde{M} - \tilde{S}$ and the edges by the components of $\tilde{S}$. The stabilizer of an edge of T is $\pi_1(S)$, and the tree T encodes the pattern seen in $\tilde{M}$. This construction is part of *Bass-Serre theory* (see e.g. [**SW79**].

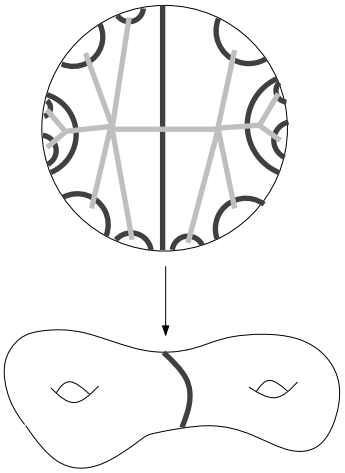


Figure 6. Construction of the Bass-Serre tree.

Now suppose that S is only immersed in M, but the preimage $\tilde{S} \subset \tilde{M}$ consists of (embedded, but possibly intersecting) copies of the universal cover of S. Michah Sageev [**Sag95**] discovered that this time there is a CAT(0) cube complex $\tilde{X}$ encoding the pattern in $\tilde{M}$.

Here is a more abstract version of Sageev's construction, due to Haglund-Paulin [**HP98**].

Let Y be a set. A *wall* is a partition of Y into two subsets. A *wall set* is a pair $(Y, \mathcal{W})$ where $\mathcal{W}$ is a collection of walls. We require that any two points of Y are separated by finitely many, and at least one, wall.

A *halfspace* determined by a wall is one of the two subsets in the partition. An ultrafilter ω is a collection of halfspaces so that

- if $W = \{A, A^c\}$ then exactly one of A, A^c is in ω,
- if $W = \{A, A^c\}$, $W' = \{B, B^c\}$, $A \subset B$ and $A \in \omega$ then $B \in \omega$.

We think of ω as making a choice of a halfspace determined by each wall. For every $y \in Y$ we have the ultrafilter ω_y that chooses the halfspace

containing y. If ω, ω' are two ultrafilters, define the *distance* $d(\omega, \omega')$ as the number of walls where ω, ω' made different choices (this is possibly infinite).

For any $y, y' \in Y$ the ultrafilters $\omega_y, \omega_{y'}$ are finite distance apart. Let V be the set of all ultrafilters finite distance apart from each ω_y. Then V is the 0-skeleton of the Sageev cube complex $\tilde{X}$ associated with the wall space. Two vertices v, v' are connected by an edge if they differ in exactly one wall. An n-cube is determined by n walls $W_1, \cdots, W_n$ and its vertices all agree on every other wall, but all 2^n choices on W_i's represent vertices of the cube.

Example 3.1. Let Y have 7 points with 3 walls, see Figure 7. There are 8 ultrafilters: 7 of the form ω_y, plus an "ideal" ultrafilter that always picks the halfspace that does not contain the middle point. The Sageev cube complex for this example is the 3-cube.

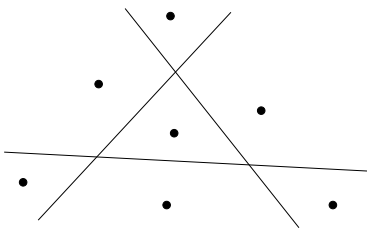


FIGURE 7. The 7 points example.

Two walls $W = \{A, A^c\}$, $W' = \{B, B^c\}$ *intersect* if all 4 intersections of halfspaces are nonempty: $A \cap B \neq \emptyset, A \cap B^c \neq \emptyset, A^c \cap B \neq \emptyset, A^c \cap B^c \neq \emptyset$.

Proposition 3.2.
- *$\tilde{X}$ is a CAT(0) cube complex.*
- *There is a natural injection $Y \to V$ and bijection $\mathcal{W} \to$ hyperplanes and the latter is stabilizer-preserving.*
- *If Y is the set of the vertices of a CAT(0) cube complex $\tilde{X}$ and walls are hyperplanes, the Sageev cube complex is $\tilde{X}$.*
- *If there is an n-cube corresponding to the walls $W_1, \cdots, W_n$ then $W_1, \cdots, W_n$ pairwise intersect.*

We will apply Sageev's construction in the context of a closed hyperbolic 3-manifold (or a surface) M and a finite collection of incompressible surfaces (or circles) S_i immersed in M. We assume that in the universal cover $\tilde{M} = \mathbb{H}^3$ the preimage of the S_i's consists of embedded copies of the universal covers of the S_i's and that they intersect transversally. Let Y' be the set of complementary components of the preimage of $\cup S_i$ in $\tilde{M}$, and for the walls take the collection of copies of universal covers of S_i. There may be distinct points in Y' not separated by any walls, so for Y take the set of equivalence classes of points in Y', with two points equivalent if they are not separated by any walls.

Proposition 3.3. [**Sag97**] *In this situation, if each $\pi_1(S_i)$ is quasi-convex in $\pi_1(M)$, the action of $\pi_1(M)$ on $\tilde{X}$ is cocompact.*

The quasi-convexity condition means the following: if W is a copy of the universal cover of S_i in $\tilde{M}$, then there is $R > 0$ so that any geodesic in $\tilde{M} = \mathbb{H}^3$ with endpoints in W stays in the R-neighborhood of W.[5] We recall a theorem from [**GMRS98**]: for any R there is k so that given k walls at least two are distance $> R$ apart (they prove this more generally in the context of hyperbolic groups and quasi-convex subgroups). In particular, the collection of walls satisfies *Scott's k-plane property: for any k distinct walls there are two that are disjoint.* Proposition 3.3 now follows from the last bullet of Proposition 3.2.

4. Special cube complexes

Here is the key definition of Haglund-Wise [**HW08**].

Definition 4.1. An NPC cube complex is *special* if

- hyperplanes are embedded and 2-sided, and
- there are no direct self-osculations nor inter-osculations.

Two oriented edges with the same initial vertex are *perpendicular* if they are connected in the link, i.e. span a square, otherwise they *osculate*. Orienting the normal bundle of a 2-sided hyperplane H assigns an orientation to every edge dual to H. A 2-sided embedded hyperplane *directly self-osculates* if it has dual osculating (oriented) edges. It *indirectly self-osculates* if there are osculating dual edges intersecting hyperplane with opposite orientations. Two hyperplanes *inter-osculate* if they have perpendicular dual edges that are square equivalent to osculating dual edges. Hyperplanes can be made 2-sided by passing to double covers. Likewise, indirect self-osculations can be removed by subdivision or passing to a double cover.

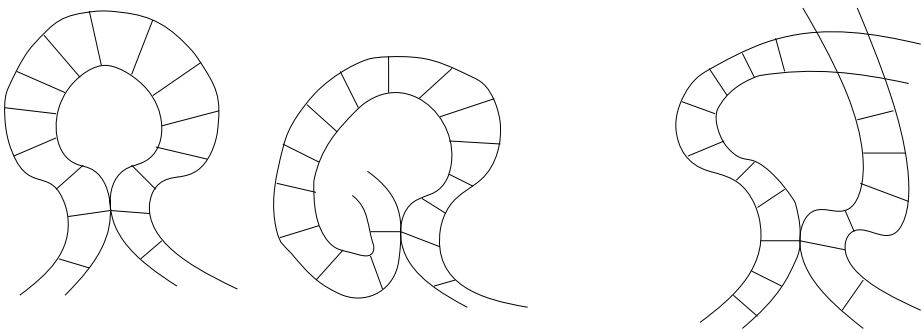


Figure 8. Direct, indirect and inter-osculations

[5] More generally, a subgroup H of a hyperbolic group G is quasi-convex if there is R so that any geodesic in the Cayley graph of G with endpoints in H stays in the R-neighborhood of H.

Achieving embedded hyperplanes and removing the other two kinds of osculations in a finite cover requires separability properties of fundamental groups of hyperplanes and their double cosets. For hyperbolic groups we have the following criterion: *A compact NPC cube complex with hyperbolic fundamental group is virtually special iff all quasi-convex subgroups are separable.* See [**HW08**]; this uses [**Min05**] (for one direction, see Proposition 4.3 below). In particular, virtual specialness of a compact NPC cube complex with hyperbolic π_1 depends only on π_1.

Notice the analogy between special cube complexes (walls are embedded) and Haken 3-manifolds.

4.1. **RAAGs.** The model special cube complexes come from Right Angled Artin Groups (RAAGs). Let Γ be a simplicial graph. Construct a cube complex X_Γ as follows. It has only one vertex v. There is one edge e_a for every vertex a of Γ, and there is one 2-cell $Q_{a,b}$ for every edge $[a,b]$ of Γ: it is obtained by attaching a square via the commutator relation $e_a e_b e_a^{-1} e_b^{-1}$. Thus $Q_{a,b}$ is a 2-torus. We then fill in higher dimensional cells: if the vertices $a, b, c \in \Gamma$ are pairwise joined by edges, then $Q_{a,b} \cup Q_{a,c} \cup Q_{b,c}$ is the 2-skeleton of a 3-torus, and we glue in a 3-cell $Q_{a,b,c}$ by identifying opposite faces of a 3-cube and gluing them to the 2-cells. Proceed similarly in higher dimensions.

For example, if Γ is a finite set, X_Γ is a wedge of circles. If Γ is a complete graph on n vertices, X_Γ is the n-torus. The fundamental group of X_Γ is the RAAG associated to Γ, given by the presentation:

$$\langle g_v, v \in \Gamma^{(0)} \mid g_v g_w = g_w g_v, [v,w] \subset \Gamma \rangle$$

that is, the generators are in 1-1 correspondence with the vertices of Γ, and two generators commute when the corresponding vertices of Γ are joined by an edge. The complex X_Γ is a special NPC cube complex.

Now RAAGs are *universal* special cube complexes, in the sense that any special cube complex X can be locally isometrically immersed in some X_Γ. The construction of Γ is simple: the vertices are the hyperplanes of X and edges represent intersecting hyperplanes. The definition of "special" is designed so that the obvious map that sends each vertex of X to the only vertex of X_Γ and each edge of X to the edge of X_Γ representing the dual hyperplane (and being careful with orientations) extends to a local isometry $X \to X_\Gamma$.

When Γ is finite, the RAAG $\pi_1(X_\Gamma)$ is linear, in fact it can be realized as a subgroup of $SL_N(\mathbb{Z})$ for large N. This follows from [**DJ00**].

Theorem 4.2. [**HW08**] *If X, Y are finite special cube complexes and $f : X \to Y$ is a locally isometric immersion, there is a finite cover $\hat{Y} \to Y$ to which f lifts and the image of X in $\hat{Y}$ is a retract of $\hat{Y}$.*

The proof is elementary and is modeled on the case of graphs (they are special cube complexes!), which is how Stallings [**Sta83**] proved the Marshall Hall theorem for free groups. See also [**Hag08**].

Sketch of the proof. First consider the special case when $Y = X_\Gamma$ and $X \to Y$ is constructed above. To construct $\hat{Y}$ look at the components of the preimages of edges. For simplicity, assume that X has no indirect self-osculations as well. Then these components are circles and arcs, each consisting of a single edge (in general, they will consist of several coherently oriented edges). Leave circle components alone. To each arc component attach an arc making it a circle and map it to the edge (circle) in Y by double cover. The retraction takes this new arc to the old arc rel endpoints. Now proceed with 2-cells. Each component of the preimage contains one 2-cell, which is a square, an annulus, or a torus. Turn it into a torus by adding 2-cells (1-skeleton is already in place), so the resulting torus will have 4, 2 or 1 2-cells and will map to the target 2-cell (also a torus) by a covering map of degree 4,2 or 1. Proceed analogously in higher dimensions.

Now in general, first immerse Y in a RAAG complex Z as above. Then construct $\hat{Z}$ as in the special case for $X \to Z$ and then let $\hat{Y}$ be the pull-back of $\hat{Z}$ to Y. □

4.2. **Consequences of being special.**

Proposition 4.3. *Let X be a compact virtually special cube complex with $G = \pi_1(X)$ hyperbolic. Then G satisfies:*

(1) quasi-convex subgroups are virtual retracts, hence separable,
(2) G is linear, in fact it embeds in $GL_n(\mathbb{Z})$,
(3) G is large or virtually abelian,
(4) G is RFRS.

Sketch of proof. For (1), the key is that a quasi-convex subgroup $H \subset G$ can be represented by a locally isometric immersion $Z \to X$ of finite special cube complexes. This follows from Haglund [**Hag08**] (the proof is a pleasant exercise in hyperbolic geometry: show that the intersection of half-spaces containing a given orbit is contained in a Hausdorff neighborhood of the orbit). Then (1) follows from Theorem 4.2 (it is not hard to see that virtual retracts in residually finite groups are separable).

For (2) use the fact that RAAGs are linear and that local isometric embeddings are π_1-injective.

For (3), find a quasi-convex free subgroup, using the Tits alternative (plus that fact that solvable subgroups of CAT(0) groups are virtually abelian) and a ping-pong argument and apply (1).

RFRS is a group-theoretic property satisfied by subgroups of RAAGs and if it holds for a closed hyperbolic 3-manifold M (or even complete, finite volume), then M virtually fibers over the circle. This is Agol's *criterion for virtual fibering* [**Ago08**]. □

5. Dehn fillings

Here we discuss the last general prerequisite. The motivation for this theorem is the celebrated theorem of Thurston: *If M is a hyperbolic manifold*

with a cusp, then all but finitely many Dehn fillings produce closed hyperbolic manifolds. Recall that M is the interior of a compact manifold whose boundary is a torus (assuming orientability). By a Dehn filling along a non-trivial circle a in the boundary torus we mean attaching a 2-handle (i.e. a thickened disk) along a and then attaching a 3-ball to the boundary 2-sphere of the resulting 3-manifold. The curve a is defined only up to isotopy, and the set of possible choices is naturally $\mathbb{Q} \cup \{\infty\}$, i.e. the slope of a, once the coordinates on the torus are chosen.

Another, more simple-minded and 2-dimensional motivation for the theorem below, is the classical small cancellation theory (for a nice exposition, see e.g. [**Str90**]).

Now let G be a hyperbolic group and $H_1, \cdots, H_k$ a collection of quasi-convex subgroups. We will assume that this collection is *almost malnormal*, i.e. $gH_ig^{-1} \cap H_j$ infinite implies $i = j$ and $g \in H_i$.

By a *Dehn filling* of G we will mean passing to the quotient

$$\overline{G} = G/\langle\langle N_1 \cup N_2 \cup \cdots \cup N_k \rangle\rangle \qquad (*)$$

where we mod out the normal closure of the union of specified normal subgroups $N_i \trianglelefteq H_i$. For simplicity we will always have that N_i has finite index in H_i.

The following is the main Dehn Filling Theorem. See [**Osi07, DGO**], and also [**GM08**] for the torsion-free case. We are stating only a special case relevant to us. For the quasi-convexity statement see [**AGM09**] and [**MMP10**].

Theorem 5.1. *Let G be a hyperbolic group and $H_1, \cdots, H_k$ a finite almost malnormal collection of quasi-convex subgroups, all contained in a quasi-convex subgroup $H < G$. Let F be a finite subset of G. Then for any sufficiently long filling* (*) the quotient $\phi : G \to \overline{G}$ satisfies:

- $Ker(\phi|H_i) = N_i$,
- $\overline{G}$ is hyperbolic.
- $\phi|F$ is injective.
- $\phi(F) \cap \phi(H_i) = \phi(F \cap H_i)$.
- The image of H in $\overline{G}$ is quasi-convex.

By a *sufficiently long filling* we mean that there exists a finite bad set $B \subset G - \{1\}$ so that the conclusions hold whenever $N_i \cap B = \emptyset$ for every i. Recall that N_i has finite index in H_i.

6. Hierarchies

We will say that a hyperbolic group G is *virtually special* if it acts with finite point stabilizers on a CAT(0) cube complex $\tilde{X}$ so that for a torsion-free finite index subgroup $H < G$ the quotient $\tilde{X}/H$ is a compact special cube complex.

The following two theorems are rather deep. The first combines theorems of Hsu-Wise and of Haglund-Wise. The reader may want to look at the

informal discussion of these theorems in [**Wis**], e.g. the case when A and B are free groups and C is cyclic.

Theorem 6.1 (The Hsu-Wise Combination Theorem [**HWc**] and the Haglund-Wise Combination Theorem [**HWa**])**.** *Suppose G is hyperbolic, $G = A *_C B$ or $G = A*_C$, C is quasi-convex and almost malnormal in G and A, B are virtually special. Then G is virtually special.*

Following Wise [**Wis**] we define the class of groups $\mathcal{AMQH}$. These are hyperbolic groups that admit an *almost malnormal quasi-convex hierarchy*:

- $1 \in \mathcal{AMQH}$
- If $A, B \in \mathcal{AMQH}$ and C is almost malnormal and quasi-convex in the hyperbolic group $G = A *_C B$ $[G = A*_C]$ then $G \in \mathcal{AMQH}$.

Thus $G \in \mathcal{AMQH}$ implies G is hyperbolic virtually special and G hyperbolic virtually special implies G is virtually in $\mathcal{AMQH}$ (by taking a cover where wall groups are malnormal).

