American Mathematical Society

CURRENT EVENTS BULLETIN

Friday, January 17, 2014, 1:00 PM to 5:00 PM

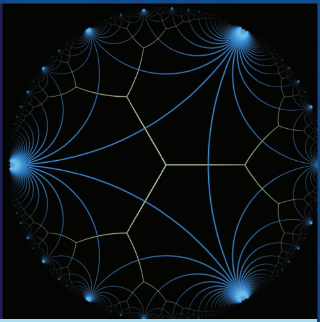
Room 310 Baltimore Convention Center
Joint Mathematics Meetings, Baltimore, MD

1:00 PM

Daniel Rothman
Massachusetts Institute of Technology

2:00 PM

Karen Vogtmann
Cornell University



Earth's Carbon Cycle: A Mathematical Perspective

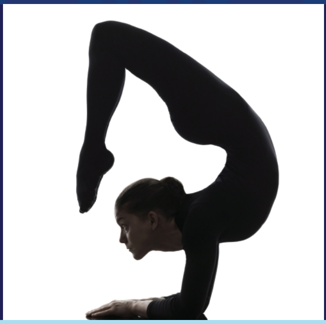
Mathematics to understand one of the great challenges to our society

The geometry of Outer space

New geometric methods advance the theory of automorphism groups of free groups

3:00 PM

Yakov Eliashberg
Stanford University



4:00 PM

Andrew Granville
Université de Montréal

Primes get closer and closer together

Recent advances in symplectic flexibility

Flexible methods (known as Gromov's h-principle generalizing the work of Nash and Smale) played important role in symplectic topology from its inception. Learn about the classic results and their new developments.

Infinitely many pairs of primes differ by no more than 70 million (and the bound's getting smaller every day)

-- Prime Twins?

-- Dunno, but at least they're close.

Organized by David Eisenbud, University of California, Berkeley

AMS
American Mathematical Society

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.

EARTH'S CARBON CYCLE: A MATHEMATICAL PERSPECTIVE

DANIEL H. ROTHMAN

ABSTRACT. The carbon cycle represents metabolism at a global scale. When viewed through a mathematical lens, observational data suggest that the cycle exhibits an underlying mathematical structure. This talk focuses on two types of emerging results: evidence of global dynamical coupling between life and the environment, and an understanding of the ways in which smaller-scale processes determine the strength of that coupling. Such insights are relevant not only to predicting future climate but also to understanding the long-term co-evolution of life and the environment.

1. Introduction

The concentration of carbon dioxide in the atmosphere is rising (Figure 1) as a consequence of the burning of fossil fuels [63, 37]. Because CO_2 traps heat, Earth's climate is expected to become warmer [51]. These observations alone make the fluxes of carbon into and out of the atmosphere a matter of intense current interest [35]. But there is a bigger story here, one that has played out since the origin of life.

Plants, fueled by the Sun's radiation, convert CO_2 to sugar and other carbohydrates by *photosynthesis*. Organisms that consume plants derive their energy from the oxidation of those photosynthetic products. Further up the food chain, organisms that consume other consumers ultimately derive their energy from the same source. In these ways, the CO_2 that had been taken out of the atmosphere and oceans by plants is returned from where it came; the process is called *respiration*. Integrated over all organisms and all environments, the loop between photosynthesis and respiration makes up the biological component of Earth's *carbon cycle*. The flux through the loop is enormous: about one hundred gigatons (1 gigaton = 10^{15} grams) of carbon pass through it each year [34], more than an order of magnitude greater than the present rate of fossil fuel emissions [10].

The loop, however, contains a tiny leak [33, 8]. About 0.1% of the organic carbon produced by photosynthesis escapes respiration and instead becomes rock. Geologic processes normally bring the rock back to the surface environment, where the ancient organic matter is oxidized; the previously buried carbon then finally reenters the atmosphere as CO_2. The extraction and burning of fossil fuels amounts to a roughly hundred-fold speed up of the leak's natural rate of reinjection.

The leak has another, more profound, consequence. To recognize it, we write the carbon cycle as a chemical reaction:

$$CO_2 + H_2O \rightleftharpoons CH_2O + O_2 \tag{1.1}$$

2010 *Mathematics Subject Classification.* Primary 86-02; Secondary 92B99.

I would like to thank T. Bosak, C. Follett, and D. Forney for collaborations on problems discussed in Section 2. This work was supported by NASA Astrobiology (NNA08CN84A and NNA13AA90A) and NSF (OCE-0930866).

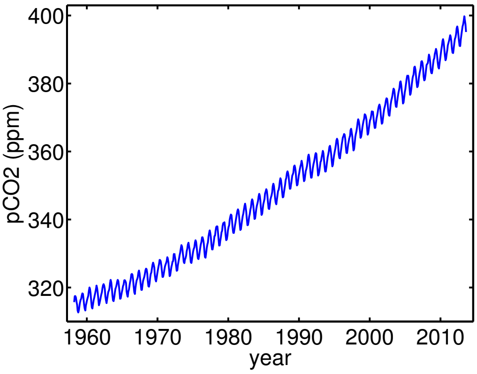


Figure 1. Partial pressure of atmospheric CO_2 at Mauna Loa Observatory, Hawaii [64]. The oscillatory component represents seasonal imbalances between photosynthesis and respiration.

The reaction from left-to-right is photosynthesis, via which CO_2 is reduced to a schematic carbohydrate (CH_2O) and free oxygen is created. The back reaction is respiration: organic matter is oxidized to CO_2, consuming the O_2 that had been produced by photosynthesis. Now note that if some CH_2O can exit the cycle and be buried as rock, a corresponding amount of O_2 must accumulate elsewhere, notably the atmosphere. The oxygen we breathe thus owes its existence to the leak [33, 8, 16, 30]. So too does the advent of complex multicellular life—animals with aerobic metabolisms—about 550 million years ago [42], roughly three billion years after the origin of life [39].

Earth's carbon cycle therefore plays a major role in determining not only CO_2 levels and climate but also life's interaction with the physical environment. Understanding the bigger story—the natural cycle, at all scales of time and space—provides a scientific foundation for understanding the consequences of the rising CO_2 levels in Figure 1. More generally, it provides a window into the complex interactions between the many components of the natural world.

Mathematics aids the pursuit. The carbon cycle is an unwieldy beast, but when one strips away insignificant complications, manifestations of elementary mathematical concepts emerge. Our focus here is phenomenological. We feature observational data and its mathematical interpretation, with the objective of providing targets for advancing theoretical understanding. Two types of problems receive special attention. Section 2 is devoted to the problem of *decomposition*: the processes by which organic matter is converted to CO_2. We illustrate ways in which the heterogeneity of the problem can be understood and feature scaling laws that appear to result from this heterogeneity. We then discuss problems of *dynamics* (Section 3), using historical records of past changes to illustrate the scope of the problem. We conclude with an appraisal of the lessons learned and the challenges ahead.