Thus $\mathcal{AMQH}$ is something like a group-theoretic version of Haken hyperbolic 3-manifolds, but it contains many non-3-manifold groups. This is crucial for the proof of Theorem 6.3 below, since in the course of the argument one passes to quotient groups that are not necessarily 3-manifold groups, even if the original group G is a 3-manifold group. But one stays in the class of virtually special groups, thanks to the following amazing theorem of Wise, which can be regarded as an addendum to the Dehn Filling Theorem.

Theorem 6.2 (Wise's Malnormal Special Quotient Theorem, [**Wis**])**.** *Under the assumptions of the Dehn Filling Theorem* 5.1*assume that G is also virtually special. Then all sufficiently deep fillings have a virtually special quotient.*

Again, the reader is invited to follow the discussion in [**Wis**] and to think about the special case when G is a free group and H_i are cyclic. By "sufficiently deep" I mean that there are finite index subgroups $H_i' < H_i$ so that if $N_i < H_i'$ then the conclusion follows.

We remark here that it is a major open problem whether hyperbolic groups are residually finite. To apply the Dehn Filling Theorem, one needs to know that the groups H_i are residually finite or else there is no way to choose the normal subgroups N_i to avoid the bad set B. The Malnormal Special Quotient Theorem makes inductive arguments possible as it keeps all groups residually finite. The following key ingredient is proved using such an induction.

Theorem 6.3 (The Agol-Groves-Manning Weak Separation Theorem [**Agob**])**.** *Let G be a hyperbolic group, H a quasi-convex and virtually special subgroup of G, and $g \in G - H$. Then there is a hyperbolic quotient $\pi : G \to \overline{G}$ so that $\pi(H)$ is finite and $\pi(g) \notin \pi(H)$.*

Idea of proof. If H is almost malnormal, the statement follows from the Dehn Filling Theorem. Now imagine that H is a surface subgroup in a hyperbolic 3-manifold group G. Then H may not be almost malnormal since for example there could be elements $\gamma \in G - H$ such that $H^\gamma \cap H$ is infinite cyclic, corresponding to copies of the universal covers of the surface intersecting in lines (but let's say these are the only obstructions to malnormality). If each such infinite cyclic subgroup is replaced by the maximal infinite cyclic subgroup containing it and then conjugates are removed, the resulting collection is finite and almost malnormal. Now apply the Dehn Filling Theorem and the Malnormal Special Quotient Theorem to pass to a hyperbolic quotient G' of G so that the image g' of g is outside the image H' of H, and so that H' is virtually special. We are now in a similar situation as before, but H' is malnormal so we can finish with another application of the Dehn Filling Theorem.

In general, the proof runs by induction on the *height* of H in G. This is the largest n such that $H^{g_1} \cap \cdots \cap H^{g_n}$ is infinite and the cosets $g_i H$ are all distinct. The Gitik-Mitra-Rips-Sageev theorem [**GMRS98**] says that the height of a quasi-convex subgroup of a hyperbolic group is always finite. The first step is to apply the Dehn Filling Theorem and the Malnormal Special Quotient Theorem to the collection of subgroups obtained from minimal infinite intersections of conjugates of H by first replacing them by maximal finite index overgroups, and then discarding conjugates. The quotient groups satisfies all the hypotheses, but the image of H has smaller height than before. □

7. Hyperbolic 3-manifolds and NPC cube complexes

The key in this program is that hyperbolic 3-manifold groups act properly on cubings. This fact was established soon after the seminal construction of Kahn-Marković [**KM**] of nearly totally geodesic immersed surfaces in closed hyperbolic 3-manifolds.

Theorem 7.1. [**BW**] *Let M be a closed hyperbolic 3-manifold. Then $G = \pi_1(M)$ is the fundamental group of a compact NPC cube complex.*

Sketch of proof. Fix a compact fundamental domain F for the action of G on $\mathbb{H}^3$. There is $\epsilon > 0$ so that every geodesic in $\mathbb{H}^3$ that intersects F has endpoints at distance $> \epsilon$ in the boundary S^2_∞ of $\mathbb{H}^3$. Fix a finite collection of round circles in S^2_∞ so that every pair of points at distance $\geq \epsilon$ is separated by at least one circle.

Kahn-Marković [**KM**] show that for every round circle in S^2_∞ there is a quasiconvex surface subgroup H whose limit set is a fractal circle contained in an arbitrarily small annular neighborhood of it and going around the annulus once. This subgroup is represented by an immersed surface in M which is nearly totally geodesic and its universal cover embeds in $\mathbb{H}^3$. Now for each circle in our collection choose a Kahn-Marković surface subgroup H_i whose limit set approximates it. Let $\tilde{X}$ be the Sageev complex associated

to this finite collection of quasi-convex subgroups, and let $X = \tilde{X}/G$. Then X is compact (see Proposition 3.3). To see that the action is free, observe that for every $g \in G - \{1\}$ some conjugate of g will have its axis in $\mathbb{H}^3$ that intersects F, so its endpoints will be separated by one of the round circles, and hence will lie on opposite sides of the plane covering one of the Kahn-Marković surfaces. A high power of g will thus take a half space properly into itself and this implies that g acts hyperbolically on $\tilde{X}$. See [**HWb**]. □

8. Wise's conjecture, Agol's theorem

Let M be a closed hyperbolic 3-manifold. As a consequence of Theorem 7.1 we have $\pi_1(M) = \pi_1(X)$ for a compact NPC cube complex X. If we could show that X is virtually special we would be done by a classical argument. For example, we could use subgroup separability and find a finite cover of M to which a Kahn-Marković surface lifts to an embedding. Alternatively, a finite cover of X has a splitting along a hyperplane and one could use a construction of Stallings ("pull back and compress") to find an incompressible surface in the corresponding finite cover of M. Thus the following theorem finishes the program.

Theorem 8.1 (Agol (2012) [**Agob**], Wise's Conjecture (2011) [**Wis**])**.** *Every finite NPC cube complex with hyperbolic π_1 is virtually special.*

Note that this fails without hyperbolicity. An extreme example is the group constructed by Burger and Mozes [**BM97**]—it is a simple group acting cocompactly on the product of two trees. Since in particular it is not residually finite, it cannot be virtually special. There is an earlier example of Wise [**Wis07**] of such a group which also doesn't have nontrivial subgroups of finite index (but isn't simple).

In the remainder of the paper I will outline a proof of Agol's theorem.

8.1. **Coloring Lemma.** An *n-coloring* of a graph is a function from the vertex set to $[n] := \{1, \cdots, n\}$ that assigns different numbers (colors) to the endpoints of each edge. Note that any countable graph with valencies bounded by k has a $(k+1)$-coloring.

Proposition 8.2. *Let $\mathcal{G}$ act on a countable simplicial graph Γ with valencies bounded by k. Then there is a $\mathcal{G}$-invariant probability measure on the space of $(k+1)$-colorings of Γ.*

Proof. The space $[n]^\Gamma$ of all functions $\Gamma^0 \to \{1, \cdots, n\}$ from the vertex set of Γ is a Cantor set and the space of n-colorings is a closed subset, and it is nonempty when $n > k$. Let μ_n be the uniform (product) measure on $[n]^\Gamma$ and note that it assigns $1/n$ to the subset where a particular pair of neighbors are assigned the same color. For $n > k+1$ define $[n]^\Gamma \to [n-1]^\Gamma$ by replacing color n at any vertex v with the minimal color not represented by any neighbors of v. Compose these to get to $[k+1]^\Gamma$. The feature of this construction is that if an edge has vertices of different colors at the start,

it also does at the end. These maps are continuous (since they are locally defined) so we may push the uniform measure to get a $\mathcal{G}$-invariant measure ν_n on $[k+1]^\Gamma$ that still assigns $\leq 1/n$ to the set where a particular edge gets one color. Take a subsequence that converges to a limiting measure ν. This is still $\mathcal{G}$-invariant and assigns 0 to the set that assigns one color to a particular edge. So the support of ν is contained in the infinite intersection, one for every edge, where the endpoints get different colors, i.e. it is a measure on the set of colorings of Γ. □

In the application below $\mathcal{G}$ will act properly (and cocompactly) on Γ. If $\mathcal{G}$ is also residually finite, there is a finite index subgroup $\mathcal{G}' < \mathcal{G}$ that acts freely with $\Gamma/\mathcal{G}'$ a simplicial graph, so there is a $\mathcal{G}'$-equivariant coloring of Γ (and hence a $\mathcal{G}$-invariant measure with support in a finite set of colorings). In fact, with a bit more work, one can arrange that $\mathcal{G}$ is hyperbolic, as follows. A stronger version of the Weak Separation Theorem 6.3 (which can be proved in the same way) states that H can be separated from any finite collection of elements outside H (instead of just one element). Then the quotient $G \to \mathcal{G}$ can be constructed with a single application of this theorem, by taking for H the free product of suitable conjugates of hyperplane groups and conjugating in the same way the elements that need to be separated from these hyperplane groups. The Coloring Lemma is a way of getting around the issue of residual finiteness of hyperbolic groups.

8.2. **Setting the stage.** Now assume X is a compact NPC cube complex with $G = \pi_1(X)$ hyperbolic. The goal is to show that X is virtually special. The proof is by induction on dimension of X, so we may assume that the wall[6] groups are virtually special. The first step is to construct a (most likely infinite) regular cover $\hat{X} \to X$ with deck group $\mathcal{G}$ so that $\hat{X}$ is special and its walls are compact. This is accomplished by using the Weak Separation Theorem 6.3. For the subgroup H one can take the free product of suitable conjugates of wall groups (for one wall in each orbit). There are finitely many elements of G that need to be separated from H; apply the Weak Separation Theorem for each such element and then intersect the kernels. The cover $\hat{X}$ is the one that corresponds to this intersection. The main feature of the walls W in $\hat{X}$ is that there are no essential annuli $(S^1 \times [0,1], S^1 \times \{0,1\}) \to (\hat{X}, W)$ (i.e. any such map that sends S^1 to an infinite order element in $\pi_1(\hat{X})$ is homotopic into (W, W)). See Figure 9.

8.3. **Decomposing into Puzzle Pieces.** Cutting $\hat{X}$ along walls results in a collection of compact pieces; we will call them *puzzle pieces*. They are simply stars of vertices with respect to the barycentric subdivision of $\hat{X}$ in which an n-cube $[-1,1]^n$ gets subdivided into 2^n subcubes of edge-length 1. A *face* of a puzzle piece is its intersection with a wall (if nonempty). The puzzle pieces will serve as building blocks for the construction of a finite cover

[6]I should really call them hyperplanes, but at this point there should be no confusion and I will use the shorter word.

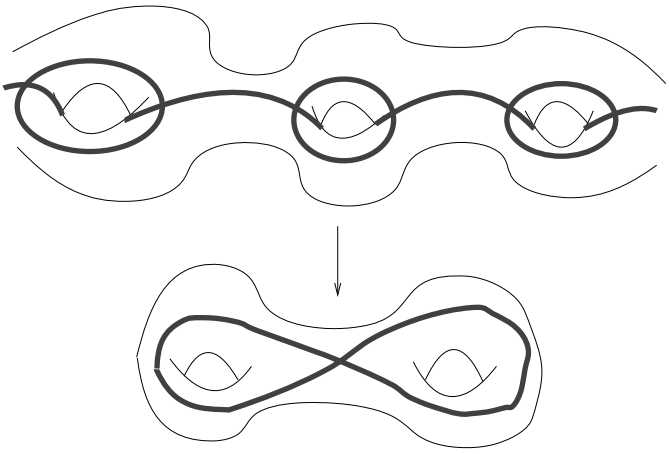


FIGURE 9. A schematic picture of $\hat{X} \to X$ and an immersed wall in X whose preimage consists of embedded, but intersecting, compact walls.

of X, but we will first color their faces in order to encode gluing information. Let Γ be the graph whose vertices are walls in $\hat{X}$ and whose edges correspond to pairs (W, W') of walls such that either $W \cap W' \neq \emptyset$ or there is an essential annulus in $\hat{X}$ with one boundary component in W and the other in W'. By [**GMRS98**] (see the discussion following Proposition 3.3) in the latter case necessarily the distance between W and W' is bounded, so Γ has bounded valence, say by k, and the deck group $\mathcal{G}$ acts on Γ cocompactly. We may apply the Coloring Lemma to find a $\mathcal{G}$-invariant probability measure ν on the space of all $(k+1)$-colorings of Γ.

Let c be a coloring. Thus each wall is assigned a color in $\{1, 2, \cdots, k+1\}$. We can cut $\hat{X}$ along walls, first along walls numbered 1, then along walls numbered 2 etc. This gives a particular hierarchy for $\hat{X}$. If W is a wall, the *descending set of walls for* (W, c) is the following finite collection of walls: it contains W, and all walls W' with $(W, W') \in \Gamma$ and $c(W') < c(W)$, and inductively, whenever V is a wall in the collection, add all walls V' with $(V, V') \in \Gamma$ and $c(V') < c(V)$.

If c, c' are two colorings and W is a wall, we will write $(W, c) \sim (W, c')$ if the descending sets of walls for (W, c) and (W, c') are identical and they receive the same color from both c and c'. Likewise, if P is a puzzle piece, we say $(P, c) \sim (P, c')$ if $(W, c) \sim (W, c')$ for every wall W that intersects P. We will write equivalence classes as $[W, c]$ and $[P, c]$.

By a *Colored Puzzle Piece* we will mean an equivalence class $[P, c]$. Informally, a colored puzzle piece is a cubical polyhedron whose faces are colored with the additional information of the coloring of the descending set of walls for each face of P. There are finitely many $\mathcal{G}$-orbits of puzzle pieces. See Figure 11.

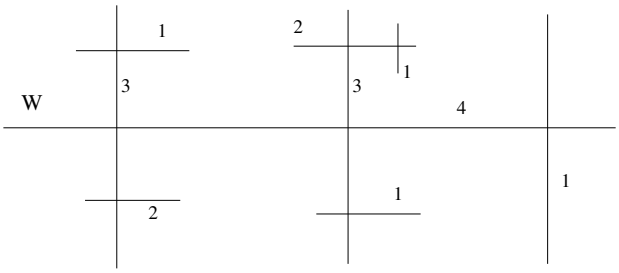


FIGURE 10. A schematic picture of a possible descending set of the wall W colored 4. For the example, we imagine that there are no essential annuli anywhere.

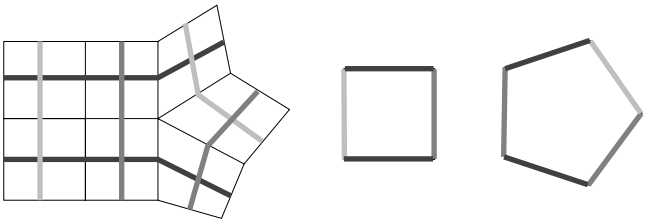


FIGURE 11. An example of a coloring of the walls and two of the resulting colored puzzle pieces. The equivalence class $[P, c]$ remembers a bit more than just the colors of the faces – this depends on the ordering of the colors and is suppressed in the figure.

8.4. **Gluing equations.** We want to take a finite collection of colored puzzle pieces with appropriate multiplicities, say ω, and glue them together to form a finite cover of X. In order to do the gluing, we need the pieces to match up, i.e. ω must satisfy the *gluing equations*.

Let F be a common face of two puzzle pieces P, Q and let c be a coloring. Then F is contained in a unique wall W and for simplicity we will write $[F, c]$ and $c(F)$ for $[W, c]$ and $c(W)$. We have a gluing equation

$$\sum_{[P,d]|(F,d)\sim(F,c)} \omega([P, d]) = \sum_{[Q,d]|(F,d)\sim(F,c)} \omega([Q, d])$$

Notice that there are finitely many $\mathcal{G}$-orbits of equations, so if we impose the requirement that the multiplicities ω are $\mathcal{G}$-invariant, there are only finitely many equations and finitely many unknowns. Also note that $\omega = \nu$ (the measure from the Coloring Lemma) is a nonnegative real solution to all of them (for a fixed F the set of colorings d with $(F, d) \sim (F, c)$ can be written as the disjoint union in two different ways and the equation is a consequence of the additivity of measures). Thus by linear algebra there is a (nonzero) nonnegative $\mathcal{G}$-invariant *integral* solution to all the equations.

8.5. **Virtual Regluing.** Now we have a finite collection of colored puzzle pieces with multiplicities satisfying the gluing equations. We will glue them back so they form a finite cover of X, but during the process we may have to pass to finite covers, so the number of colored puzzle pieces at the end of the process will be some multiple of the original.

In the first step, glue colored puzzle pieces along faces colored $k+1$. That is, to glue $[P, c]$ and $[Q, d]$ along a common (up to the action of $\mathcal{G}$) face F, we require $(F, c) \sim (F, d)$, $c(F) = d(F) = k + 1$ (and that F receives opposite transverse orientations from P and Q). That we can match up all faces of our puzzle pieces colored $k + 1$ follows from the fact that multiplicities satisfy the gluing equations. Notice that the gluing also respects the colors of the faces adjacent to F resulting in bigger colored puzzle pieces, whose faces now carry colors $1, 2, \cdots, k$.

The idea is now to continue this process of gluing. There is another issue that comes up, see Figure 12. The faces colored with k to be glued together are finite covers of the same complex Y_k, but may not be "the same".

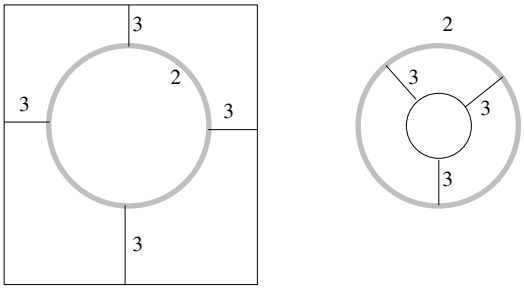


FIGURE 12. An example with $k = 3$. Imagine having 3 copies of the puzzle piece on the left and 4 copies of the piece on the right (both are annuli). The red "faces", colored 2, cover the circle Y_2 with total degree 0, but no gluing is possible.

This is resolved by passing to a finite cover which is *regular* with respect to Y_k. That such a cover can be extended to a cover of the puzzle pieces is Agol's *Gluing Theorem* and it follows for example from Theorem 4.2 after noting that these pieces are equipped with a hierarchy by construction, so they are virtually special. The extra information about cylinders encoded in Γ and about descending set of walls encoded in $[P, c]$ is used to ensure that this hierarchy is almost malnormal.

Proceeding inductively, glue in this way all the colored puzzle pieces, along faces colored from highest to lowest color. The resulting complex is a finite cover of X and it comes with a hierarchy, so it is virtually special. Note here that the hierarchy is in fact almost malnormal, since by construction there are no essential annuli with both boundary components mapped to the same color. This finishes the proof of Agol's theorem.

References

[AGM09] Ian Agol, Daniel Groves, and Jason Fox Manning. Residual finiteness, QCERF and fillings of hyperbolic groups. *Geom. Topol.*, 13(2):1043–1073, 2009.

[Agoa] Ian Agol. Tameness of hyperbolic 3-manifolds. math/0405568.

[Agob] Ian Agol. The virtual Haken conjecture, with an appendix by Agol, Groves and Manning. arXiv:1204.2810.

[Ago08] Ian Agol. Criteria for virtual fibering. *J. Topol.*, 1(2):269–284, 2008.

[AR90] I. R. Aitchison and J. H. Rubinstein. An introduction to polyhedral metrics of nonpositive curvature on 3-manifolds. In *Geometry of low-dimensional manifolds, 2 (Durham, 1989)*, volume 151 of *London Math. Soc. Lecture Note Ser.*, pages 127–161. Cambridge Univ. Press, Cambridge, 1990.

[BCM] Jeffrey F. Brock, Richard D. Canary, and Yair N. Minsky. The classification of Kleinian surface groups, II: The Ending Lamination Conjecture. math/0412006.

[BH99] Martin R. Bridson and André Haefliger. *Metric spaces of non-positive curvature*, volume 319 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1999.

[BM97] Marc Burger and Shahar Mozes. Finitely presented simple groups and products of trees. *C. R. Acad. Sci. Paris Sér. I Math.*, 324(7):747–752, 1997.

[BW] Nicolas Bergeron and Daniel T. Wise. A boundary criterion for cubulation. arXiv:0908.3609.

[CG06] Danny Calegari and David Gabai. Shrinkwrapping and the taming of hyperbolic 3-manifolds. *J. Amer. Math. Soc.*, 19(2):385–446, 2006.

[CLR97] D. Cooper, D. D. Long, and A. W. Reid. Essential closed surfaces in bounded 3-manifolds. *J. Amer. Math. Soc.*, 10(3):553–563, 1997.

[DGO] Francois Dahmani, Vincent Guirardel, and Denis Osin. Hyperbolically embedded subgroups and rotating families in groups acting on hyperbolic spaces. arXiv:1111.7048.

[DJ00] Michael W. Davis and Tadeusz Januszkiewicz. Right-angled Artin groups are commensurable with right-angled Coxeter groups. *J. Pure Appl. Algebra*, 153(3):229–235, 2000.

[GM08] Daniel Groves and Jason Fox Manning. Dehn filling in relatively hyperbolic groups. *Israel J. Math.*, 168:317–429, 2008.

[GMRS98] Rita Gitik, Mahan Mitra, Eliyahu Rips, and Michah Sageev. Widths of subgroups. *Trans. Amer. Math. Soc.*, 350(1):321–329, 1998.

[Gro87] M. Gromov. Hyperbolic groups. In *Essays in group theory*, volume 8 of *Math. Sci. Res. Inst. Publ.*, pages 75–263. Springer, New York, 1987.

[Hag08] Frédéric Haglund. Finite index subgroups of graph products. *Geom. Dedicata*, 135:167–209, 2008.

[Hak62] Wolfgang Haken. Über das Homöomorphieproblem der 3-Mannigfaltigkeiten. I. *Math. Z.*, 80:89–120, 1962.

[HP98] Frédéric Haglund and Frédéric Paulin. Simplicité de groupes d'automorphismes d'espaces à courbure négative. In *The Epstein birthday schrift*, volume 1 of *Geom. Topol. Monogr.*, pages 181–248 (electronic). Geom. Topol. Publ., Coventry, 1998.

[HWa] Frédéric Haglund and Daniel T. Wise. A combination theorem for cube complexes. Annals of Math., to appear.

[HWb] G. Christopher Hruska and Daniel T. Wise. Finiteness properties of cubulated groups. arXiv:1209.1074.

[HWc] Tim Hsu and Daniel T. Wise. Cubulating malnormal amalgams. preprint.

[HW08] Frédéric Haglund and Daniel T. Wise. Special cube complexes. *Geom. Funct. Anal.*, 17(5):1551–1620, 2008.

[KM] Jeremy Kahn and Vladimir Markovic. Immersing almost geodesic surfaces in a closed hyperbolic three manifold. arXiv:0910.5501.

[Lea] Ian Leary. A metric Kan-Thurston theorem. Journal of Topology, to appear.

[Min05] Ashot Minasyan. Some properties of subsets of hyperbolic groups. *Comm. Algebra*, 33(3):909–935, 2005.

[MMP10] Jason Fox Manning and Eduardo Martínez-Pedroza. Separation of relatively quasiconvex subgroups. *Pacific J. Math.*, 244(2):309–334, 2010.

[Osi07] Denis V. Osin. Peripheral fillings of relatively hyperbolic groups. *Invent. Math.*, 167(2):295–326, 2007.

[Sag] Michah Sageev. $CAT(0)$ cube complexes and groups. PCMI lecture notes, to appear.

[Sag95] Michah Sageev. Ends of group pairs and non-positively curved cube complexes. *Proc. London Math. Soc. (3)*, 71(3):585–617, 1995.

[Sag97] Michah Sageev. Codimension-1 subgroups and splittings of groups. *J. Algebra*, 189(2):377–389, 1997.

[Sco78] Peter Scott. Subgroups of surface groups are almost geometric. *J. London Math. Soc. (2)*, 17(3):555–565, 1978.

[Sta83] John R. Stallings. Topology of finite graphs. *Invent. Math.*, 71(3):551–565, 1983.

[Str90] Ralph Strebel. Appendix. Small cancellation groups. In *Sur les groupes hyperboliques d'après Mikhael Gromov (Bern, 1988)*, volume 83 of *Progr. Math.*, pages 227–273. Birkhäuser Boston, Boston, MA, 1990.

[SW79] Peter Scott and Terry Wall. Topological methods in group theory. In *Homological group theory (Proc. Sympos., Durham, 1977)*, volume 36 of *London Math. Soc. Lecture Note Ser.*, pages 137–203. Cambridge Univ. Press, Cambridge, 1979.

[Thu82] William P. Thurston. Three-dimensional manifolds, Kleinian groups and hyperbolic geometry. *Bull. Amer. Math. Soc. (N.S.)*, 6(3):357–381, 1982.

[Wal68] Friedhelm Waldhausen. On irreducible 3-manifolds which are sufficiently large. *Ann. of Math. (2)*, 87:56–88, 1968.

[Wis] Daniel T. Wise. From riches to raags: 3-manifolds, right-angled Artin groups, and cubical geometry. CBMS lecture notes, to appear.

[Wis07] Daniel T. Wise. Complete square complexes. *Comment. Math. Helv.*, 82(4):683–724, 2007.

Department of Mathematics, University of Utah, Salt Lake City, UT 84112

CLUSTER ALGEBRAS: AN INTRODUCTION

LAUREN K. WILLIAMS

Dedicated to Andrei Zelevinsky on the occasion of his 60th birthday

Abstract. Cluster algebras are commutative rings with a set of distinguished generators having a remarkable combinatorial structure. They were introduced by Fomin and Zelevinsky in 2000 in the context of Lie theory, but have since appeared in many other contexts, from Poisson geometry to triangulations of surfaces and Teichmüller theory. In this expository paper we give an introduction to cluster algebras, and illustrate how this framework naturally arises in Teichmüller theory. We then sketch how the theory of cluster algebras led to a proof of the Zamolodchikov periodicity conjecture in mathematical physics.

Contents

1. Introduction

Cluster algebras were conceived by Fomin and Zelevinsky [13] in the spring of 2000 as a tool for studying total positivity and dual canonical bases in Lie theory. However, the theory of cluster algebras has since taken on a life of its own, as connections and applications have been discovered to diverse areas of mathematics including quiver representations, Teichmüller theory, tropical geometry, integrable systems, and Poisson geometry.

In brief, a *cluster algebra* $\mathcal{A}$ of rank n is a subring of an *ambient field* $\mathcal{F}$ of rational functions in n variables. Unlike "most" commutative rings, a cluster algebra is not presented at the outset via a complete set of generators and relations. Instead, from the initial data of a *seed* – which includes n distinguished generators called *cluster variables* plus an *exchange matrix* – one uses an iterative procedure called *mutation* to produce the rest of the cluster variables. In particular, each new cluster variable is a rational expression in the initial cluster variables. The cluster algebra is then defined to be the subring of $\mathcal{F}$ generated by all cluster variables.

The set of cluster variables has a remarkable combinatorial structure: this set is a union of overlapping algebraically independent n-subsets of $\mathcal{F}$ called *clusters*,

1991 *Mathematics Subject Classification.* 13F60, 30F60, 82B23, 05E45.
The author is partially supported by a Sloan Fellowship and an NSF Career award.

which together have the structure of a simplicial complex called the *cluster complex*. The clusters are related to each other by birational transformations of the following kind: for every cluster $\mathbf{x}$ and every cluster variable $x \in \mathbf{x}$, there is another cluster $\mathbf{x}' = (\mathbf{x} - \{x\}) \cup \{x'\}$, with the new cluster variable x' determined by an *exchange relation* of the form

$$xx' = y^+M^+ + y^-M^-.$$

Here y^+ and y^- lie in a *coefficient semifield* $\mathbb{P}$, while M^+ and M^- are monomials in the elements of $\mathbf{x} - \{x\}$. In the most general class of cluster algebras, there are two dynamics at play in the exchange relations: that of the monomials, and that of the coefficients, both of which are encoded in the exchange matrix.

The aim of this article is threefold: to give an elementary introduction to the theory of cluster algebras; to illustrate how the framework of cluster algebras naturally arises in diverse areas of mathematics, in particular Teichmüller theory; and to illustrate how the theory of cluster algebras has been an effective tool for solving outstanding conjectures, in particular the Zamolodchikov periodicity conjecture from mathematical physics.

To this end, in Section 2 we introduce the notion of cluster algebra, beginning with the simple but somewhat restrictive definition of a cluster algebra defined by a quiver. After giving a detailed example (the *type A cluster algebra*, and its realization as the coordinate ring of the Grassmannian $Gr_{2,d}$), we give a more general definition of cluster algebra, in which both the cluster variables and coefficient variables have their own dynamics.

In Section 3 we explain how cluster algebras had appeared implicitly in Teichmüller theory, long before the introduction of cluster algebras themselves. We start by associating a cluster algebra to any *bordered surface with marked points*, following work of Fock-Goncharov [8], Gekhtman-Shapiro-Vainshtein [20], and Fomin-Shapiro-Thurston [11]. This construction specializes to the type A example from Section 2 when the surface is a disk with marked points on the boundary. We then explain how a cluster algebra from a bordered surface is related to the decorated Teichmüller space of the corresponding *cusped surface*. Finally we briefly discuss the Teichmüller space of a surface with oriented geodesic boundary and two related spaces of laminations, and how natural coordinate systems on these spaces are related to cluster algebras.

In Section 4 we discuss Zamolodchikov's *periodicity conjecture* for *Y-systems* [46]. Although this conjecture arose from Zamolodchikov's study of the thermodynamic Bethe ansatz in mathematical physics, Fomin-Zelevinsky realized that it could be reformulated in terms of the dynamics of coefficient variables in a cluster algebra [15]. We then discuss how Fomin-Zelevinsky used fundamental structural results for finite type cluster algebras to prove the periodicity conjecture for Dynkin diagrams [15], and how Keller used deep results from the categorification of cluster algebras to prove the corresponding conjecture for pairs of Dynkin diagrams [28, 29].

Acknowledgements: This paper is intended to accompany my upcoming talk at the Current Events Bulletin Session at the Joint Mathematics Meetings in San Diego, in January 2013. I would like to thank the organizers for the impetus to prepare this document. In addition, I am indebted to Bernhard Keller, Tomoki Nakanishi, and Dylan Thurston for useful conversations, and to Keller for the use of several figures. Finally, I am grateful for the hospitality of MSRI during the Fall 2012 program on cluster algebras, which provided an ideal environment for writing this paper.

2. What is a cluster algebra?

In this section we will define the notion of cluster algebra, first introduced by Fomin and Zelevinsky in [13]. For the purpose of acquainting the reader with the basic notions, in Section 2.1 we will give the simple but somewhat restrictive definition of a *cluster algebra defined by a quiver*, also called a *skew-symmetric cluster algebra of geometric type*. We will give a detailed example in Section 2.2, and then present a more general definition of cluster algebra in Section 2.3.

2.1. **Cluster algebras from quivers.**

Definition 2.1 (*Quiver*)**.** A *quiver* Q is an oriented graph given by a set of vertices Q_0, a set of arrows Q_1, and two maps $s : Q_1 \to Q_0$ and $t : Q_1 \to Q_0$ taking an arrow to its source and target, respectively.

A quiver Q is *finite* if the sets Q_0 and Q_1 are finite. A *loop* of a quiver is an arrow α whose source and target coincide. A 2-*cycle* of a quiver is a pair of distinct arrows β and γ such that $s(\beta) = t(\gamma)$ and $t(\beta) = s(\gamma)$.

Definition 2.2 (*Quiver Mutation*)**.** Let Q be a finite quiver without loops or 2-cycles. Let k be a vertex of Q. Following [13], we define the *mutated quiver* $\mu_k(Q)$ as follows: it has the same set of vertices as Q, and its set of arrows is obtained by the following procedure:

(1) for each subquiver $i \to k \to j$, add a new arrow $i \to j$;
(2) reverse all allows with source or target k;
(3) remove the arrows in a maximal set of pairwise disjoint 2-cycles.

Exercise 2.3. Mutation is an involution, that is, $\mu_k^2(Q) = Q$ for each vertex k.

Figure 1 shows two quivers which are obtained from each other by mutating at the vertex 1. We say that two quivers Q and Q' are *mutation-equivalent* if one can get from Q to Q' by a sequence of mutations.

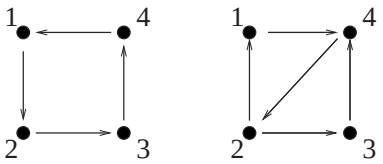


Figure 1. Two mutation-equivalent quivers.

Definition 2.4. Let Q be a finite quiver with no loops or 2-cycles and whose vertices are labeled $1, 2, \ldots, m$. Then we may encode Q by an $m \times m$ skew-symmetric *exchange matrix* $B(Q) = (b_{ij})$ where $b_{ij} = -b_{ji} = \ell$ whenever there are ℓ arrows from vertex i to vertex j. We call $B(Q)$ the *signed adjacency matrix* of the quiver.

Exercise 2.5. Check that when one encodes a quiver Q by a matrix as in Definition 2.4, the matrix $B(\mu_k(Q)) = (b'_{ij})$ is again an $m \times m$ skew-symmetric matrix, whose

entries are given by

$$b'_{ij} = \begin{cases} -b_{ij} & \text{if } i = k \text{ or } j = k; \\ b_{ij} + b_{ik}b_{kj} & \text{if } b_{ik} > 0 \text{ and } b_{kj} > 0; \\ b_{ij} - b_{ik}b_{kj} & \text{if } b_{ik} < 0 \text{ and } b_{kj} < 0; \\ b_{ij} & \text{otherwise.} \end{cases} \tag{2.1}$$

Definition 2.6 (*Labeled seeds*)**.** Choose $m \geq n$ positive integers. Let $\mathcal{F}$ be an *ambient field* of rational functions in n independent variables over $\mathbb{Q}(x_{n+1}, \ldots, x_m)$. A *labeled seed* in $\mathcal{F}$ is a pair $(\mathbf{x}, Q)$, where

- $\mathbf{x} = (x_1, \ldots, x_m)$ forms a free generating set for $\mathcal{F}$, and
- Q is a quiver on vertices $1, 2, \ldots, n, n+1, \ldots, m$, whose vertices $1, 2, \ldots, n$ are called *mutable*, and whose vertices $n+1, \ldots, m$ are called *frozen*.