2. Decomposition

After photosynthesis creates organic compounds from CO_2, organisms proceed to decompose those compounds to gain energy. As a result, the organic carbon *decays*.

The process plays out at all scales of space and time and in all environments suitable for life. Surprisingly, a common structure emerges despite intrinsic heterogeneity.

2.1. **Fast time scales, on land.**

Forests are widely regarded as sinks for CO_2. But all that plant matter eventually decays, providing a CO_2 source. How can we characterize the decay? Typical studies measure the dry mass of carbon, $g(t)$, in a reasonably controlled setting. We proceed to consider the example of "litter bags" [27] of plant matter left on the ground over a period of years.

Figure 2a shows data obtained from one such experiment. To understand the data, consider first a simple model where $\dot{g} = -kg$, with k a rate constant dependent upon the plant matter composition, environmental conditions, and the microbial community. The resulting exponential decay would imply a straight line in Figure 2a. However the data in Figure 2a do not appear to describe a straight line, suggesting that the simple model is inadequate.

Another approach [13], sometimes referred to as *disordered kinetics* [66, 54, 52], proceeds from an assumption of heterogeneity. Some bits of a leaf may decay slowly, other bits quickly. This reasoning suggests the existence of a probability density function $p(k)$ of decay constants. Linear superposition then yields the decay

$$\frac{g(t)}{g(0)} = \int_0^\infty p(k)e^{-kt}\mathrm{d}k. \tag{2.1}$$

In other words, decay is characterized by the Laplace transform of $p(k)$. In principle, the inverse Laplace transform of $g(t)$ provides $p(k)$. Because this inverse problem is ill-posed, solutions are found by regularization [22].

Figure 2b shows the distribution of rate constants that characterize the decay in Figure 2a, plotted as $\rho(x) = p[k(x)]\mathrm{d}k/\mathrm{d}x$, where $x = \ln k$. The result is clearly consistent with a Gaussian, suggesting that $p(k)$ is lognormal with parameters μ and σ^2 representing the mean and variance, respectively, of $\ln k$. Figures 2c and 2d show the average of 182 different solutions, for different plant matter in different forests throughout North America, each rescaled to have zero mean and unit variance on the $\ln k$ axis. The result is again Gaussian, suggesting that a lognormal distribution of time scales is a general characteristic of plant-matter decay.

Why is the lognormal ubiquitous? The simplest answer appeals to unspecified multiplicative processes and the central limit theorem [5]. A more mechanistic interpretation derives from the following argument [61, 46, 22]. We suppose that the decay of a particular component of plant matter requires the satisfaction of various conditions, such as the presence of water, the presence of a specific micro-organism, the expression of particular enzymes, etc. Then if the probability P of decomposing that component during a small time Δt is the product of independent probabilities of satisfying many requirements, the rate constant $k = P/\Delta t$ is lognormal.

This natural manifestation of the central limit theorem, a kind of elementary universality [36], suggests a common form of plant-matter decay. The parameters μ and σ turn out to be related to climatic conditions and the composition of the plant matter [22]. To understand their practical importance, consider first the average rate constant

$$\langle k \rangle = \int_0^\infty kp(k)\mathrm{d}k, \tag{2.2}$$

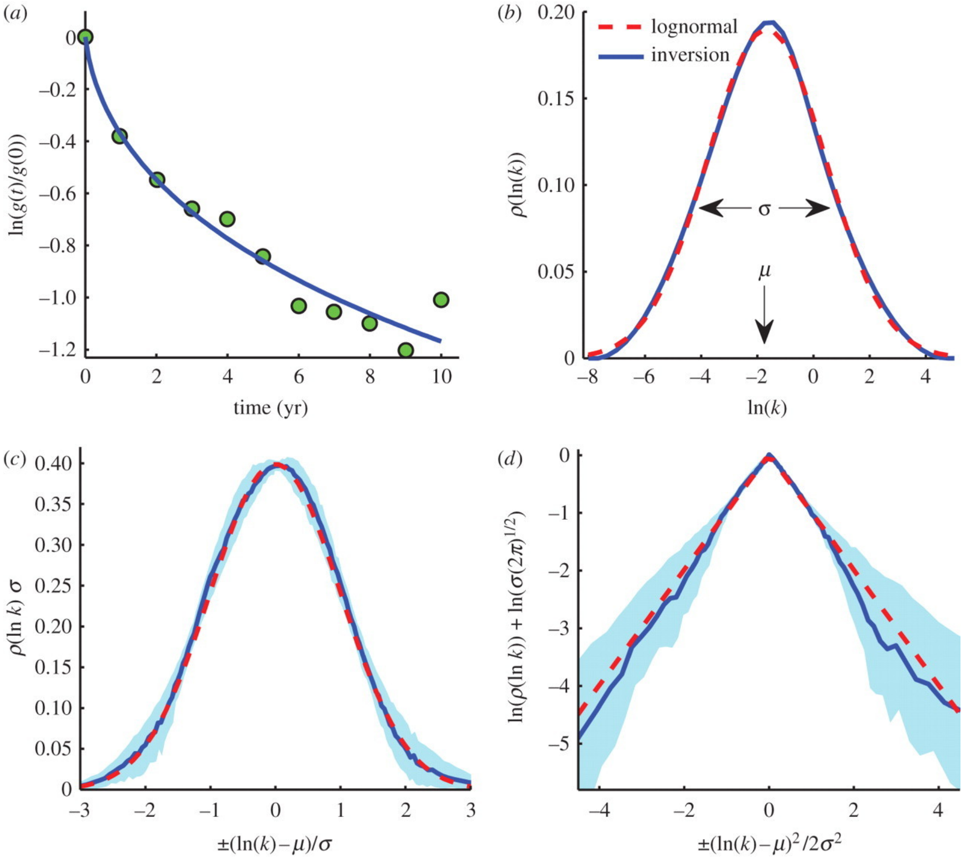


FIGURE 2. Analysis of plant-matter decay [22]. (a) Experimental results from a single sample. The smooth curve is the Laplace transform of the solid (blue) curve in (b). (b) The solid curve (blue) is the solution $\rho(\ln k)$ obtained by inversion of the data in (a). The dashed curve (red) is the best-fitting Gaussian, with variance σ^2 and mean μ. (c) The solid curve (blue) is the average of 182 rescaled solutions $\rho(\ln k)$. The dashed curve (red) is a Gaussian with zero mean and unit variance; the shaded area represents the scatter of solutions. (d) Logarithmic transformation of the results of (c). The dashed (red) straight lines indicate an exact lognormal distribution.

which is also the apparent rate constant $\dot{g}/g$ at $t = 0$. Inserting the lognormal distribution for $p(k)$, we find

$$\langle k \rangle = e^{\mu+\sigma^2/2}. \tag{2.3}$$

One might think that the inverse of the mean rate, $\langle k \rangle^{-1}$, would estimate the mean lifetime of organic matter. But it does not; the average time to decay is instead the *turnover time* $\tau = \langle k^{-1} \rangle$. When $p(k)$ is lognormal [22],

$$\tau = e^{-\mu+\sigma^2/2} = \langle k \rangle^{-1} e^{\sigma^2}. \tag{2.4}$$

We see that the mean lifetime τ equals $\langle k \rangle^{-1}$ only when $\sigma = 0$; otherwise τ increases exponentially with σ^2. Typical values of σ for plant-matter decay range from about 1–2 [22].