We refer to $\mathbf{x}$ as the (labeled) *extended cluster* of a labeled seed $(\mathbf{x}, Q)$. The variables $\{x_1, \ldots, x_n\}$ are called *cluster variables*, and the variables $c = \{x_{n+1}, \ldots, x_m\}$ are called *frozen* or *coefficient variables*.

Definition 2.7 (*Seed mutations*)**.** Let $(\mathbf{x}, Q)$ be a labeled seed in $\mathcal{F}$, and let $k \in \{1, \ldots, n\}$. The *seed mutation* μ_k in direction k transforms $(\mathbf{x}, Q)$ into the labeled seed $\mu_k(\mathbf{x}, Q) = (\mathbf{x}', \mu_k(Q))$, where the cluster $\mathbf{x}' = (x'_1, \ldots, x'_m)$ is defined as follows: $x'_j = x_j$ for $j \neq k$, whereas $x'_k \in \mathcal{F}$ is determined by the *exchange relation*

$$x'_k \, x_k = \prod_{\substack{\alpha \in Q_1 \\ s(\alpha) = k}} x_{t(\alpha)} + \prod_{\substack{\alpha \in Q_1 \\ t(\alpha) = k}} x_{s(\alpha)} \, . \tag{2.2}$$

Remark 2.8. Note that arrows between two frozen vertices of a quiver do not affect seed mutation (they do not affect the mutated quiver or the exchange relation). For that reason, one may omit arrows between two frozen vertices. Correspondingly, when one represents a quiver by a matrix, one often omits the data corresponding to such arrows. The resulting matrix B is hence an $m \times n$ matrix rather than an $m \times m$ one.

Example 2.9. Let Q be the quiver on two vertices 1 and 2 with a single arrow from 1 to 2. Let $((x_1, x_2), Q)$ be an initial seed. Then if we perform seed mutations in directions 1, 2, 1, 2, and 1, we get the sequence of labeled seeds shown in Figure 2. Note that up to relabeling of the vertices of the quiver, the initial seed and final seed coincide.

Definition 2.10 (*Patterns*)**.** Consider the *n-regular tree* $\mathbb{T}_n$ whose edges are labeled by the numbers $1, \ldots, n$, so that the n edges emanating from each vertex receive different labels. A *cluster pattern* is an assignment of a labeled seed $\Sigma_t = (\mathbf{x}_t, Q_t)$ to every vertex $t \in \mathbb{T}_n$, such that the seeds assigned to the endpoints of any edge $t \overset{k}{\text{---}} t'$ are obtained from each other by the seed mutation in direction k. The components of $\mathbf{x}_t$ are written as $\mathbf{x}_t = (x_{1;t}, \ldots, x_{n;t})$.

Clearly, a cluster pattern is uniquely determined by an arbitrary seed.

Definition 2.11 (*Cluster algebra*)**.** Given a cluster pattern, we denote

$$\mathcal{X} = \bigcup_{t \in \mathbb{T}_n} \mathbf{x}_t = \{x_{i,t} \, : \, t \in \mathbb{T}_n \, , \; 1 \leq i \leq n\} \, , \tag{2.3}$$

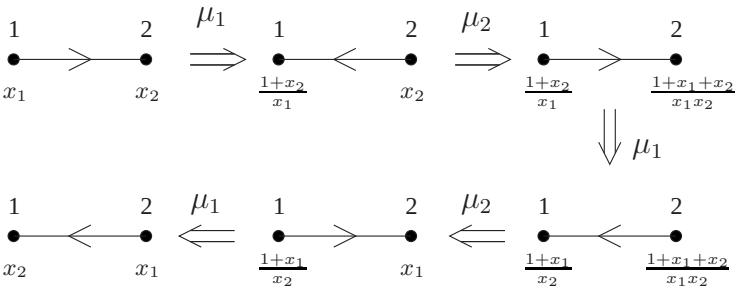


FIGURE 2. Seeds and seed mutations in type A_2.

the union of clusters of all the seeds in the pattern. The elements $x_{i,t} \in \mathcal{X}$ are called *cluster variables*. The *cluster algebra* $\mathcal{A}$ associated with a given pattern is the $\mathbb{Z}[c]$-subalgebra of the ambient field $\mathcal{F}$ generated by all cluster variables: $\mathcal{A} = \mathbb{Z}[c][\mathcal{X}]$. We denote $\mathcal{A} = \mathcal{A}(\mathbf{x}, Q)$, where $(\mathbf{x}, Q)$ is any seed in the underlying cluster pattern. In this generality, $\mathcal{A}$ is called a *cluster algebra from a quiver*, or a *skew-symmetric cluster algebra of geometric type.* We say that $\mathcal{A}$ has *rank* n because each cluster contains n cluster variables.

2.2. **Example: the type A cluster algebra.** In this section we will construct a cluster algebra using the combinatorics of triangulations of a d-gon (a polygon with d vertices). We will subsequently identify this cluster algebra with the homogeneous coordinate ring of the Grassmannian $Gr_{2,d}$ of 2-planes in a d-dimensional vector space.

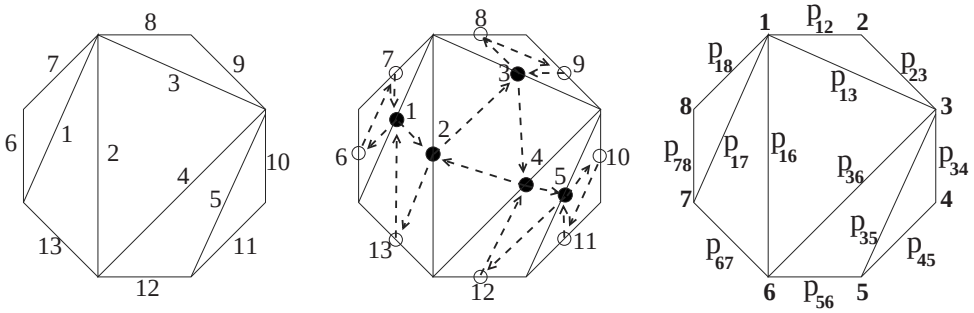


FIGURE 3. A triangulation T of an octagon, the quiver $Q(T)$, and the labeling of T by Plücker coordinates.

Definition 2.12 (*The quiver* $Q(T)$)**.** Consider a d-gon ($d \geq 3$), and choose any triangulation T. Label the $d-3$ diagonals of T with the numbers $1, 2, \ldots, d-3$, and label the d sides of the polygon by the numbers $d-2, d-1, \ldots, 2d-3$. Put a frozen vertex at the midpoint of each side of the polygon, and put a mutable vertex at the midpoint of each diagonal of the polygon. These $2d-3$ vertices are the vertices $Q_0(T)$ of $Q(T)$; label them according to the labeling of the diagonals and sides of the polygon. Now within each triangle of T, inscribe a new triangle on the vertices $Q_0(T)$, whose edges are oriented clockwise. The edges of these inscribed triangles comprise the set of arrows $Q_1(T)$ of $Q(T)$.

See the left and middle of Figure 3 for an example of a triangulation T together with the corresponding quiver $Q(T)$. The frozen vertices are indicated by hollow circles and the mutable vertices are indicated by shaded circles. The arrows of the quiver are indicated by dashed lines.

Definition 2.13 (*The cluster algebra associated to a d-gon*)**.** Let T be any triangulation of a d-gon, let $m = 2d-3$, and let $n = d-3$. Set $\mathbf{x} = (x_1, \ldots, x_m)$. Then $(\mathbf{x}, Q(T))$ is a labeled seed and it determines a cluster algebra $\mathcal{A}(T) = \mathcal{A}(\mathbf{x}, Q(T))$.

Remark 2.14. The quiver $Q(T)$ depends on the choice of triangulation T. However, we will see in Proposition 2.17 that the cluster algebra $\mathcal{A}(T)$ does not depend on T, only on the number d.

Definition 2.15 (*Flips*)**.** Consider a triangulation T which contains a diagonal t. Within T, the diagonal t is the diagonal of some quadrilateral. Then there is a new triangulation T' which is obtained by replacing the diagonal t with the other diagonal of that quadrilateral. This local move is called a *flip*.

Consider the graph whose vertex set is the set of triangulations of a d-gon, with an edge between two vertices whenever the corresponding triangulations are related by a flip. It is well-known that this "flip-graph" is connected, and moreover, is the 1-skeleton of a convex polytope called the associahedron. See Figure 4 for a picture of the flip-graph of the hexagon.

Exercise 2.16. Let T be a triangulation of a polygon, and let T' be the new triangulation obtained from T by flipping the diagonal k. Then the quiver associated to T' is the same as the quiver obtained from $Q(T)$ by mutating at k: $Q(T') = \mu_k(Q(T))$.

Proposition 2.17. Given a triangulation T of a d-gon, the cluster variables and clusters of $\mathcal{A}(T)$ are in bijection with the diagonals and triangulations of the d-gon. Moreover, if T_1 and T_2 are two triangulations of a d-gon, then the cluster algebras $\mathcal{A}(T_1)$ and $\mathcal{A}(T_2)$ are isomorphic.

Proof. This follows from Exercise 2.16 and the fact that the flip-graph is connected. □

Since the cluster algebra associated to a triangulation of a d-gon depends only on d, we will refer to this cluster algebra as $\mathcal{A}_{d-3}$. We've chosen to index this cluster algebra by $d-3$ because this cluster algebra has rank $d-3$.

2.2.1. *The homogeneous coordinate ring of the Grassmannian $Gr_{2,d}$.* The cluster algebra associated to a d-gon can be identified with the coordinate ring $\mathbb{C}[Gr_{2,d}]$ of (the affine cone over) the Grassmannian $Gr_{2,d}$ of 2-planes in a d-dimensional vector space. To see this connection, let us choose a new labeling for the cluster variables. Label the vertices of the polygon from 1 to d in order around the boundary, and label each diagonal and side of the polygon according to its endpoints, see the right of Figure 3. Using this notation, exchange relations have the following simple form.

Exercise 2.18. Let T be a triangulation of a d-gon, and consider any (triangulated) quadrilateral within T. Let $i < j < k < \ell$ be the four vertices of the quadrilateral. Then the exchange relation in $\mathcal{A}_{d-3}$ corresponding to the mutation at the diagonal of this quadrilateral is the following:

$$p_{ik}p_{j\ell} = p_{ij}p_{k\ell} + p_{i\ell}p_{jk}. \tag{2.4}$$

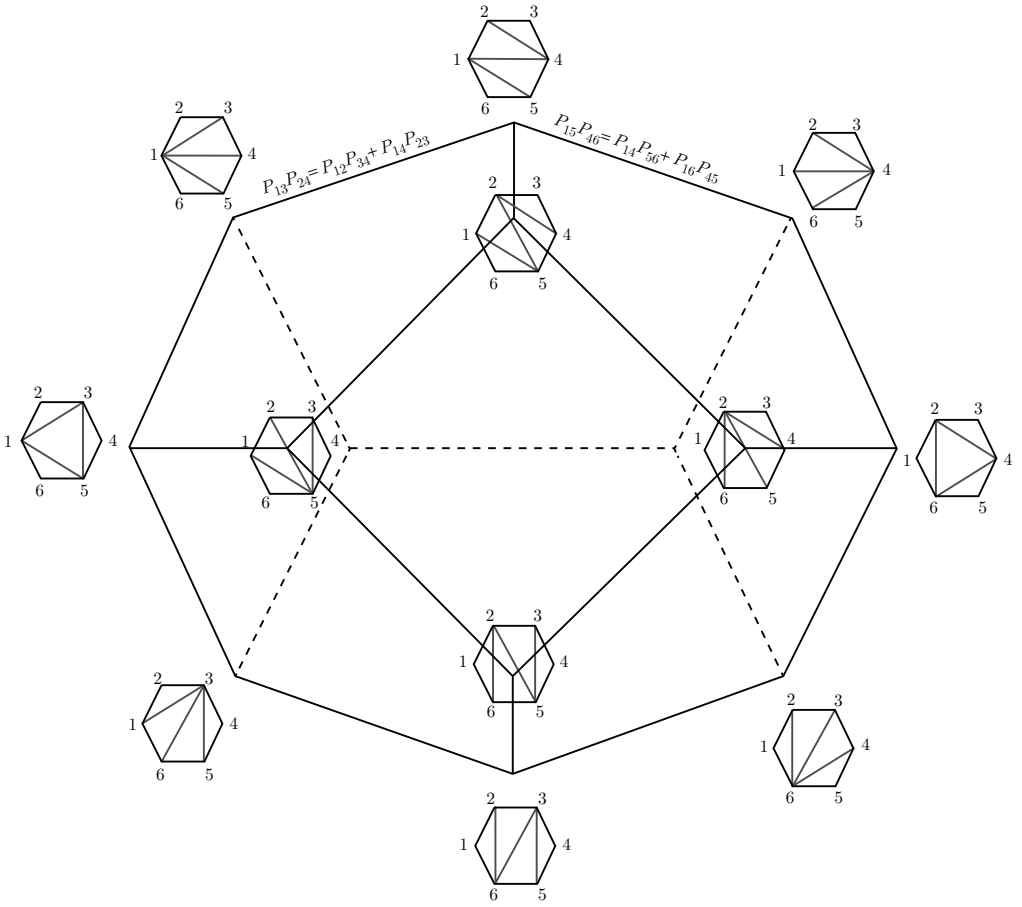


FIGURE 4. The exchange graph of the cluster algebra of type A_3, which coincides with the 1-skeleton of the associahedron.

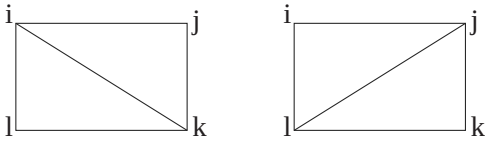


FIGURE 5. A flip in a quadrilateral and the corresponding exchange relation $p_{ik}p_{j\ell} = p_{ij}p_{k\ell} + p_{i\ell}p_{jk}$.

Recall that the coordinate ring $\mathbb{C}[Gr_{2,d}]$ is generated by *Plücker coordinates* p_{ij} for $1 \leq i < j \leq d$. The relations among the Plücker coordinates are generated by the *three-term Plücker relations*: for any $1 \leq i < j < k < \ell \leq d$, one has (2.4). Therefore it follows from Exercise 2.18 that one can identify the coordinate ring of $Gr_{2,d}$ with the cluster algebra $\mathcal{A}_{d-3}$ associated to the d-gon.

Remark 2.19. One may generalize this example of the type A cluster algebra in several ways. First, one may replace $Gr_{2,d}$ by an arbitrary Grassmannian, or partial flag variety. It turns out that the coordinate ring $\mathbb{C}[Gr_{k,d}]$ of any Grassmannian has

the structure of a cluster algebra [41], and more generally, so does the coordinate ring of any partial flag variety $SL_m(\mathbb{C})/P$ [18]. Second, one may generalize this example by replacing the d-gon – topologically a disk with d marked points on the boundary – by an orientable Riemann surface S (with or without boundary) together with some marked points M on S. One may still consider triangulations of (S, M), and use the combinatorics of these triangulations to define a cluster algebra. This cluster algebra is closely related to the *decorated Teichmulller space* associated to (S, M). We will take up this theme in Section 3.

2.3. **Cluster algebras revisited.** We now give a more general definition of cluster algebra, following [16], in which the coefficient variables have their own dynamics. In Section 4 we will see that the dynamics of coefficient variables is closely related to Zamolodchikov's Y-systems.

To define a cluster algebra $\mathcal{A}$ we first choose a *semifield* $(\mathbb{P}, \oplus, \cdot)$, i.e., an abelian multiplicative group endowed with a binary operation of *(auxiliary) addition* $\oplus$ which is commutative, associative, and distributive with respect to the multiplication in $\mathbb{P}$. The group ring $\mathbb{Z}\mathbb{P}$ will be used as a *ground ring* for $\mathcal{A}$. One important choice for $\mathbb{P}$ is the tropical semifield (see Definition 2.20); in this case we say that the corresponding cluster algebra is of *geometric type*.

Definition 2.20 (*Tropical semifield*)**.** Let $\mathrm{Trop}(u_1, \ldots, u_m)$ be an abelian group, which is freely generated by the u_j's, and written multiplicatively. We define $\oplus$ in $\mathrm{Trop}(u_1, \ldots, u_m)$ by

$$\prod_j u_j^{a_j} \oplus \prod_j u_j^{b_j} = \prod_j u_j^{\min(a_j, b_j)}, \tag{2.5}$$

and call $(\mathrm{Trop}(u_1, \ldots, u_m), \oplus, \cdot)$ a *tropical semifield*. Note that the group ring of $\mathrm{Trop}(u_1, \ldots, u_m)$ is the ring of Laurent polynomials in the variables u_j .

As an *ambient field* for $\mathcal{A}$, we take a field $\mathcal{F}$ isomorphic to the field of rational functions in n independent variables (here n is the *rank* of $\mathcal{A}$), with coefficients in $\mathbb{Q}\mathbb{P}$. Note that the definition of $\mathcal{F}$ does not involve the auxiliary addition in $\mathbb{P}$.

Definition 2.21 (*Labeled seeds*)**.** A *labeled seed* in $\mathcal{F}$ is a triple $(\mathbf{x}, \mathbf{y}, B)$, where

- $\mathbf{x} = (x_1, \ldots, x_n)$ is an n-tuple from $\mathcal{F}$ forming a *free generating set* over $\mathbb{Q}\mathbb{P}$, that is, $x_1, \ldots, x_n$ are algebraically independent over $\mathbb{Q}\mathbb{P}$, and $\mathcal{F} = \mathbb{Q}\mathbb{P}(x_1, \ldots, x_n)$.
- $\mathbf{y} = (y_1, \ldots, y_n)$ is an n-tuple from $\mathbb{P}$, and
- $B = (b_{ij})$ is an $n \times n$ integer matrix which is *skew-symmetrizable*, that is, there exist positive integers $d_1, \ldots, d_n$ such that $d_i b_{ij} = -d_j b_{ij}$.

We refer to $\mathbf{x}$ as the (labeled) *cluster* of a labeled seed $(\mathbf{x}, \mathbf{y}, B)$, to the tuple $\mathbf{y}$ as the *coefficient tuple*, and to the matrix B as the *exchange matrix*.

We obtain (*unlabeled*) *seeds* from labeled seeds by identifying labeled seeds that differ from each other by simultaneous permutations of the components in $\mathbf{x}$ and $\mathbf{y}$, and of the rows and columns of B.

In what follows, we use the notation $[x]_+ = \max(x, 0)$.

Definition 2.22 (*Seed mutations*)**.** Let $(\mathbf{x}, \mathbf{y}, B)$ be a labeled seed in $\mathcal{F}$, and let $k \in \{1, \ldots, n\}$. The *seed mutation* μ_k in direction k transforms $(\mathbf{x}, \mathbf{y}, B)$ into the labeled seed $\mu_k(\mathbf{x}, \mathbf{y}, B) = (\mathbf{x}', \mathbf{y}', B')$ defined as follows:

- The entries of $B' = (b'_{ij})$ are given by

$$b'_{ij} = \begin{cases} -b_{ij} & \text{if } i = k \text{ or } j = k; \\ b_{ij} + b_{ik}b_{kj} & \text{if } b_{ik} > 0 \text{ and } b_{kj} > 0; \\ b_{ij} - b_{ik}b_{kj} & \text{if } b_{ik} < 0 \text{ and } b_{kj} < 0; \\ b_{ij} & \text{otherwise.} \end{cases} \tag{2.6}$$

- The coefficient tuple $\mathbf{y}' = (y'_1, \ldots, y'_n)$ is given by

$$y'_j = \begin{cases} y_k^{-1} & \text{if } j = k; \\ y_j y_k^{[b_{kj}]_+} (y_k \oplus 1)^{-b_{kj}} & \text{if } j \neq k. \end{cases} \tag{2.7}$$

- The cluster $\mathbf{x}' = (x'_1, \ldots, x'_n)$ is given by $x'_j = x_j$ for $j \neq k$, whereas $x'_k \in \mathcal{F}$ is determined by the *exchange relation*

$$x'_k = \frac{y_k \prod x_i^{[b_{ik}]_+} + \prod x_i^{[-b_{ik}]_+}}{(y_k \oplus 1) x_k} \,. \tag{2.8}$$

We say that two exchange matrices B and B' are *mutation-equivalent* if one can get from B to B' by a sequence of mutations.

If we forget the cluster variables, then we refer to the resulting seeds and operation of mutation as *Y-seeds* and *Y-seed mutation*.

Definition 2.23 (*Y-seed mutations*)**.** Let $(\mathbf{y}, B)$ be a labeled seed in which we have omitted the cluster $\mathbf{x}$, and let $k \in \{1, \ldots, n\}$. The *Y-seed mutation* μ_k in direction k transforms $(\mathbf{y}, B)$ into the *labeled Y-seed* $\mu_k(\mathbf{y}, B) = (\mathbf{y}', B')$, where $\mathbf{y}'$ and B' are as in Definition 2.22.

Definition 2.24 (*Patterns*)**.** Consider the *n-regular tree* $\mathbb{T}_n$ whose edges are labeled by the numbers $1, \ldots, n$, so that the n edges emanating from each vertex receive different labels. A *cluster pattern* is an assignment of a labeled seed $\Sigma_t = (\mathbf{x}_t, \mathbf{y}_t, B_t)$ to every vertex $t \in \mathbb{T}_n$, such that the seeds assigned to the endpoints of any edge $t \overset{k}{\text{---}} t'$ are obtained from each other by the seed mutation in direction k. The components of Σ_t are written as:

$$\mathbf{x}_t = (x_{1;t}, \ldots, x_{n;t}), \quad \mathbf{y}_t = (y_{1;t}, \ldots, y_{n;t}), \quad B_t = (b^t_{ij}). \tag{2.9}$$

One may view a cluster pattern as a *discrete dynamical system* on an n-regular tree. If one ignores the coefficients (i.e. set each coefficient tuple equal to $(1, \ldots, 1)$), then we refer to the evolution of the cluster variables as *cluster dynamics*. On the other hand, ignoring the cluster variables, we refer to the evolution of the coefficient variables as *coefficient dynamics*.

Definition 2.25 (*Cluster algebra*)**.** Given a cluster pattern, we denote

$$\mathcal{X} = \bigcup_{t \in \mathbb{T}_n} \mathbf{x}_t = \{x_{i,t} : t \in \mathbb{T}_n, \ 1 \leq i \leq n\}, \tag{2.10}$$

the union of clusters of all the seeds in the pattern. The elements $x_{i,t} \in \mathcal{X}$ are called *cluster variables*. The *cluster algebra* $\mathcal{A}$ associated with a given pattern is the $\mathbb{ZP}$-subalgebra of the ambient field $\mathcal{F}$ generated by all cluster variables: $\mathcal{A} = \mathbb{ZP}[\mathcal{X}]$. We denote $\mathcal{A} = \mathcal{A}(\mathbf{x}, \mathbf{y}, B)$, where $(\mathbf{x}, \mathbf{y}, B)$ is any seed in the underlying cluster pattern.

We now explain the relationship between Definition 2.11 – the definition of cluster algebra we gave in Section 2.1 – and Definition 2.25. There are two apparent differences between the definitions. First, in Definition 2.11, the dynamics of mutation was encoded by a quiver, while in Definition 2.25, the dynamics of mutation was encoded by a skew-symmetrizable matrix B. Clearly if B is not only skew-symmetrizable but also skew-symmetric, then B can be regarded as the signed adjacency matrix of a quiver. In that case mutation of B reduces to the mutation of the corresponding quiver, and the two notions of exchange relation coincide. Second, in Definition 2.11, the coefficient variables are "frozen" and do not mutate, while in Definition 2.25, the coefficient variables y_i have a dynamics of their own. It turns out that if in Definition 2.25 the semifield $\mathbb{P}$ is the tropical semifield (and B is skew-symmetric), then Definitions 2.11 and 2.25 are equivalent. This is a consequence of the following exercise.

Exercise 2.26. Let $\mathbb{P} = \mathrm{Trop}(x_{n+1}, \ldots, x_m)$ be the tropical semifield with generators $x_{n+1}, \ldots, x_m$, and consider a cluster algebra as defined in Definition 2.25. Since the coefficients $y_{j;t}$ at the seed $\Sigma_t = (\mathbf{x}_t, \mathbf{y}_t, B_t)$ are Laurent monomials in $x_{n+1}, \ldots, x_m$, we may define the integers b'^t_{ij} for $j \in \{1, \ldots, n\}$ and $n < i \leq m$ by

$$y_{j;t} = \prod_{i=n+1}^{m} x_i^{b'^t_{ij}}.$$

This gives a natural way of including the exchange matrix B_t as the principal $n \times n$ submatrix into a larger $m \times n$ matrix $\widetilde{B}_t = (b'^t_{ij})$ where $1 \leq i \leq m$ and $1 \leq j \leq n$, whose matrix elements b'^t_{ij} with $i > n$ encode the coefficients $y_j = y_{j;t}$.

Check that with the above conventions, the exchange relation (2.8) reduces to the exchange relation (2.2), and that the Y-seed mutation rule (2.7) implies that the extended exchange matrix $\widetilde{B}_t$ mutates according to (2.1).

2.4. **Structural properties of cluster algebras.** In this section we will explain various structural properties of cluster algebras. Throughout this section $\mathcal{A}$ will be an arbitrary cluster algebra as defined in Section 2.3.

From the definitions, it is clear that any cluster variable can be expressed as a rational function in the variables of an arbitrary cluster. However, the remarkable *Laurent phenomenon*, proved in [13, Theorem 3.1], asserts that each such rational function is actually a Laurent polynomial.

Theorem 2.27 (*Laurent Phenomenon*)**.** *The cluster algebra $\mathcal{A}$ associated with a seed $\Sigma = (\mathbf{x}, \mathbf{y}, B)$ is contained in the Laurent polynomial ring $\mathbb{ZP}[\mathbf{x}^{\pm 1}]$, i.e. every element of $\mathcal{A}$ is a Laurent polynomial over $\mathbb{ZP}$ in the cluster variables from $\mathbf{x} = (x_1, \ldots, x_n)$.*

Let $\mathcal{A}$ be a cluster algebra, Σ be a seed, and x be a cluster variable of $\mathcal{A}$. Let $[x]^{\mathcal{A}}_{\Sigma}$ denote the Laurent polynomial which expresses x in terms of the cluster variables from Σ; it is called the *cluster expansion* of x in terms of Σ. The longstanding *Positivity Conjecture* [13] says that the coefficients that appear in such Laurent polynomials are positive.

Conjecture 2.28 (*Positivity Conjecture*)**.** For any cluster algebra $\mathcal{A}$, any seed Σ, and any cluster variable x, the Laurent polynomial $[x]^{\mathcal{A}}_{\Sigma}$ has coefficients which are nonnegative integer linear combinations of elements in $\mathbb{P}$.

While Conjecture 2.28 is open in general, it has been proved in some special cases, see for example [2], [35], [36], [5].

One of Fomin-Zelevinsky's motivations for introducing cluster algebras was the desire to understand the canonical bases of quantum groups due to Lusztig and Kashiwara [32, 27]. See [19] for some recent results connecting cluster algebras and canonical bases. Some of the conjectures below, including Conjectures 2.30 and 2.32, are motivated in part by the conjectural connection between cluster algebras and canonical bases.

Definition 2.29 (*Cluster monomial*)**.** A *cluster monomial* in a cluster algebra $\mathcal{A}$ is a monomial in cluster variables, all of which belong to the same cluster.

Conjecture 2.30. Cluster monomials are linearly independent.

The best result to date towards Conjecture 2.30 is the following.

Theorem 2.31. [3] *In a cluster algebra defined by a quiver, the cluster monomials are linearly independent.*

The following conjecture implies both Conjecture 2.28 and Conjecture 2.30.

Conjecture 2.32 (*Strong Positivity Conjecture*)**.** Any cluster algebra has an additive basis $\mathbb{B}$ which

- includes the cluster monomials, and
- has nonnegative structure constants, that is, when one writes the product of any two elements in $\mathbb{B}$ in terms of $\mathbb{B}$, the coefficients are positive.

One of the most striking results about cluster algebras is that the classification of the *finite type* cluster algebras is parallel to the Cartan-Killing classification of complex simple Lie algebras. In particular, finite type cluster algebras are classified by Dynkin diagrams.

Definition 2.33 (*Finite type*)**.** We say that a cluster algebra is of *finite type* if it has finitely many seeds.

It turns out that the classification of finite type cluster algebras is parallel to the Cartan-Killing classification of complex simple Lie algebras [14]. More specifically, define the *diagram* $\Gamma(B)$ associated to an $n \times n$ exchange matrix B to be a weighted directed graph on nodes $v_1, \ldots, v_n$, with v_i directed towards v_j if and only if $b_{ij} > 0$. In that case, we label this edge by $|b_{ij}b_{ji}|$.

Theorem 2.34. [14, Theorem 1.8] *The cluster algebra $\mathcal{A}$ is of finite type if and only if it has a seed $(\mathbf{x}, \mathbf{y}, B)$ such that $\Gamma(B)$ is an orientation of a finite type Dynkin diagram.*

If the conditions of Theorem 2.34 hold, we say that $\mathcal{A}$ is of *finite type*. And in that case if $\Gamma(B)$ is an orientation of a Dynkin diagram of type X (here X belongs to one of the infinite series A_n, B_n, C_n, D_n, or to one of the exceptional types E_6, E_7, E_8, F_4, G_2), we say that the cluster algebra $\mathcal{A}$ is of type X.

We define the *exchange graph* of a cluster algebra to be the graph whose vertices are the (unlabeled) seeds, and whose edges connect pairs of seeds which are connected by a mutation. When a cluster algebra is of finite type, its exchange graph has a remarkable combinatorial structure.

Theorem 2.35. [4] *Let $\mathcal{A}$ be a cluster algebra of finite type. Then its exchange graph is the* 1*-skeleton of a convex polytope called a* generalized associahedron.

The polytopes in Theorem 2.35 are called generalized associahedra because when $\mathcal{A}$ is a cluster algebra of type A, its exchange graph is the 1-skeleton of the usual associahedron, see Figure 4.

3. Cluster algebras in Teichmüller theory

In this section we will explain how cluster algebras had already appeared implicitly in Teichmüller theory, before the introduction of cluster algebras themselves. In particular, we will associate a cluster algebra to any *bordered surface with marked points*, following work of Fock-Goncharov [8], Gekhtman-Shapiro-Vainshtein [20], and Fomin-Shapiro-Thurston [11]. This construction provides a natural generalization of the type A cluster algebra from Section 2.2, and realizes the *lambda lengths* (also called *Penner coordinates*) on the decorated Teichmüller space associated to a cusped surface, which Penner had defined in 1987 [37]. We will also briefly discuss the Teichmüller space of a surface with oriented geodesic boundary and related spaces of laminations, and how these spaces are related to cluster theory. For more details on the Teichmüller and lamination spaces, see [9].

3.1. Surfaces, arcs, and triangulations.

Definition 3.1 (*Bordered surface with marked points*)**.** Let S be a connected oriented 2-dimensional Riemann surface with (possibly empty) boundary. Fix a nonempty set M of *marked points* in the closure of S with at least one marked point on each boundary component. The pair (S, M) is called a *bordered surface with marked points*. Marked points in the interior of S are called *punctures*.

For technical reasons we require that (S, M) is not a sphere with one, two or three punctures; a monogon with zero or one puncture; or a bigon or triangle without punctures.

Let g denote the genus of S, p the number of punctures, b the number of boundary components, and c the number of marked points on the boundary.

Definition 3.2 (*Arcs and boundary segments*)**.** An *arc* γ in (S, M) is a curve in S, considered up to isotopy, such that: the endpoints of γ are in M; γ does not cross itself, except that its endpoints may coincide; except for the endpoints, γ is disjoint from M and from the boundary of S; and γ does not cut out an unpunctured monogon or an unpunctured bigon.

A *boundary segment* is a curve that connects two marked points and lies entirely on the boundary of S without passing through a third marked point.

Let $A(S, M)$ and $B(S, M)$ denote the sets of arcs and boundary segments in (S, M). Note that $A(S, M)$ and $B(S, M)$ are disjoint.

Definition 3.3 (*Compatibility of arcs, and triangulations*)**.** We say that arcs γ and γ' are *compatible* if there exist curves α and α' isotopic to γ and γ', such that α and α' do not cross. A *triangulation* is a maximal collection of pairwise compatible arcs (together with all boundary segments). The arcs of a triangulation cut the surface into *triangles*.

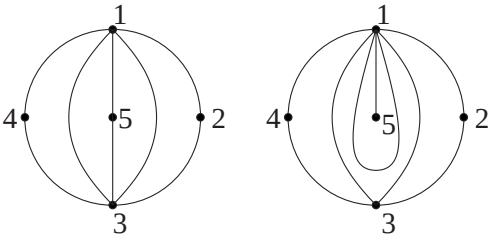


FIGURE 6. Two triangulations of a once-punctured polygon. The triangulation at the right contains a self-folded triangle.

There are two types of triangles: triangles that have three distinct sides, and *self-folded triangles* that have only two. Note that a self-folded triangle consists of a loop, together with an arc to an enclosed puncture, called a *radius*, see Figure 6.

Definition 3.4 (*Flips*)**.** A *flip* of a triangulation T replaces a single arc $\gamma \in T$ by a (unique) arc $\gamma' \neq \gamma$ that, together with the remaining arcs in T, forms a new triangulation.

In Figure 6, the triangulation at the right is obtained from the triangulation at the left by flipping the arc between the marked points 3 and 5. However, a radius inside a self-folded triangle in T cannot be flipped (see e.g. the arc between 1 and 5 at the right).

Proposition 3.5. [22, 23, 34] Any two triangulations of a bordered surface are related by a sequence of flips.

3.2. **Decorated Teichmüller space.** In this section we assume that the reader is familiar with some basics of hyperbolic geometry.