2.2. **Intermediate time scales, at sea.** About half of annual photosynthetic primary production occurs at sea, in the upper hundred meters of the ocean where sunlight penetrates [34]. About 90% of the production is consumed in those shallow waters. The remainder sinks, over a time scale of about a month, during which about 90% of the sinking fraction is consumed [31].

There are about 700 gigatons (Gt) of organic carbon in the oceans [34], which presumably represents a steady-state balance between production and respiration. Nearly all of the organic matter in the ocean is effectively dissolved [26]. Each year, about 50 Gt of carbon are fixed by marine photosynthetic organisms. The assumption of a steady state then allows the turnover time of marine organic carbon to be straightforwardly calculated:

$$\tau = \frac{700\ \text{Gt}}{50\ \text{Gt/yr}} = 14\ \text{yr}. \tag{2.5}$$

Modern radiocarbon methods provide an estimate of the mean age of the dissolved phase. The mean age turns out to be about 5000 yr [18], more than 300 times greater than the mean lifetime τ! What happened?

To gain insight, we imagine that the same model (2.1) of disordered kinetics that describes terrestrial plant-matter decay also applies to the oceans, and seek an expression for the mean age. Taking a to be the age of a parcel of organic carbon—i.e., the duration of time since its photosynthetic creation—the steady-state age distribution $p_a(a)$ is given by [11]

$$p_a(a) = \frac{\langle e^{-ka} \rangle}{\langle k^{-1} \rangle}, \tag{2.6}$$

where angle brackets again represent averages taken with respect to $p(k)$. The mean age $\bar{a} = \int_0^\infty a p_a(a) \mathrm{d}a$; thus[1]

$$\bar{a} = \frac{\langle k^{-2} \rangle}{\langle k^{-1} \rangle}. \tag{2.7}$$

If $p(k)$ is lognormal,

$$\bar{a} = \tau e^{\sigma^2}, \tag{2.8}$$

showing that the ratio of the mean age to the turnover time increases exponentially with the variance of $\ln k$. Taking $\tau = 14$ yr and $\bar{a} = 5000$ yr, we obtain $\sigma = 2.4$, roughly consistent with that found for plant matter decay. Although the assumption of lognormality remains unjustified, this calculation makes clear that, if indeed equation (2.1) applies, long-tailed rate distributions can produce huge disparities between the mean age of a reservoir and its mean lifetime. In this respect, it may be useful to note that if decay is characterized by a single rate k_0 with no dispersion, then $\bar{a} = \tau = k_0^{-1}$ [11]. Discrepancies between age and turnover time are therefore a quantitative signature of the kinetic complexity of the carbon cycle.

[1] For an alternative derivation using the *von Foerster equation* of population dynamics [48], see Ref. [21].

2.3. **Slow time scales, in sediments.** The 1% of marine primary production that survives decay long enough to settle on the seafloor may nevertheless be degraded once it is buried within sediment. How that process evolves is a subject of much discussion and debate [32]. There is, however, some observational certainty: the longer a parcel of organic matter survives, the slower its decay becomes. Remarkably, this *aging* process is characterized by an empirical scaling law.

The scaling law concerns the evolution of the effective first-order rate constant

$$k_{\text{eff}} = -\frac{\mathrm{d}\ln g}{\mathrm{d}t}. \tag{2.9}$$

Over time scales ranging from days to millions of years, observations suggest that [43, 44]

$$k_{\text{eff}} \propto t^{-1}. \tag{2.10}$$

Perhaps the simplest way of understanding this observation is to appeal again to multiplicative processes and lognormal disordered kinetics. We approximate the lognormal as a uniform distribution in log space, such that $\rho(\ln k)$ is constant and $p(k) \propto 1/k$ between minimum and maximum rates k_{min} and k_{max}, respectively. This "log-uniform" approximation is valid when $(\ln k - \mu)/2\sigma^2 \ll 1$ [46]. Inserting it into (2.1), we find

$$g(t)/A \sim E_1(k_{\text{min}}t) - E_1(k_{\text{max}}t), \tag{2.11}$$

where $E_1(x) = \int_1^\infty k^{-1}e^{-kt}\mathrm{d}k$ is the exponential integral [6] and A is a constant. When $k_{\text{max}} \gg k_{\text{min}}$, the second exponential integral can be neglected at long times $t \gg 1/k_{\text{max}}$. An asymptotic expansion [6] of the first integral for $t \ll k_{\text{min}}^{-1}$ then yields [56]

$$g(k_{\text{min}}t) \sim -\gamma - \ln k_{\text{min}}t, \qquad k_{\text{max}}^{-1} < t \ll k_{\text{min}}^{-1} \tag{2.12}$$

where $\gamma = 0.5772\ldots$ is Euler's constant. Equation (2.12) fits much observational data [56]. Inserting it into (2.9), we reproduce the $1/t$ scaling of equation (2.10) with a weak logarithmic correction [56].

2.4. **Immobilization.** As already indicated in the introduction, a small fraction of organic matter—about 0.1%—escapes decay. Informally, one says it is *buried*; more accurately, it is *immobilized* deep within sediment where it can no longer be degraded. Such a slow process would naively seem to be insignificant. Yet, as stated above, this small leak out of the carbon cycle is directly responsible for the oxygenation of the atmosphere [33, 8, 16, 30].

Astonishingly, observational data suggest the existence of another scaling law. The quantity of interest is the *immobilized fraction*

$$\phi = \frac{\text{mass of organic carbon immobilized}}{\text{mass of organic carbon entering sediments}}, \tag{2.13}$$

also called the "burial efficiency" [9]. Typically, $\phi \sim 0.1$. There is, however, significant natural variability.