Definition 3.6 (*Teichmüller space*)**.** Let (S, M) be a bordered surface with marked points. The (cusped) *Teichmüller space* $\mathcal{T}(S, M)$ consists of all complete finite-area hyperbolic metrics with constant curvature -1 on $S \setminus M$, with geodesic boundary at $\partial S \setminus M$, considered up to $\mathrm{Diff}_0(S, M)$, diffeomorphisms of S fixing M that are homotopic to the identity. (Thus there is a cusp at each point of M: points at M "go off to infinity," while the area remains bounded.)

For a given hyperbolic metric in $\mathcal{T}(S, M)$, each arc can be represented by a unique geodesic. Since there are cusps at the marked points, such a geodesic segment is of infinite length. So if we want to measure the "length" of a geodesic arc between two marked points, we need to renormalize.

To do so, around each cusp p we choose a *horocycle*, which may be viewed as the set of points at an equal distance from p. Although the cusp is infinitely far away from any point in the surface, there is still a well-defined way to compare the distance to p from two different points in the surface. A horocycle can also be characterized as a curve perpendicular to every geodesic to p. See Figure 7 for a depiction of some points and horocycles, drawn in the hyperbolic plane.

The notion of horocycle leads to the following definition.

Definition 3.7 (*Decorated Teichmüller space*)**.** A point in a decorated Teichmüller space $\widetilde{\mathcal{T}}(S, M)$ is a hyperbolic metric as above together with a collection of horocyles h_p, one around each cusp corresponding to a marked point $p \in M$.

One may parameterize decorated Teichmüller space using *lambda lengths* or *Penner coordinates*, as introduced and developed by Penner [37, 38].

Definition 3.8 (*Lambda lengths*)**.** [38] Fix $\sigma \in \widetilde{\mathcal{T}}(S, M)$. Let γ be an arc or a boundary segment. Let γ_σ denote the geodesic representative of γ (relative to σ). Let $\ell(\gamma) = \ell_\sigma(\gamma)$ be the signed distance along γ_σ between the horocycles at either end of γ (positive if the two horocycles do not intersect, and negative if they do). The *lambda length* $\lambda(\gamma) = \lambda_\sigma(\gamma)$ of γ is defined by

$$\lambda(\gamma) = \exp(\ell(\gamma)/2). \tag{3.1}$$

Given $\gamma \in A(S, M) \cup B(S, M)$, one may view the lambda length

$$\lambda(\gamma) : \sigma \mapsto \lambda_\sigma(\gamma)$$

as a function on the decorated Teichmüller space $\widetilde{\mathcal{T}}(S, M)$. Let n denote the number of arcs in a triangulation of (S, M); recall that c denotes the number of marked points on the boundary of S. Penner showed that if one fixes a triangulation T, then the lambda lengths of the arcs of T and the boundary segments can be used to parameterize $\widetilde{\mathcal{T}}(S, M)$:

Theorem 3.9. *For any triangulation T of (S, M), the map*

$$\prod_{\gamma \in T \cup B(S,M)} \lambda(\gamma) : \widetilde{\mathcal{T}}(S, M) \to \mathbb{R}_{>0}^{n+c}$$

is a homeomorphism.

Note that the first versions of Theorem 3.9 were due to Penner [37, Theorem 3.1], [38, Theorem 5.10], but the formulation above is from [12, Theorem 7.4].

The following "Ptolemy relation" is an indication that lambda lengths on decorated Teichmüller space are part of a related cluster algebra.

Proposition 3.10. [37, Proposition 2.6(a)] Let $\alpha, \beta, \gamma, \delta \in A(S, M) \cup B(S, M)$ be arcs or boundary segments (not necessarily distinct) that cut out a quadrilateral in S; we assume that the sides of the quadrilateral, listed in cyclic order, are $\alpha, \beta, \gamma, \delta$. Let η and θ be the two diagonals of this quadrilateral. Then the corresponding lambda lengths satisfy the Ptolemy relation

$$\lambda(\eta)\lambda(\theta) = \lambda(\alpha)\lambda(\gamma) + \lambda(\beta)\lambda(\delta).$$

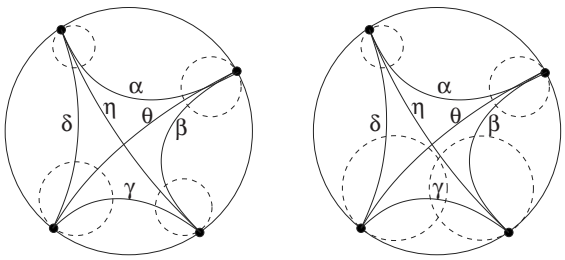


FIGURE 7. Four points and corresponding horocycles, together with the arcs forming the geodesics between them, drawn in the hyperbolic plane. At the left, all lengths $\ell(\alpha), \ldots, \ell(\theta)$ are positive; but at the right, $\ell(\gamma)$ is negative.

3.3. **The cluster algebra associated to a surface.** To make precise the connection between decorated Teichmüller space and cluster algebras, let us fix a triangulation T of (S, M), and explain how to associate an exchange matrix B_T to T [8, 10, 20, 11]. For simplicity we assume that T has no self-folded triangles. It is not hard to see that all of the bordered surfaces we are considering admit a triangulation without self-folded triangles.

Definition 3.11 (*Exchange matrices associated to a triangulation*)**.** Let T be a triangulation of (S, M). Let $\tau_1, \tau_2, \ldots, \tau_n$ be the n arcs of T, and $\tau_{n+1}, \ldots, \tau_{n+c}$ be the c boundary segments of (S, M). We define

$$\begin{aligned} b_{ij} =& \#\{\text{triangles with sides } \tau_i \text{ and } \tau_j, \text{ with } \tau_j \text{ following } \tau_i \text{ in clockwise order}\} - \\ & \#\{\text{triangles with sides } \tau_i \text{ and } \tau_j, \text{ with } \tau_j \text{ following } \tau_i \text{ in counterclockwise order}\}. \end{aligned}$$

Then we define the *exchange matrix* $B_T = (b_{ij})_{1 \le i \le n, 1 \le j \le n}$ and the *extended exchange matrix* $\widetilde{B}_T = (b_{ij})_{1 \le i \le n+c, 1 \le j \le n}$.

In Figure 7 there is a triangle with sides α, β, and η, and our convention is that β follows α in clockwise order. Note that in order to speak about clockwise order, one must be working with an oriented surface. That is why Definition 3.1 requires S to be oriented.

We leave the following as an exercise for the reader; alternatively, see [11].

Exercise 3.12. Extend Definition 3.11 to the case that T has self-folded triangles, so that the exchange matrices transform compatibly with mutation.

Remark 3.13. Note that each entry b_{ij} of the exchange matrix (or extended exchanged matrix) is either $0, \pm 1$, or ± 2, since every arc τ is in at most two triangles.

Exercise 3.14. Note that B_T is skew-symmetric, and hence can be viewed as the signed adjacency matrix associated to a quiver. Verify that this quiver generalizes the quiver $Q(T)$ from Definition 2.12 associated to a triangulation of a polygon.

The following result follows easily from the definition of B_T, and the fact that any two triangulations of (S, M) can be connected by a sequence of flips.

Theorem 3.15. *Let (S, M) be a bordered surface and let $T = (\tau_1, \ldots, \tau_n)$ be a triangulation of (S, M). Let $\mathbf{x}_T = (x_{\tau_1}, \ldots, x_{\tau_n})$, and let $\mathcal{A} = \mathcal{A}(\mathbf{x}_T, B_T)$ be the corresponding cluster algebra. Then we have the following:*

- *Each arc $\gamma \in A(S, M)$ gives rise to a cluster variable x_γ;*
- *Each triangulation T of (S, M) gives rise to a seed $\Sigma_T = (x_T, B_T)$ of $\mathcal{A}$;*
- *If T' is obtained from T by flipping at τ_k, then $B_{T'} = \mu_k(B_T)$.*

It follows that the cluster algebra $\mathcal{A}$ does not depend on the triangulation T, but only on (S, M). Therefore we refer to this cluster algebra as $\mathcal{A}(S, M)$.

Remark 3.16. Theorem 3.15 gives an inclusion of arcs into the set of cluster variables of $\mathcal{A}(S, M)$. This inclusion is a bijection if and only if (S, M) has no punctures. In [11], Fomin-Shapiro-Thurston introduced *tagged arcs* and *tagged triangulations*, which generalize arcs and triangulations, and are in bijection with cluster variables and clusters of $\mathcal{A}(S, M)$, see [11, Theorem 7.11]. To each tagged triangulation one may associate an exchange matrix, which as before has all entries equal to 0, ± 1, or ± 2.

Combining Theorem 3.15 with Theorem 3.9 and Proposition 3.10, we may identify the cluster variable x_γ with the corresponding lambda length $\lambda(\gamma)$, and therefore view such elements of the cluster algebra $\mathcal{A}$ as functions on $\widetilde{\mathcal{T}}(S, M)$. In particular, when one performs a flip in a triangulation, the lambda lengths associated to the arcs *transform according to cluster dynamics*.

It is natural to consider whether there is a nice system of coordinates on Teichmüller space $\mathcal{T}(S, M)$ itself (as opposed to its decorated version.) Indeed, if one fixes a point of $\mathcal{T}(S, M)$ and a triangulation T of (S, M), one may represent T by geodesics and lift it to an *ideal triangulation* of the upper half plane. Note that every arc of the triangulation is the diagonal of a unique quadrilateral. The four points of this quadrilateral have a unique invariant under the action of $PSL_2(\mathbb{R})$, the *cross-ratio*. One may compute the cross-ratio by sending three of the four points to 0, -1, and ∞. Then the position x of the fourth point is the cross-ratio. The collection of cross-ratios associated to the arcs in T comprise a system of coordinates on $\mathcal{T}(S, M)$. And when one performs a flip in the triangulation, the coordinates *transform according to coefficient dynamics*, see [9, Section 4.1].

3.4. **Spaces of laminations and their coordinates.** Several compactifications of Teichmüller space have been introduced. The most widely used compactification is due to W. Thurston [43, 44]; the points at infinity of this compactification correspond to *projective measured laminations.*

Informally, a measured lamination on (S, M) is a finite collection of non-self-intersecting and pairwise non-intersecting weighted curves in $S \setminus M$, considered up to homotopy, and modulo a certain equivalence relation. It is not hard to see why such a lamination L might correspond to a limit point of Teichmüller space $\mathcal{T}(S, M)$: given L, one may construct a family of metrics on the surface "converging to L," by cutting at each curve in L and inserting a "long neck." As the necks get longer and longer, the length of an arbitrary curve in the corresponding metric becomes dominated by the number of times that curve crosses L. Therefore L represents the limit of this family of metrics.

Interestingly, two versions of the space of laminations on (S, M) – the space of *rational bounded measured laminations*, and the space of *rational unbounded measured laminations* – are closely connected to cluster theory. In both cases, one may fix a triangulation T and then use appropriate coordinates to get a parameterization of the space. The appropriate coordinates for the space of rational bounded measured laminations are *intersection numbers*. Let T' be a triangulation obtained from T by performing a flip. It turns out that when one replaces T by T', the rule for how intersection numbers change is given by a *tropical* version of *cluster dynamics*. On the other hand, the appropriate coordinates for the space of rational unbounded measured laminations are *shear coordinates*. When one replaces T by T', the rule for how shear coordinates change is given by a *tropical* version of *coefficient dynamics*. See [9] for more details.

In Table 1, we summarize the properties of the two versions of Teichmüller space, and the two versions of the space of laminations, together with their coordinates.

3.5. **Applications of Teichmüller theory to cluster theory.** The connection between Teichmüller theory and cluster algebras has useful applications to cluster algebras, some of which we discuss below.

Space	Coordinates	Coordinate transformations
Decorated Teichmüller space	Lambda lengths	Cluster dynamics
Teichmüller space	Cross-ratios	Coefficient dynamics
Bounded measured laminations	Intersection numbers	Tropical cluster dynamics
Unbounded measured laminations	Shear coordinates	Tropical coefficient dynamics

TABLE 1. Teichmüller and lamination spaces.

As mentioned in Remark 3.16, the combinatorics of (tagged) arcs and (tagged) triangulations gives a concrete way to index cluster variables and clusters in a cluster algebra from a surface. Additionally, the combinatorics of unbounded measured laminations gives a concrete way to encode the coefficient variables for a cluster algebra, whose coefficient system is of geometric type [12]. Recall from Section 2.1 or Exercise 2.26 that the coefficient system is determined by the bottom $m-n$ rows of the initial extended exchange matrix $\widetilde{B}$. However, after one has mutated away from the initial cluster, one would like an explicit way to read off the resulting coefficient variables (short of performing the corresponding sequence of mutations). In [12], the authors demonstrated that one may encode the initial extended exchange matrix by a triangulation *together with a lamination*, and that one may compute the coefficient variables (even after mutating away from the initial cluster) by using *shear coordinates*.

Note that for a general cluster algebra, there is no explicit way to index cluster variables or clusters, or to encode the coefficients. A cluster variable is simply a rational function of the initial cluster variables that is obtained after some arbitrary and arbitrarily long sequence of mutations. Having a concrete index set for the cluster variables and clusters, as in [11, Theorem 7.11], is a powerful tool. Indeed, this was a key ingredient in [35], which proved the Positivity Conjecture for all cluster algebras from surfaces.

The connection between Teichmüller theory and cluster algebras from surfaces has also led to important structural results for such cluster algebras. We say that a cluster algebra has *polynomial growth* if the number of distinct seeds which can be obtained from a fixed initial seed by at most n mutations is bounded from above by a polynomial function of n. A cluster algebra has *exponential growth* if the number of such seeds is bounded from below by an exponentially growing function of n. In [11], Fomin-Shapiro-Thurston classified the cluster algebras from surfaces according to their growth: there are six infinite families which have polynomial growth, and all others have exponential growth.

Another structural result relates to the classification of mutation-finite cluster algebras. We say that a matrix B (and the corresponding cluster algebra) is *mutation-finite* (or is of *finite mutation type*) if its mutation equivalence class is finite, i.e. only finitely many matrices can be obtained from B by repeated matrix mutations. Felikson-Shapiro-Tumarkin gave a classification of all skew-symmetric mutation-finite cluster algebras in [7]. They showed that these cluster algebras are the union of the following classes of cluster algebras:

- Rank 2 cluster algebras;
- Cluster algebras from surfaces;
- One of 11 exceptional types.

Note that the above classification may be extended to all mutation-finite cluster algebras (not necessarily skew-symmetric), using *cluster algebras from orbifolds* [6].

Exercise 3.17. Show that any cluster algebra from a surface is mutation-finite. *Hint: use Remark 3.16.*

4. Cluster algebras and the Zamolodchikov periodicity conjecture

The thermodynamic Bethe ansatz is a tool for understanding certain conformal field theories. In a paper from 1991 [46], the physicist Al. B. Zamolodchikov studied the thermodynamic Bethe ansatz equations for ADE-related diagonal scattering theories. He showed that if one has a solution to these equations, it should also be a solution of a set of functional relations called a *Y-system*. Furthermore, he remarked that based on numerical tests, the solutions to the Y-system appeared to be periodic. This phenomenon is called the *Zamolodchikov periodicity conjecture*, and has important consequences for the corresponding field theory. Although this conjecture arose in mathematical physics, we will see that it can be reformulated and proved using the framework of cluster algebras.

Note that Zamolodchikov initially stated his conjecture for the Y-system of a simply-laced Dynkin diagram. The notion of Y-system and the periodicity conjecture were subsequently generalized by Ravanini-Valleriani-Tateo [40], Kuniba-Nakanishi [30], Kuniba-Nakanishi-Suzuki [31], Fomin-Zelevinsky [15], etc. We will first present Zamolodchikov's periodicity conjecture for Dynkin diagrams Δ (not necessarily simply-laced), and then present its extension to pairs (Δ, Δ') of Dynkin diagrams. Note that the latter conjecture reduces to the former in the case that $\Delta' = A_1$. The conjecture was proved for (A_n, A_1) by Frenkel-Szenes [17] and Gliozzi-Tateo [21]; for (Δ, A_1) (where Δ is an arbitrary Dynkin diagram) by Fomin-Zelevinsky [15]; and for (A_n, A_m) by Volkov [45] and independently by Szenes [42]. Finally in 2008, Keller proved the conjecture in the general case [28, 29], using cluster algebras and their additive categorification via triangulated categories. Another proof was subsequently given by Inoue-Iyama-Keller-Kuniba-Nakanishi [24, 25].

In Sections 4.1 and 4.2 we will state the periodicity conjecture for Dynkin diagrams and pairs of Dynkin diagrams, respectively, and explain how the conjectures may be formulated in terms of cluster algebras. In Section 4.3 we will discuss how techniques from the theory of cluster algebras were used to prove the conjectures.

4.1. Zamolodchikov's Periodicity Conjecture for Dynkin diagrams.

Let Δ be a Dynkin diagram with vertex set I. Let A denote the incidence matrix of Δ, i.e. if C is the Cartan matrix of Δ and J the identity matrix of the same size, then $A = 2J - C$. Let h denote the Coxeter number of Δ, see Table 2.

Δ	A_n	B_n	C_n	D_n	E_6	E_7	E_8	F_4	G_2
h	$n+1$	$2n$	$2n$	$2n-2$	12	18	30	12	6

Table 2. Coxeter numbers.

Theorem 4.1 (*Zamolodchikov's periodicity conjecture*)**.** *Consider the recurrence relation*

$$Y_i(t+1)Y_i(t-1) = \prod_{j \in I} (Y_j(t)+1)^{a_{ij}}, \quad t \in \mathbb{Z}. \tag{4.1}$$

All solutions to this system are periodic in t with period dividing $2(h+2)$, i.e. $Y_i(t+2(h+2)) = Y_i(t)$ for all i and t.

The system of equations in (4.1) is called a *Y-system.*

Note that any Dynkin diagram is a tree, and hence its set I of vertices is the disjoint union of two sets I_+ and I_- such that there is no edge between any two vertices of I_+ nor between any two vertices of I_-. Define $\epsilon(i)$ to be 1 or -1 based on whether $i \in I_+$ or $i \in I_-$. Let $\mathbb{Q}(u)$ be the field of rational functions in the variables u_i for $i \in I$. For $\epsilon = \pm 1$, define an automorphism τ_ϵ by setting

$$\tau_\epsilon(u_i) = \begin{cases} u_i \prod_{j \in I} (u_j + 1)^{a_{ij}} & \text{if } \epsilon(i) = \epsilon \\ u_i^{-1} & \text{otherwise.} \end{cases}$$

One may reformulate Zamolodchikov's periodicity conjecture in terms of τ_ϵ, as we will see below in Lemma 4.4. First note that the variables $Y_i(k)$ on the left-hand side of (4.1) have a fixed "parity" $\epsilon(i)(-1)^k$. Therefore the Y-system decomposes into two independent systems, an even one and an odd one, and it suffices to prove periodicity for one of them. Without loss of generality, we may therefore assume that

$$Y_i(k+1) = Y_i(k)^{-1} \text{ whenever } \epsilon(i) = (-1)^k.$$

If we combine this assumption with (4.1), we obtain

$$Y_i(k+1) = \begin{cases} Y_i(k) \prod_{j \in I} (Y_j(k) + 1)^{a_{ij}} & \text{if } \epsilon(i) = (-1)^{k+1} \\ Y_i(k)^{-1} & \text{if } \epsilon(i) = (-1)^k. \end{cases} \tag{4.2}$$

Example 4.2. Let Δ be the Dynkin diagram of type A_2, on nodes 1 and 2, where $I_- = \{1\}$ and $I_+ = \{2\}$. The incidence matrix of the Dynkin diagram is

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

If we set $Y_1(0) = u_1$, $Y_2(0) = u_2$, then the recurrence for $Y_i(k)$ in (4.2) yields:

$$\begin{aligned} Y_1(0) &= u_1\,, & Y_2(0) &= u_2 \\ Y_1(1) &= u_1^{-1}\,, & Y_2(1) &= u_2(1+u_1) \\ Y_1(2) &= \frac{1+u_2+u_1u_2}{u_1}\,, & Y_2(2) &= \frac{1}{u_2(1+u_1)} \\ Y_1(3) &= \frac{u_1}{1+u_2+u_1u_2}\,, & Y_2(3) &= \frac{1+u_2}{u_1u_2} \\ Y_1(4) &= u_2^{-1}\,, & Y_2(4) &= \frac{u_1u_2}{1+u_2} \\ Y_1(5) &= u_2\,, & Y_2(5) &= u_1. \end{aligned}$$

By symmetry, it's clear that $Y_1(10) = u_1$ and $Y_2(10) = u_2$ and this system has period $10 = 2(3+2)$, as predicted by Theorem 4.1.

The following lemma follows easily from induction and the definition of τ_ϵ.

Lemma 4.3. *Set $Y_i(0) = u_i$ for $i \in I$. Then for all $k \in \mathbb{Z}_{\geq 0}$ and $i \in I$, we have $Y_i(k) = (\tau_- \tau_+ \dots \tau_\pm)(u_i)$, where the number of factors τ_+ and τ_- equals k.*

Let us define an automorphism of $\mathbb{Q}(u)$ by

$$\phi = \tau_- \tau_+. \tag{4.3}$$

Then we have the following.

Lemma 4.4. *The Y-system from* (4.1) *is periodic with period dividing* $2(h+2)$ *if and only if* ϕ *has finite order dividing* $h+2$.

To connect Zamolodchikov's conjecture to cluster algebras, let us revisit the notion of Y-seed mutation from Definition 2.23. We will assume that B is skew-symmetric, and hence can be encoded by a finite quiver Q without loops or 2-cycles. Let $(\mathbb{P}, \oplus, \cdot)$ be $\mathbb{Q}$ with the usual operations of addition and multiplication. Then $\mu_k(\mathbf{y}, Q) = (\mathbf{y}', Q')$, where

$$y'_j = \begin{cases} y_k^{-1} & \text{if } j = k \\ y_j(1+y_k)^m & \text{if there are } m \geq 0 \text{ arrows } j \to k \\ y_j(1+y_k^{-1})^{-m} & \text{if there are } m \geq 0 \text{ arrows } k \to j. \end{cases}$$

Comparing the formula for Y-seed mutation with the definition of the automorphisms τ_ϵ suggests a connection, which we make precise in Exercise 4.6. First we will define a *restricted Y-pattern*.

Definition 4.5 (*Restricted Y-pattern*)**.** Let $(\mathbb{P}, \oplus, \cdot)$ be $\mathbb{Q}$ with the usual operations of addition and multiplication. Let Q denote a finite quiver without loops or 2-cycles with vertex set $\{1, \ldots, n\}$, let $\mathbf{y} = (y_1, \ldots, y_n)$ and let $(\mathbf{y}, Q)$ be the corresponding Y-seed. Let $\mathbf{v}$ be a sequence of vertices $v_1, \ldots, v_N$ of Q, with the property that the composed mutation

$$\mu_{\mathbf{v}} = \mu_{v_N} \cdots \mu_{v_2} \mu_{v_1}$$

transforms Q into itself. Then clearly the same holds for the same sequence in reverse $\mu_{\mathbf{v}}^{-1}$. We define the *restricted Y-pattern* associated with Q and $\mu_{\mathbf{v}}$ to be the sequence of Y-seeds obtained from the initial Y-seed $(\mathbf{y}, Q)$ be applying all integer powers of $\mu_{\mathbf{v}}$.

Exercise 4.6. Let Δ be a simply-laced Dynkin diagram with n vertices, and vertex set $I = I_+ \cup I_-$ as above. Let Q denote the unique "bipartite" orientation of Δ such that each vertex in I_+ is a source and each vertex in I_- is a sink.

(1) Then the composed mutation $\mu_+ = \prod_{i \in I_-} \mu_i$ is well-defined, in other words, any sequence of mutations on the vertices in I_- yields the same result. Similarly $\mu_- = \prod_{i \in I_+} \mu_i$ is well-defined.
(2) The composed mutation $\mu_- \mu_+$ transforms Q into itself. Similarly for $\mu_+ \mu_-$.
(3) The automorphism $\tau_- \tau_+$ has finite order m if and only if the restricted Y-pattern associated with Q and $\mu_- \mu_+$ is periodic with period m.

Combining Exercise 4.6 with Lemma 4.3, we see that Zamolodchikov's periodicity conjecture is equivalent to verifying the periodicity of the restricted Y-pattern from Exercise 4.6 (3). See Figure 8 for an example of Y-seed mutation on the Dynkin diagram of type A_2. Compare the labeled Y-seeds here with Example 4.2.

4.2. **The periodicity conjecture for pairs of Dynkin diagrams.** In this section we let Δ and Δ' be Dynkin diagrams, with vertex sets I and I', and incidence matrices A and A'.

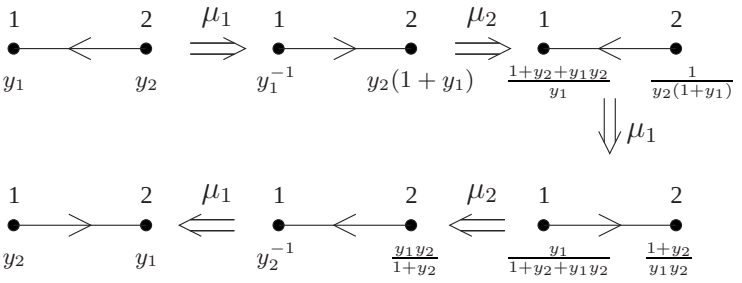


FIGURE 8. Y-seeds and Y-seed mutations in type A_2.

Theorem 4.7 (*The periodicity conjecture for pairs of Dynkin diagrams*)**.** *Consider the recurrence relation*

$$Y_{i,i'}(t+1)Y_{i,i'}(t-1) = \frac{\prod_{j\in I}(Y_{j,i'}(t)+1)^{a_{ij}}}{\prod_{j'\in I'}(Y_{i,j'}(t)^{-1}+1)^{a'_{i',j'}}}, \quad t \in \mathbb{Z}. \tag{4.4}$$

All solutions to this system are periodic in t with period dividing $2(h+h')$.

Note that if Δ' is of type A_1, then Theorem 4.7 reduces to Theorem 4.1.

Just as we saw for Theorem 4.1, it is possible to reformulate Theorem 4.7 in terms of certain automorphisms. Write $I = I_+ \cup I_-$ and $I' = I'_+ \cup I'_-$ as before. For a vertex (i,i') of the product $I \times I'$, define $\epsilon(i,i')$ to be 1 or -1 based on whether (i,i') lies in $(I_+ \times I'_+) \cup (I_- \times I'_-)$ or not. Let $\mathbb{Q}(u)$ be the field of rational functions in the variables $u_{ii'}$ for $i \in I$ and $i' \in I'$, and define an automorphism of $\mathbb{Q}(u)$ by

$$\phi = \tau_-\tau_+, \text{ where} \tag{4.5}$$

$$\tau_\epsilon(u_{ii'}) = \begin{cases} u_{ii'}\prod_{j\in I}(u_{ji'}+1)^{a_{ij}}\prod_{j'\in I'}(u_{ij'}^{-1}+1)^{-a'_{i'j'}} & \text{if } \epsilon(i,i') = \epsilon \\ u_{ii'}^{-1} & \text{otherwise.} \end{cases}$$

As before, we may assume that $Y_{i,i'}(k+1) = Y_{i,i'}(k)^{-1}$ whenever $\epsilon(i,i') = (-1)^k$. One may then reformulate the periodicity conjecture as follows.

Lemma 4.8. *The periodicity conjecture for pairs of Dynkin diagrams holds if and only if ϕ has finite order dividing $h+h'$.*

Now let us explain how to relate the periodicity conjecture for pairs of Dynkin diagrams to cluster algebras (more specifically, restricted Y-patterns). We first need to define some operations on quivers.

Let Q and Q' be two finite quivers on vertex sets I and I' which are bipartite, i.e. each vertex is a source or a sink. The *tensor product* $Q \otimes Q'$ is the quiver on vertex set $I \times I'$, where the number of arrows from a vertex (i,i') to a vertex (j,j')

(1) is zero if $i \neq j$ and $i' \neq j'$;
(2) equals the number of arrows from j to j' if $i = i'$;
(3) equals the number of arrows from i to i' if $j = j'$.

The *square product* $Q \square Q'$ is the quiver obtained from $Q \otimes Q'$ by reversing all arrows in the full subquivers of the form $\{i\} \times Q'$ and $Q \times \{i'\}$, where i is a sink of Q and i' a source of Q'. See Figure 9 for an example of the square product of the following

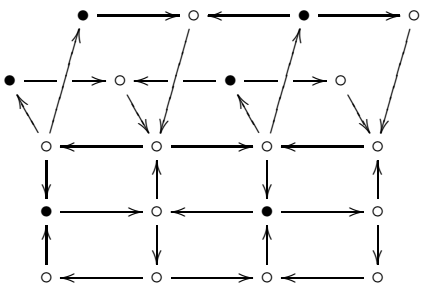


FIGURE 9. The quiver $\vec{A}_4 \square \vec{D}_5$

quivers.

$$\vec{A}_4 : \ 1 \longleftarrow 2 \longrightarrow 3 \longleftarrow 4 \ ,$$

$$\vec{D}_5 : \ 1 \longleftarrow 2 \longrightarrow 3 \begin{array}{l} \nearrow 4 \\ \searrow 5. \end{array}$$

Exercise 4.9. Let Δ and Δ' be simply-laced Dynkin diagrams with vertex sets I and I'. We write $I = I_+ \cup I_-$ and $I' = I'_+ \cup I'_-$ as usual, and choose the corresponding bipartite orientations Q and Q' of Δ and Δ' so that each vertex in I_+ or I'_+ is a source and each vertex in I_- or I'_- is a sink.

(1) Given two elements σ, σ' of $\{+,-\}$, the following composed mutation

$$\mu_{\sigma,\sigma'} = \prod_{i \in I_\sigma,\ i' \in I'_{\sigma'}} \mu_{(i,i')}$$

of $Q \square Q'$ is well-defined, that is, the order in the product does not matter.

(2) The composition $\mu_\square = \mu_{-,-}\ \mu_{+,+}\ \mu_{-,+}\ \mu_{+,-}$ transforms Q into itself.

(3) The automorphism ϕ has finite order m if and only if the restricted Y-seed associated with $Q \square Q'$ and $\mu_\square$ is periodic with period m.

Combining Exercise 4.9 with Lemma 4.8, we have the following:

Lemma 4.10. *The periodicity conjecture holds for Δ and Δ' if and only if the restricted Y-seed associated with $Q \square Q'$ and $\mu_\square$ is periodic with period dividing $h + h'$.*

4.3. **On the proofs of the periodicity conjecture.** In this section we discuss how techniques from the theory of cluster algebras were used to prove Theorems 4.1 and 4.7.

First note that the proofs of Theorem 4.1 and Theorem 4.7 can be reduced to the case that the Dynkin diagrams are simply-laced, using standard "folding" arguments. Second, as illustrated in Exercises 4.6 and 4.9, the periodicity conjecture for simply-laced Dynkin diagrams may be reformulated in terms of cluster algebras. Specifically, the conjecture is equivalent to verifying the periodicity of certain restricted Y-patterns.

Fomin-Zelevinsky's proof of Theorem 4.1 used ideas which are now fundamental to the structure theory of finite type cluster algebras, including a bijection between cluster variables and "almost-positive" roots of the corresponding root system. They showed that this bijection, together with a "tropical" version of Theorem 4.1, implies Theorem 4.1. Moreover, they gave an explicit solution to each Y-system, in terms of certain *Fibonacci polynomials*. The Fibonacci polynomials are (up to a twist) special cases of *F-polynomials*, which in turn are important objects in cluster algebras, and control the dynamics of both cluster and coefficient variables [16]. Note however that the Fomin-Zelevinsky proof does not apply to Theorem 4.7, because the cluster algebras associated with products $Q \square Q'$ are not in general of finite type.

Keller's proof of Theorem 4.7 used the *additive categorification* (via triangulated categories) of cluster algebras. To give some background on categorification, in 2003, Marsh-Reineke-Zelevinsky [33] discovered that when Δ is a simply-laced Dynkin diagram, there is a close resemblance between the combinatorics of the cluster variables and those of the *tilting modules* in the category of representations of the quiver; this initiated the theory of *additive categorification* of cluster algebras. In this theory, one seeks to construct module or triangulated categories associated to quivers so as to obtain a correspondence between rigid objects of the categories and the cluster monomials in the cluster algebras. The required correspondence sends direct sum decompositions of rigid objects to factorizations of the associated cluster monomials. One may then hope to use the rich structure of these categories to prove results on cluster algebras which seem beyond the scope of purely combinatorial methods.

Recall from Section 4.2 that the periodicity conjecture for pairs of Dynkin diagrams is equivalent to the periodicity of the automorphism ϕ from (4.5), which in turn is equivalent to the periodicity of a restricted Y-pattern associated to $Q \square Q'$. Keller's central construction from [29] was a triangulated 2-Calabi-Yau category $\mathcal{C}$ with a cluster-tilting object T, whose endoquiver (quiver of its endomorphism algebra) is closely related to $Q \square Q'$; the category $\mathcal{C}$ is a generalized cluster category in the sense of Amiot [1]. Since $\mathcal{C}$ is 2-Calabi-Yau, results of Iyama-Yoshino [26] imply that there is a well-defined mutation operation for the cluster-tilting objects. Keller defined the *Zamolodchikov transformation* $Za : \mathcal{C} \to \mathcal{C}$, which one may think of as a categorification of the automorphism ϕ, and proved that Za is periodic of period $h + h'$. By "decategorification," it follows that ϕ is periodic of period $h + h'$, and hence the periodicity conjecture for pairs of Dynkin diagrams is true.

The Inoue-Iyama-Keller-Kuniba-Nakanishi proof of Theorem 4.7 also used categorification, in particular the work of Plamondon [39]. Moreover, just as in the case of the Fomin-Zelevinsky proof of Theorem 4.7, one crucial ingredient in their proof was a "tropical" version of Theorem 4.7.