One cause of variability relates to the existence of oxygen in the sedimentary environment. Most seafloor environments are oxic. However O_2 penetrates only to shallow depths (e.g., a few cm beneath the sediment-water interface) because it is consumed by microbes feeding on detritus. Once O_2 is depleted, anaerobic microbial communities continue to consume the remaining organic matter in deep sediments. However, there is growing evidence that some kinds of organic matter—or organic

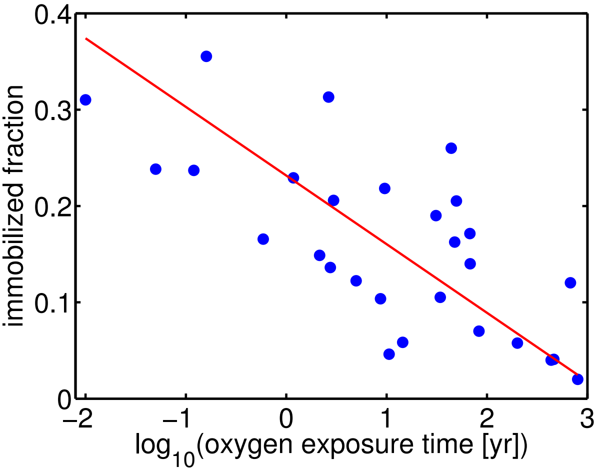


Figure 3. The immobilization fraction ϕ as a function of the logarithm of the oxygen exposure time t_{ox} [28], compared to the best-fitting straight line. The measurements are obtained from analyses of ocean-bottom sediments. Further evidence of the logarithmic time dependence can be found in Ref. [24].

matter that is physically protected by its association with mineral surfaces [38]—can be degraded only in the presence of O_2 [28]. This observation may derive from a unique ability of aerobic micro-organisms to manufacture the hydrolytic enzymes and/or other chemical agents required for degradation of otherwise inert organic matter.

Figure 3 supports this picture. Here the immobilized fraction ϕ is plotted as a function of the logarithm of the *oxygen exposure time* t_{ox} [28]. The time t_{ox} is calculated by dividing the depth (beneath the seafloor) of O_2 penetration by the rate at which sediment is deposited on the seafloor. A remarkably simple result appears to emerge: ϕ decreases linearly with $\log t_{\mathrm{ox}}$ [28].

This observation can be understood in terms of the following reaction-diffusion problem [65, 56]. Aerobic microbes emit a constant flux of enzymes that diffuse away. The enzymes hydrolyze organic matter upon contact, but eventually enzymes become inactive. Under the assumptions that the consumption of organic matter is limited by hydrolysis and that hydrolysis is diffusion limited, the steady-state concentration of enzymes is inhomogeneously distributed in space. The probability of finding a particular concentration c of enzymes at a given location is then found to scale approximately like $1/c$ within wide limits. If we assume that the degradation proceeds locally at a rate proportional to c, we derive once again the logarithmic decay of equation (2.12), where k_{min} is related to the characteristic distance between microbes [56].

Note, however, that we have just described decay, not preservation. To understand preservation, we assume, following Hedges and Keil [32], that there are two kinds of organic matter: that which is only degraded aerobically, and that which will always eventually degrade anaerobically. The quantity $\phi(t_{\mathrm{ox}})$ therefore represents the decay of the portion that requires the oxic environment, as a function of the time t_{ox} of exposure to that environment. We therefore find, as with (2.12),

$$\phi(t_{\text{ox}}) \simeq \text{const.} - \log t_{\text{ox}}, \tag{2.14}$$

consistent with Figure 3.

2.5. **Logarithmic time.** The decomposition models discussed above are essentially null hypotheses. Viewed as a set, they suggest that the carbon cycle's back reaction—respiration—follows the rhythm of a logarithmic clock. That is, equal amounts of organic carbon are converted to equal amounts of CO_2 during equal amounts of the logarithm of time since the organic carbon was produced. In this way, local microbial metabolisms acting on biological time scales exert an influence globally at geologic time scales. The fast and slow processes of the carbon cycle then appear as two ends of a continuum, in some ways more alike than they are different.

3. Dynamics

The analyses of Section 2 provide a basis for understanding and specifying rates of global respiration. Such rates are required for parameterizing models of the form [23]

$$\dot{x} = j_i - j_o, \tag{3.1}$$

where, for example, x represents the mass of CO_2 in the atmosphere, and j_i and j_o represent fluxes of CO_2 into the atmosphere (e.g., respiration and fossil fuel burning) and fluxes out (e.g., primary production and absorption into the oceans), respectively. A basic problem concerns feedbacks; i.e., the dependence of the fluxes on x. For example, if the climate becomes warmer, frozen ground in the arctic will thaw, exposing ancient organic matter to decomposition [69]. Consequently CO_2 levels would rise, leading to more warming, etc. On the other hand, the same processes lead to more primary production due to the release of nutrients, causing more carbon to be stored in, say, new forest growth. Which flux will dominate?

Answering such questions is crucially important for understanding the carbon cycle's dynamics, including the identification of steady states, the specification of rates of relaxation towards those steady states, and the analysis of stability. Here we illustrate these issues by considering examples of dynamical change in the modern and ancient carbon cycle.

3.1. **Impulse response.** The rising CO_2 levels in Figure 1 motivate a simple question. If we inject, say, 1 gigaton of carbon in the atmosphere today, how does the system respond to this impulse?

In the 1950s and early 1960s, tests of nuclear weapons produced a significant increase in the concentration of ^{14}C in the atmosphere. The tests were essentially ended by the Nuclear Test Ban Treaty in October, 1963. Figure 4 shows that the ^{14}C content of the atmosphere then decayed exponentially, like $e^{-\omega\tau}$, where the time constant $\omega^{-1} \simeq 17$ yr.

Although ^{14}C is an unstable isotope of carbon, its half-life, about 5730 yr, is far too large to account for the exponential decay. The exponential decay instead largely represents the equilibration of the atmospheric ^{14}C concentration with that of the ocean. Using the "thin-film" model of Broecker and Peng [14], one considers the ocean to be in a pre-existing steady state balance between the input of "natural" ^{14}C from the atmosphere and its radiogenic decay to ^{12}C. The so-called *bomb spike*

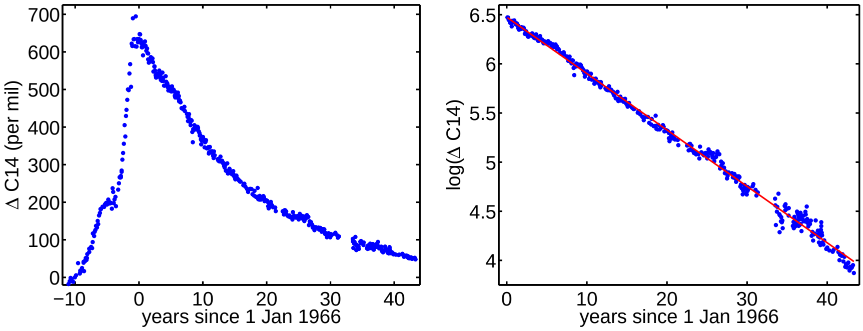


FIGURE 4. Evolution of ^{14}C in the atmosphere [17]. The quantity plotted, Δ^{14}C, is effectively proportional to the atmospheric ^{14}C concentration minus a baseline concentration. *Left*: Δ^{14}C as a function of time. *Right*: $\log(\Delta^{14}\mathrm{C})$ as a function of time, compared to a straight line, showing that the decay is exponential. The slope corresponds to the best-fitting time constant, about 17 yr.

of ^{14}C then diffuses into the ocean through a thin boundary later at the sea surface. Such a simple model predicts not only the exponential decay but it also provides a reasonable estimate of the time constant.

The decay of the bomb spike presents the simplest example of a global impulse response in the carbon cycle. But it provides only a partial view of the problem, not only because of the short time scale, but also because ^{14}C is effectively a passive tracer with an inconsequential concentration. The reality is vastly more complicated—and more interesting [1, 3]. As CO_2 is absorbed into the oceans, the oceans become more acidic, which makes the oceans increasingly less able to absorb more CO_2. This results in a new equilibrium state, on a time scale of about 200–2000 yr, characterized by a 20–35% net increase in atmospheric CO_2 levels. At longer time scales, slow geochemical processes act to restore, or nearly restore, the original equilibrium.

3.2. **Stability.** The persistence of elevated CO_2 levels despite absorption into the oceans suggests not only that perturbations need not fully decay, but also that nonlinearities within the carbon cycle could conceivably amplify a perturbation. Indeed, the geologic record contains much evidence of past disturbances in the carbon cycle. Whether these result from internal excitations or external forcings remains a question. We proceed to review some examples.

3.2.1. *Glacial cycles.* Figure 5 displays one of the most fascinating—and enigmatic—datasets in climate science: the fluctuations of atmospheric CO_2 concentrations and surface temperature in Antarctica for the last 420,000 years, obtained from the Vostok ice core [50]. Two features immediately capture one's attention: the ∼100,000 yr periodicity of the fluctuations, and the extraordinarily tight correlation of the two signals.

The periodicity of the glacial cycles is widely considered to be due to variations in the Earth's orbital parameters [47]. (For an alternative view, see Ref. [68].) The eccentricity of Earth's orbit varies slightly, between about 0 and 0.05, with a period of about 100 Kyr. Also, the Earth's spin axis precesses with a period of

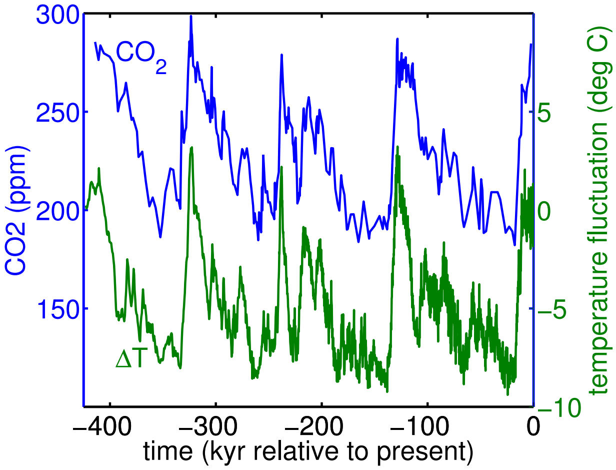


FIGURE 5. Fluctuations of the atmospheric CO_2 concentration (blue upper curve) and temperature (green lower curve) in Antarctica for the last 420,000 years, obtained from analyses of the Vostok ice core [50].

about 26 Kyr, and the tilt, or obliquity, of the spin axis varies with a period of about 41 Kyr. Neither the changes in obliquity nor the changes in precession have any effect on the total annual flux of solar energy received by the Earth. However changes in eccentricity change the annually averaged distance from the sun; consequently eccentricity directly influences annually averaged insolation (the solar radiation energy received per unit area). One would therefore be tempted to immediately ascribe the 100 Kyr periodicity of the temperature record in Figure 5 to the eccentricity variations. However the perturbation of insolation due to changes in eccentricity varies to leading order like eccentricity squared [7]; the resulting small change cannot by itself account for the climatic shifts. Many treatments of the problem, beginning with the original hypothesis of Milanković [45], suggest that the controlling variable is not global insolation but instead the insolation in the northern hemisphere, where most glacial ice is located. For example, Ref. [19] demonstrates an impressive correlation between insolation at summer solstice (derived from obliquity and eccentricity variations) and the derivative of the global ice volume (inferred from geochemical measurements).

Whatever the role of the orbital changes in the glacial cycles, there remains the question of what drives the synchronous changes in atmospheric CO_2 concentrations [4, 62]. Because there is no reason to imagine that the orbital changes directly affect CO_2, it would appear that CO_2 levels are in some sense responding to the climatic changes. For example, because CO_2 is more soluble in colder water, the oceans must absorb more CO_2 during glacial climates. Correspondingly, atmospheric CO_2 concentrations must decline, by about 30 ppm [62], which is only about one-third of that seen in Figure 5. On the other hand, we know that declining CO_2 levels should result in cooler climates [51]; thus it seems likely that the changes in the carbon cycle and climate are amplifying each other to produce the glacial cycles of Figure 5. Consequently the carbon cycle, when suitably forced, contains within

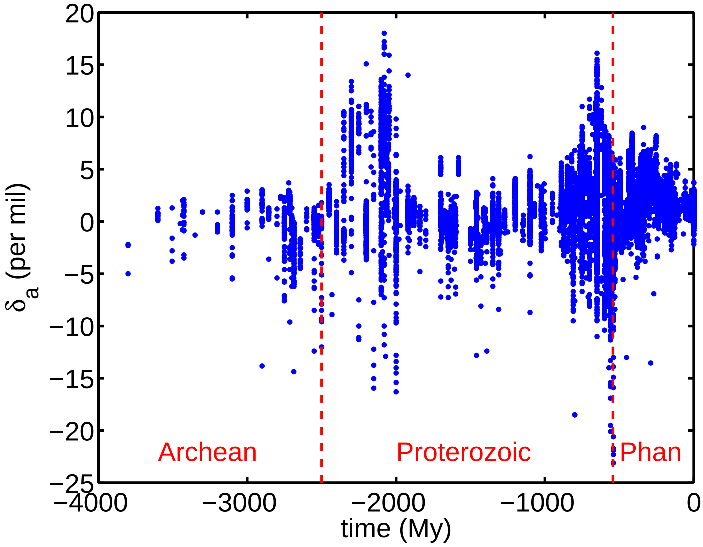


FIGURE 6. Evolution of the carbon isotopic composition of carbonate rocks for the last 3.8 billion years [60]. The red dashed lines demarcate the Archean, Proterozoic, and Phanerozoic eons. The large fluctuations at the beginning and end of the Proterozoic may be associated with the early and late stages, respectively, of the oxygenation of the atmosphere. The later set of fluctuations immediately precedes the evolution of complex, multicellular life.

it mechanisms for losing the stability of its steady state at time scales of 10^4–10^5 yr [2]. Moreover the extraordinary synchronization and apparent periodicity of the two signals in Figure 5 suggest that whatever the feedback mechanisms are, they are simple enough to be reliably repeatable.