References

1. Claire Amiot, *Cluster categories for algebras of global dimension 2 and quivers with potential*, Ann. Inst. Fourier (Grenoble) **59** (2009), no. 6, 2525–2590. MR 2640929 (2011c:16026)
2. Philippe Caldero and Markus Reineke, *On the quiver Grassmannian in the acyclic case*, J. Pure Appl. Algebra **212** (2008), no. 11, 2369–2380. MR 2440252 (2009f:14102)
3. Giovanni Cerulli Irelli, Bernhard Keller, Daniel Labardini-Fragoso, and Pierre-Guy Plamondon, *Linear independence of cluster monomials for skew-symmetric cluster algebras*, ArXiv Mathematics e-prints (2012), arXiv:1203.1307.

4. Frédéric Chapoton, Sergey Fomin, and Andrei Zelevinsky, *Polytopal realizations of generalized associahedra*, Canad. Math. Bull. **45** (2002), no. 4, 537–566, Dedicated to Robert V. Moody. MR 1941227 (2003j:52014)
5. Philippe Di Francesco and Rinat Kedem, *Q-systems, heaps, paths and cluster positivity*, Comm. Math. Phys. **293** (2010), no. 3, 727–802. MR 2566162 (2010m:13032)
6. Anna Felikson, Michael Shapiro, and Pavel Tumarkin, *Cluster algebras of finite mutation type via unfoldings*, Int. Math. Res. Not. IMRN (2012), no. 8, 1768–1804. MR 2920830
7. ———, *Skew-symmetric cluster algebras of finite mutation type*, J. Eur. Math. Soc. (2012), no. 14, 1135–1180.
8. Vladimir Fock and Alexander Goncharov, *Moduli spaces of local systems and higher Teichmüller theory*, Publ. Math. Inst. Hautes Études Sci. (2006), no. 103, 1–211. MR 2233852 (2009k:32011)
9. ———, *Dual Teichmüller and lamination spaces*, Handbook of Teichmüller theory. Vol. I, IRMA Lect. Math. Theor. Phys., vol. 11, Eur. Math. Soc., Zürich, 2007, pp. 647–684. MR 2349682 (2008k:32033)
10. ———, *Cluster ensembles, quantization and the dilogarithm*, Ann. Sci. Éc. Norm. Supér. (4) **42** (2009), no. 6, 865–930. MR 2567745 (2011f:53202)
11. Sergey Fomin, Michael Shapiro, and Dylan Thurston, *Cluster algebras and triangulated surfaces. I. Cluster complexes*, Acta Math. **201** (2008), no. 1, 83–146. MR 2448067 (2010b:57032)
12. Sergey Fomin and Dylan Thurston, *Cluster algebras and triangulated surfaces. part ii: Lambda lengths*, ArXiv Mathematics e-prints (2012), arXiv:1210.5569.
13. Sergey Fomin and Andrei Zelevinsky, *Cluster algebras. I. Foundations*, J. Amer. Math. Soc. **15** (2002), no. 2, 497–529 (electronic). MR 1887642 (2003f:16050)
14. ———, *Cluster algebras. II. Finite type classification*, Invent. Math. **154** (2003), no. 1, 63–121. MR 2004457 (2004m:17011)
15. ———, *Y-systems and generalized associahedra*, Ann. of Math. (2) **158** (2003), no. 3, 977–1018. MR 2031858 (2004m:17010)
16. ———, *Cluster algebras. IV. Coefficients*, Compos. Math. **143** (2007), no. 1, 112–164. MR 2295199 (2008d:16049)
17. Edward Frenkel and András Szenes, *Thermodynamic Bethe ansatz and dilogarithm identities. I*, Math. Res. Lett. **2** (1995), no. 6, 677–693. MR 1362962 (97a:11182)
18. Christof Geiss, Bernard Leclerc, and Jan Schröer, *Partial flag varieties and preprojective algebras*, Ann. Inst. Fourier (Grenoble) **58** (2008), no. 3, 825–876. MR 2427512 (2009f:14104)
19. ———, *Preprojective algebras and cluster algebras*, Trends in representation theory of algebras and related topics, EMS Ser. Congr. Rep., Eur. Math. Soc., Zürich, 2008, pp. 253–283. MR 2484728 (2009m:16024)
20. Michael Gekhtman, Michael Shapiro, and Alek Vainshtein, *Cluster algebras and Weil-Petersson forms*, Duke Math. J. **127** (2005), no. 2, 291–311. MR 2130414 (2006d:53103)
21. Ferdinando Gliozzi and Roberto Tateo, *Thermodynamic Bethe ansatz and three-fold triangulations*, Internat. J. Modern Phys. A **11** (1996), no. 22, 4051–4064. MR 1403679 (97e:82014)
22. John Harer, *The virtual cohomological dimension of the mapping class group of an orientable surface*, Invent. Math. **84** (1986), no. 1, 157–176. MR 830043 (87c:32030)
23. Allen Hatcher, *On triangulations of surfaces*, Topology Appl. **40** (1991), no. 2, 189–194. MR 1123262 (92f:57020)
24. Rei Inoue, Osama Iyama, Bernhard Keller, Atsuo Kuniba, and Tomoki Nakanishi, *Periodicities of t and y-systems, dilogarithm identities, and cluster algebras i: Type b_r.*, ArXiv Mathematics e-prints (2010), arXiv:1001.1880, to appear in Publ. RIMS.
25. ———, *Periodicities of t and y-systems, dilogarithm identities, and cluster algebras ii: Types c_r, f_4, and g_2.*, ArXiv Mathematics e-prints (2010), arXiv:1001.1881, to appear in Publ. RIMS.
26. Osamu Iyama and Yuji Yoshino, *Mutation in triangulated categories and rigid Cohen-Macaulay modules*, Invent. Math. **172** (2008), no. 1, 117–168. MR 2385669 (2008k:16028)
27. Masaki Kashiwara, *On crystal bases of the Q-analogue of universal enveloping algebras*, Duke Math. J. **63** (1991), no. 2, 465–516. MR 1115118 (93b:17045)
28. Bernhard Keller, *Cluster algebras, quiver representations and triangulated categories*, Triangulated categories, London Math. Soc. Lecture Note Ser., vol. 375, Cambridge Univ. Press, Cambridge, 2010, pp. 76–160. MR 2681708 (2011h:13033)
29. ———, *The periodicity conjecture for pairs of dynkin diagrams*, Ann. Math. **177** (2013).

30. Atsuo Kuniba and Tomoki Nakanishi, *Spectra in conformal field theories from the Rogers dilogarithm*, Modern Phys. Lett. A **7** (1992), no. 37, 3487–3494. MR 1192727 (94c:81185)
31. Atsuo Kuniba, Tomoki Nakanishi, and Junji Suzuki, *Functional relations in solvable lattice models. I. Functional relations and representation theory*, Internat. J. Modern Phys. A **9** (1994), no. 30, 5215–5266. MR 1304818 (96h:82003)
32. George Lusztig, *Canonical bases arising from quantized enveloping algebras*, J. Amer. Math. Soc. **3** (1990), no. 2, 447–498. MR 1035415 (90m:17023)
33. Robert Marsh, Markus Reineke, and Andrei Zelevinsky, *Generalized associahedra via quiver representations*, Trans. Amer. Math. Soc. **355** (2003), no. 10, 4171–4186. MR 1990581 (2004g:52014)
34. Lee Mosher, *Tiling the projective foliation space of a punctured surface*, Trans. Amer. Math. Soc. **306** (1988), no. 1, 1–70. MR 927683 (89f:57014)
35. Gregg Musiker, Ralf Schiffler, and Lauren Williams, *Positivity for cluster algebras from surfaces*, Adv. Math. **227** (2011), no. 6, 2241–2308. MR 2807089 (2012f:13052)
36. Hiraku Nakajima, *Quiver varieties and cluster algebras*, Kyoto J. Math. **51** (2011), no. 1, 71–126. MR 2784748 (2012f:13053)
37. Robert Penner, *The decorated Teichmüller space of punctured surfaces*, Comm. Math. Phys. **113** (1987), no. 2, 299–339. MR 919235 (89h:32044)
38. ———, *Decorated Teichmüller theory of bordered surfaces*, Comm. Anal. Geom. **12** (2004), no. 4, 793–820. MR 2104076 (2006a:32018)
39. Pierre-Guy Plamondon, *Cluster algebras via cluster categories with infinite-dimensional morphism spaces*, Compos. Math. **147** (2011), no. 6, 1921–1934. MR 2862067
40. Francesco Ravanini, Angelo Valleriani, and Roberto Tateo, *Dynkin TBAs*, Internat. J. Modern Phys. A **8** (1993), no. 10, 1707–1727. MR 1216231 (94h:81149)
41. Joshua Scott, *Grassmannians and cluster algebras*, Proc. London Math. Soc. (3) **92** (2006), no. 2, 345–380. MR 2205721 (2007e:14078)
42. András Szenes, *Periodicity of Y-systems and flat connections*, Lett. Math. Phys. **89** (2009), no. 3, 217–230. MR 2551180 (2011b:81116)
43. William Thurston, *On the geometry and dynamics of diffeomorphisms of surfaces*, Bull. Amer. Math. Soc. (N.S.) **19** (1988), no. 2, 417–431. MR 956596 (89k:57023)
44. William P. Thurston, *The geometry and topology of three-manifolds*, Princeton University notes, 1980.
45. Alexandre Yu. Volkov, *On the periodicity conjecture for Y-systems*, Comm. Math. Phys. **276** (2007), no. 2, 509–517. MR 2346398 (2008m:17021)
46. Al. B. Zamolodchikov, *On the thermodynamic Bethe ansatz equations for reflectionless ADE scattering theories*, Phys. Lett. B **253** (1991), no. 3-4, 391–394. MR 1092210 (92a:81196)

Department of Mathematics, University of California, Berkeley, CA 94720
E-mail address: williams@math.berkeley.edu

CURRENT EVENTS BULLETIN
Previous speakers and titles

For PDF files of talks, and links to Bulletin of the AMS articles, see
http://www.ams.org/ams/current-events-bulletin.html.

January 6, 2012 (Boston, MA)

Jeffrey Brock, Brown University
Assembling surfaces from random pants: the surface-subgroup and Ehrenpreis conjectures

Daniel Freed, University of Texas at Austin
The cobordism hypothesis: quantum field theory + homotopy invariance = higher algebra

Gigliola Staffilani, Massachusetts Institute of Technology
Dispersive equations and their role beyond PDE

Umesh Vazirani, University of California, Berkeley
How does quantum mechanics scale?

January 6, 2011 (New Orleans, LA)

Luca Trevisan, Stanford University
Khot's unique games conjecture: its consequences and the evidence for and against it

Thomas Scanlon, University of California, Berkeley
Counting special points: logic, Diophantine geometry and transcendence theory

Ulrike Tillmann, Oxford University
Spaces of graphs and surfaces

David Nadler, Northwestern University
The geometric nature of the Fundamental Lemma

January 15, 2010 (San Francisco, CA)

Ben Green, University of Cambridge
Approximate groups and their applications: work of Bourgain, Gamburd, Helfgott and Sarnak

David Wagner, University of Waterloo
Multivariate stable polynomials: theory and applications

Laura DeMarco, University of Illinois at Chicago
The conformal geometry of billiards

Michael Hopkins, Harvard University
On the Kervaire Invariant Problem

January 7, 2009 (Washington, DC)

Matthew James Emerton, Northwestern University
Topology, representation theory and arithmetic: Three-manifolds and the Langlands program

Olga Holtz, University of California, Berkeley
Compressive sensing: A paradigm shift in signal processing

Michael Hutchings, University of California, Berkeley
From Seiberg-Witten theory to closed orbits of vector fields: Taubes's proof of the Weinstein conjecture

Frank Sottile, Texas A & M University
Frontiers of reality in Schubert calculus

January 8, 2008 (San Diego, California)

Günther Uhlmann, University of Washington
Invisibility

Antonella Grassi, University of Pennsylvania
Birational Geometry: Old and New

Gregory F. Lawler, University of Chicago
Conformal Invariance and 2-d Statistical Physics

Terence C. Tao, University of California, Los Angeles
Why are Solitons Stable?

January 7, 2007 (New Orleans, Louisiana)

Robert Ghrist, University of Illinois, Urbana-Champaign
Barcodes: The persistent topology of data

Akshay Venkatesh, Courant Institute, New York University
Flows on the space of lattices: work of Einsiedler, Katok and Lindenstrauss

Izabella Laba, University of British Columbia
From harmonic analysis to arithmetic combinatorics

Barry Mazur, Harvard University
The structure of error terms in number theory and an introduction to the Sato-Tate Conjecture

January 14, 2006 (San Antonio, Texas)

Lauren Ancel Myers, University of Texas at Austin
Contact network epidemiology: Bond percolation applied to infectious disease prediction and control

Kannan Soundararajan, University of Michigan, Ann Arbor
Small gaps between prime numbers

Madhu Sudan, MIT
Probabilistically checkable proofs

Martin Golubitsky, University of Houston
Symmetry in neuroscience

January 7, 2005 (Atlanta, Georgia)

Bryna Kra, Northwestern University
The Green-Tao Theorem on primes in arithmetic progression: A dynamical point of view

Robert McEliece, California Institute of Technology
Achieving the Shannon Limit: A progress report

Dusa McDuff, SUNY at Stony Brook
Floer theory and low dimensional topology

Jerrold Marsden, Shane Ross, California Institute of Technology
New methods in celestial mechanics and mission design

László Lovász, Microsoft Corporation
Graph minors and the proof of Wagner's Conjecture

January 9, 2004 (Phoenix, Arizona)

Margaret H. Wright, Courant Institute of Mathematical Sciences, New York University
The interior-point revolution in optimization: History, recent developments and lasting consequences

Thomas C. Hales, University of Pittsburgh
What is motivic integration?

Andrew Granville, Université de Montréal
It is easy to determine whether or not a given integer is prime

John W. Morgan, Columbia University
Perelman's recent work on the classification of 3-manifolds

January 17, 2003 (Baltimore, Maryland)

Michael J. Hopkins, MIT
Homotopy theory of schemes

Ingrid Daubechies, Princeton University
Sublinear algorithms for sparse approximations with excellent odds

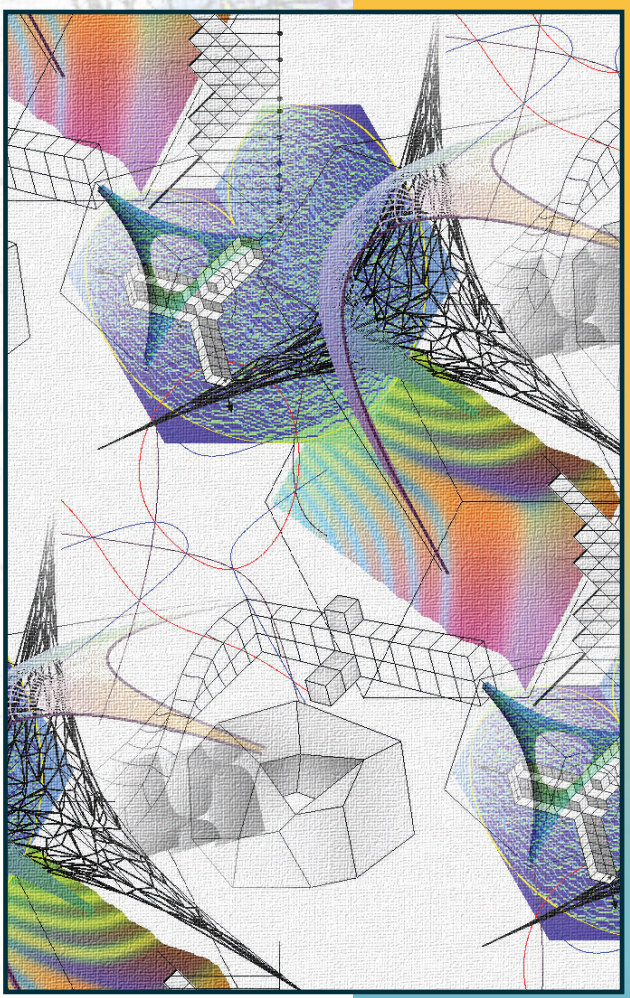
Edward Frenkel, University of California, Berkeley
Recent advances in the Langlands Program

Daniel Tataru, University of California, Berkeley
The wave maps equation

2013 CURRENT EVENTS BULLETIN

Committee

Hélène Barcelo, *Mathematical Sciences Research Institute*
Jeffrey Brock, *Brown University*
David Eisenbud, *University of California, Berkeley, Chair*
Dan Freed, *University of Texas, Austin*
Eric Friedlander, *University of Southern California*
Susan Friedlander, *University of Southern California*
Andrew Granville, *Université de Montréal*
Richard Karp, *University of California, Berkeley*
Isabella Laba, *University of British Columbia*
David Nadler, *University of California, Berkeley*
Thomas Scanlon, *University of California, Berkeley*
Gigliola Staffilani, *Massachusetts Institute of Technology*
Ulrike Tillmann, *Oxford University*
Luca Trevisan, *Stanford University*
Yuri Tschinkel, *New York University*
Akshay Venkatesh, *Stanford University*
Andrei Zelevinsky, *Northeastern University*



The back cover graphic is reprinted courtesy of Andrei Okounkov.
Cover graphic associated with Sam Payne's talk courtesy of Matthew Baker, Georgia Institute of Technology.
Cover graphic associated with Mladen Bestvina's talk was created by Silvio Levy.