3.2.2. *Long-term evolution.* Figure 6 is a record of the carbon cycle's fluctuations at the longest possible time scales, from 3.8 billion years ago (approximately the time of the origin of life [58, 12]) to present. This record contains evidence of both stability and instability. To see why, we first digress to explain the quantity plotted.

Carbon occurs as two stable isotopes, ^{12}C and ^{13}C. However the relative partitioning of ^{12}C and ^{13}C in different global "reservoirs" need not be constant. In particular, the production of organic carbon from CO_2 by photosynthesis slightly favors the lighter isotope, so that organic carbon contains a smaller fraction of ^{13}C than the CO_2 from which it was produced [29]. In Section 2.4 we discussed the long-term immobilization of organic carbon in rocks. Inorganic carbon is also sequestered in rock, as carbonate. If the carbon cycle is in a steady state where the average isotopic composition of carbon in the earth's surface environment is constant, the depletion of ^{13}C in organic carbon requires that inorganic carbon be relatively enriched. Geochemists measure these quantities in terms of the departure of the abundance ratio $R_x = (^{13}\mathrm{C}/^{12}\mathrm{C})_x$ for carbon of type x. Such isotopic data is then reported in terms of the departure of this ratio from a standard ratio R_{std}

as the quantity

$$\delta_x = \frac{R_x - R_{\text{std}}}{R_{\text{std}}} \times 1000, \tag{3.2}$$

where multiplication by 1000 means that the units are parts per thousand, or *per mil.* Figure 6 is a plot of this quantity over geologic time, for carbonate (inorganic) carbon, signified by δ_a. That the average value of δ_a is near zero reflects the choice of R_{std}.

There is also an analogous measure for organic carbon, δ_o (not shown), which for present purposes may be taken to be less than δ_a by a constant. Because the relative abundance of ^{13}C is much smaller than that of ^{12}C, the abundance of ^{12}C approximates the abundance of all carbon, so that the average isotopic composition $\bar{\delta}$ of all carbon can be expressed as

$$\bar{\delta} = (1-f)\delta_a + f\delta_o, \qquad 0 \leq f \leq 1, \tag{3.3}$$

where f is the fraction of buried carbon that is organic.

With these details out of the way, we can now arrive at two simple yet profound interpretations of Figure 6. The first is to identify the characteristic value of δ_a with a characteristic value of the organic burial fraction f. One finds[2] $f \simeq 0.2$ [29]; thus for every five moles of carbon that are sequestered as rock, one is organic and four are inorganic. There has been a tendency to regress to this 1:4 ratio—the essential stoichiometry of the carbon cycle— ever since the origin of life. Neither the value of the ratio nor its long-term stability have been explained.

Somehow the stoichiometry has remained approximately constant despite tremendous changes in both the biosphere and geosphere. Among the most important of those changes is the oxygenation of the atmosphere and oceans. Various lines of evidence suggest that initial rise of oxygen occurred around 2.4 billion years ago [16]. Because larger, multicellular, organisms with aerobic metabolisms evolved nearly two billion years later [39], it is thought that a second rise in oxygen occurred around 600 million years ago [49]. These two periods bookend the *Proterozoic* eon in Figure 6. Because the carbon cycle is intimately connected to the accumulation of oxygen via the immobilization of organic matter (Sections 1 and 2.4), it is tempting to identify the large fluctuations in δ_a at the beginning and end of the Proterozoic with periods of instability. Dynamical systems passing through a bifurcation may be expected to exhibit large fluctuations [41]. Possibly one or both periods of relative instability in Figure 6 represent such changes [57].

This brings us to a conundrum. The logarithmic relation between immobilization and oxygen-exposure time (Figure 3) can be interpreted as a negative feedback for O_2. That is, as O_2 levels increase, so too does oxygen exposure time, and thus O_2 "production" (proportional to the immobilized fraction) decreases [28]. How then, could O_2 levels have increased from a state that was already stable? Possibly, the early rise in O_2 was controlled by different mechanisms [30], one of which has been described as a saddle-node bifurcation [25]. The late-Proterozoic rise, however, almost surely has some relation to carbon burial [16, 49], suggesting that the dependence of burial on O_2 may not be as simple as that depicted by Figure 3. Alternatively, the controls on O_2 production (i.e., burial) may have changed over

[2] Typically, $\bar{\delta} = -6$ per mil and $\delta_a \simeq 0$ per mil. The fraction $f \simeq 0.2$ results from assuming that δ_o is about 30 per mil smaller than δ_a.

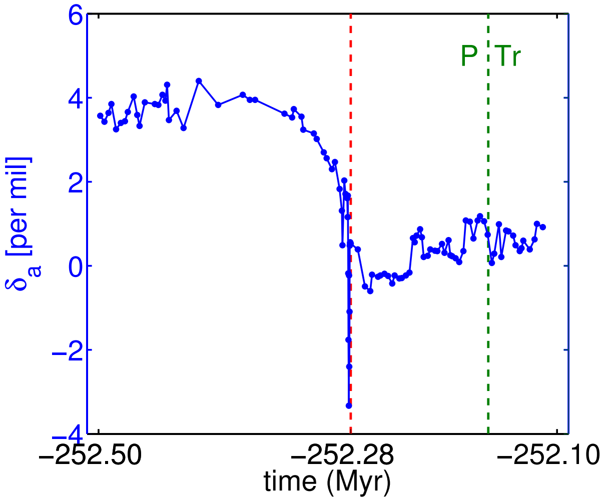


FIGURE 7. Evolution of the carbon isotopic composition of late-Permian carbonate rocks at Meishan, China [15], with dates from Ref. [59]. The peak extinction activity occurs 252.28 million years ago (red dashed line). The later line separates the Permian (P) and Triassic (Tr) periods. The rapid downward acceleration preceding the extinction suggests a form similar to a singular blow-up, and thus a nonlinear instability in the carbon cycle [55].

geologic time. If so, one is left wondering why the burial fraction f has essentially remained constant since the origin of life.

3.2.3. *Mass extinctions.* Our final example concerns mass extinctions—those periods of Earth history where a large fraction of marine animal species went extinct [53]. The most severe extinction occurred about 252.28 million years ago [59] at the end of the Permian [20]. In common with all major events in the history of life [67], the geochemical record suggests that the end-Permian extinction was accompanied by a major perturbation of the carbon cycle [40].

A recent high-resolution record is depicted in Figure 7. The accelerating decline of δ_a occurs within less than ten thousand years [59] as the time of the extinction approaches. The cause of this geologically rapid change remains mysterious, but it is widely viewed to be related to the injection of isotopically light carbon into the oceans and atmosphere [20, 40, 59]. Consequently the downward trend leading to the extinction represents increasing CO_2 levels [59]. The sharp increase in CO_2 appears to be faster than exponential, and similar to a singular blow-up scaling like $1/(t_c - t)$, where t_c is the time of the extinction [55]. This observation suggests that the carbon cycle contains within it sufficient complexity to undergo a nonlinear instability. That the instability is associated with the greatest extinction in Earth history suggests that the rising CO_2 levels of Figure 1 may affect far more than our climate.

4. Conclusion

Much attention in climate science is focused on the response of climate to changes in CO_2 levels. A major unknown, however, is the biosphere's response. Does it amplify changes in CO_2, damp them out, or remain neutral? And at what time scale? These questions address the need to develop a deeper understanding of the carbon cycle, from which we may then better predict the consequences of the trend in Figure 1.

To a considerable extent, the problem is observational. That is, a better understanding of the carbon cycle will follow from more observations of its past and present behavior. Data, however, require understanding. This review has focused on observational data that have been, or appear likely to be, understood by mathematical theory.

An important lesson emerges: despite the carbon cycle's complexity, it exhibits behavior that is simple enough to comprehend mathematically. Such mathematical understanding then leads to basic insights. For example, the apparent universality of the lognormal rate distribution suggests a new interpretation of plant-matter decay: rather than being a collection of special cases, the decay of leaves becomes a manifestation of general principles. We also learn how to relate kinetics at small scales of space and time to long-term global fluxes.

Much remains to be done. Because the carbon cycle represents the coupling between life and the environment—metabolism at a global scale—its mathematical description inherits the difficulties of biology in addition to physical climate science. Thus theoretical understanding of dynamics, so crucial to advancing knowledge of how the carbon cycle works, remains more qualitative than quantitative. Such problems present scientific opportunities with no shortage of social significance. Mathematics will surely play a central role in future progress.

References

1. D. Archer, *The long thaw*, Princeton University Press, Princeton, N.J., 2009.
2. D. Archer, *The global carbon cycle*, Princeton University Press, Princeton, N. J., 2010.
3. D. Archer, M. Eby, V. Brovkin, A. Ridgwell, L. Cao, U. Mikolajewicz, K. Caldeira, K. Matsumoto, G. Munhoven, A. Montenegro, and K. Tokos, *Atmospheric lifetime of fossil fuel carbon dioxide*, Annual Review of Earth and Planetary Sciences **37** (2009), 117–134 (English).
4. D. Archer, A. Winguth, D. Lea, and N. Mahowald, *What caused the glacial/interglacial atmospheric pCO_2 cycles?*, Reviews of Geophysics **38** (2000), no. 2, 159–189.
5. J. Atchison and J. A. C. Brown, *The lognormal distribution*, Cambridge University Press, 1963.
6. C. M. Bender and S. A. Orszag, *Advanced mathematical methods for scientists and engineers*, McGraw Hill Book Company, New York, 1978.
7. A. Berger and M. F. Loutre, *Precession, eccentricity, obliquity, insolation, and paleoclimates*, Long-Term Climatic Variations (J.-C Duplessy and M.-T. Spyridakis, eds.), NATO ASI Series, vol. 122, Springer-Verlag, Heidelberg, 1994, pp. 107–151.
8. R. A. Berner, *The Phanerozoic carbon cycle: CO_2 and O_2*, Oxford University Press, New York, 2004.
9. J. N. Betts and H. D. Holland, *The oxygen content of ocean bottom waters, the burial efficiency of organic carbon, and the regulation of atmospheric oxygen*, Palaeogeography, Palaeoclimatology, Palaeoecology (Global and Planetary Change Section) **97** (1991), 5–18.
10. T.A. Boden, G. Marland, and R.J. Andres, *Global, regional, and national fossil-fuel CO_2 emissions*, Tech. report, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, TN, 2013, doi 10.3334/CDIAC/00001_V2013.

11. B. Bolin and H. Rodhe, *A note on the concepts of age distribution and transit time in natural reservoirs*, Tellus **25** (1973), 58–62.
12. T. Bosak, A. H. Knoll, and A. P. Petroff, *The meaning of stromatolites*, Annual Review of Earth and Planetary Sciences **41** (2012), 21–44.
13. B. P. Boudreau and B. R. Ruddick, *On a reactive continuum representation of organic matter diagenesis*, American Journal of Science **291** (1991), 507–538.
14. W.S. Broecker and T.H. Peng, *Gas-exchange rates between air and sea*, Tellus **26** (1974), no. 1-2, 21–35 (English).
15. C. Cao, G. D. Love, L. E. Hays, W. Wang, S. Shen, and R. E. Summons, *Biogeochemical evidence for euxinic oceans and ecological disturbance presaging the end-Permian mass extinction event*, Earth and Planetary Science Letters **281** (2009), no. 3-4, 188 – 201.
16. D.C. Catling and M.W. Claire, *How Earth's atmosphere evolved to an oxic state: a status report*, Earth and Planetary Science Letters **237** (2005), no. 1-2, 1–20.
17. K. I. Currie, G. Brailsford, S. Nichol, A. Gomez, R. Sparks, K. R. Lassey, and K. Riedel, *Tropospheric $^{14}CO_2$ at Wellington, New Zealand: the worlds longest record*, Biogeochemistry (2009), 5–22.
18. E. R. M. Druffel, P. M. Williams, J. E. Bauer, and J. R. Ertel, *Cycling of dissolved and particularte organic matter in the open ocean*, Journal of Geophysical Research **97** (1992), 15639–15659.
19. S. Edvardsson, K.G. Karlsson, and M. Engholm, *Accurate spin axes and solar system dynamics: Climatic variations for the earth and mars*, Astronomy and Astrophysics **384** (2002), no. 2, 689–701.
20. D. H. Erwin, *Extinction: How life on Earth nearly ended 250 million years ago*, Princeton University Press, Princeton, N.J., 2006.
21. C. L. Follett and D. H. Rothman, *Heterogeneity and its effect on the cycling of marine organic matter*, in preparation.
22. D. C. Forney and D. H. Rothman, *Common structure in the heterogeneity of plant-matter decay*, Journal of The Royal Society Interface **9** (2012), no. 74, 2255–2267.
23. P. Friedlingstein, P. Cox, R. Betts, L. Bopp, W. Von Bloh, V. Brovkin, P. Cadule, S. Doney, M. Eby, I. Fung, et al., *Climate-carbon cycle feedback analysis: Results from the C4MIP model intercomparison*, Journal of Climate **19** (2006), no. 14, 3337–3353.
24. Y. Gelinas, J. A. Baldock, and J. I. Hedges, *Organic carbon composition of marine sediments: effect of oxygen exposure on oil generation potential*, Science **294** (2001), 145–148.
25. C. Goldblatt, T. M. Lenton, and A. J. Watson, *Bistability of atmospheric oxygen and the Great Oxidation*, Nature **443** (2006), 683–686.
26. D. A. Hansell, C. A. Carlson, D. J. Repeta, and R. Schlitzer, *Dissolved organic matter in the ocean: A controversy stimulates new insights*, Oceanography **22** (2009), 202–211.
27. M. E. Harmon, W. L. Silver, B. Fasth, H. Chen, I. C. Burke, W. J. Parton, S. C. Hart, and W. S. Currie, *Long-term patterns of mass loss during the decomposition of leaf and fine root litter: an intersite comparison*, Global Change Biology **15** (2009), no. 5, 1320–1338.
28. H. E. Hartnett, R. G. Keil, J. I. Hedges, and A. H. Devol, *Influence of oxygen exposure time on organic carbon preservation in continental margin sediments*, Nature **391** (1998), 572–574.
29. J. M. Hayes, H. Strauss, and A. J. Kaufman, *The abundance of ^{13}C in marine organic matter and isotopic fractionation in the global biogeochemical cycle of carbon during the past 800 Ma*, Chemical Geology **161** (1999), 103–125.
30. J.M. Hayes and J.R. Waldbauer, *The carbon cycle and associated redox processes through time*, Philosophical Transactions of the Royal Society B: Biological Sciences **361** (2006), no. 1470, 931–950.
31. J. I Hedges, *Global biogeochemical cycles: progress and problems*, Marine Chemistry **39** (1992), 67–93.
32. J. I. Hedges and R. G. Keil, *Sedimentary organic matter preservation: an assessment and speculative synthesis*, Marine Chemistry **49** (1995), 81–115.
33. H. D. Holland, *The chemistry of the atmosphere and oceans*, John Wiley & Sons, New York, 1978.
34. W. T. Holser, M. Schidlowski, F. T. Mackenzie, and J. B. Maynard, *Biogeochemical cycles of carbon and sulfur*, Chemical Cycles in the Evolution of the Earth (C. B. Gregor, R. M. Garrels, F. T. Mackenzie, and J. B. Maynard, eds.), John Wiley & Sons, New York, 1988, pp. 105–173.

35. IPCC, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, http://www.ipcc.ch/, 2013.
36. L. P. Kadanoff, *Statistical physics: statics, dynamics and renormalization*, World Scientific, 2000.
37. C. D. Keeling, *The Suess effect: ^{13}Carbon-^{14}Carbon interrelations*, Environment International **2** (1979), no. 4, 229–300.
38. R. G. Keil and L. M. Mayer, *Mineral matrices and organic matter*, Treatise in Geochemistry, 2nd edition, Elsevier, 2013, p. in press.
39. A. H. Knoll, *Life on a young planet*, Princeton University Press, Princeton, 2003.
40. C. Korte and H. W. Kozur, *Carbon-isotope stratigraphy across the permian-triassic boundary: A review*, Journal of Asian Earth Sciences **39** (2010), 215–235.
41. C. Kuehn, *A mathematical framework for critical transitions: Bifurcations, fast–slow systems and stochastic dynamics*, Physica D: Nonlinear Phenomena **240** (2011), no. 12, 1020–1035.
42. C. M. Marshall, *Explaining the Cambrian "explosion" of animals*, Annu. Rev. Earth Planet. Sci. **34** (2006), 355–284.
43. J. J. Middelburg, *A simple rate model for organic matter decomposition in marine sediments*, Geochimica et Cosmochimica Acta **53** (1989), 1577–1581.
44. J. J. Middelburg and F. J. R. Meysman, *Burial at sea*, Science (Washington) **316** (2007), no. 5829, 1325–1326.
45. M. Milankovič, *Canon of insolation and the ice-age problem*, Zavod za udžbenike i nastavna sredstva, 1998.
46. E. W. Montroll and M. F. Shlesinger, *On $1/f$ noise and other distributions with long tails*, Proc. Natl. Acad. Sci. USA **79** (1982), 3380–3383.
47. R. A. Muller and G. J. Macdonald, *Ice ages and astronomical causes*, Springer, New York, 2000.
48. J. D. Murray, *Mathematical biology, second, corrected, edition*, Springer, 1993.
49. L. Och and G.A. Shields, *The Neoproterozoic oxygenation event: Environmental perturbations and biogeochemical cycling*, Earth-Science Reviews **110** (2012), 26–57.
50. JR Petit, J Jouzel, D Raynaud, NI Barkov, JM Barnola, I Basile, M Bender, J Chappellaz, M Davis, G Delaygue, M Delmotte, VM Kotlyakov, M Legrand, VY Lipenkov, C Lorius, L Pepin, C Ritz, E Saltzman, and M Stievenard, *Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica*, Nature **399** (1999), no. 6735, 429–436.
51. R. T. Pierrehumbert, *Principles of planetary climate*, Cambridge University Press, 2010.
52. A. Plonka, *Dispersive kinetics*, Kluwer, Boston, 2001.
53. D. M. Raup and J. J. Sepkoski Jr, *Mass extinctions in the marine fossil record*, Science **215** (1982), no. 4539, 1501–1503.
54. J. Ross and M. O. Vlad, *Nonlinear kinetics and new approaches to complex reaction mechanisms*, Annual Review of Physical Chemistry **50** (1999), 51–78.
55. D. H. Rothman, *Singular blow-up in the end-Permian carbon cycle*, Abstract B53B-08 presented at 2010 Fall Meeting, American Geophysical Union, San Francisco, Calif., 13-17 Dec, 2010.
56. D. H. Rothman and D. C. Forney, *Physical model for the decay and preservation of marine organic carbon*, Science **316** (2007), 1325–1328.
57. D. H. Rothman, J. M. Hayes, and R. E. Summons, *Dynamics of the Neoproterozoic carbon cycle*, Proceedings of the National Academy of Sciences USA **100** (2003), 8124–8129.
58. M. Schidlowski, *Carbon isotopes as biogechemical recorders of life over 3.8 Ga of Earth history: evolution of a concept*, Precambrian Research **106** (2001), 117–134.
59. S.-Z. Shen, J. L. Crowley, Y. Wang, S. A. Bowring, D. H. Erwin, P. M. Sadler, C.-Q Cao, D. H. Rothman, C. M. Henderson, J. Ramezani, H. Zhang, Y. Shen, X.-D. Wang, W. Wang, Lin Mu, W.-Z. Li, Y.-G. Tang, X.-L. Liu, L.-J Liu, Y. Zeng, Y.-F Jiang, and Y.-G. Jin, *Calibrating the end-Permian mass extinction*, Science **334** (2011), 1367–1372.
60. G. Shields and J. Veizer, *Precambrian marine carbonate isotope database: Version 1.1*, Geochemistry Geophysics Geosystems **3** (2002), Art. No. 1031.
61. W. Shockley, *On the statistics of individual variations of productivity in research laboratories*, Proceedings of the IRE **45** (1957), no. 3, 279–290.
62. D. M. Sigman and E. A. Boyle, *Glacial/interglacial variations in atmospheric carbon dioxide*, Nature **407** (2000), no. 6806, 859–869.

63. H. E. Suess, *Radiocarbon concentration in modern wood*, Science **122** (1955), no. 3166, 415–417.
64. P. Tans and R. Keeling, October 2013, NOAA/ESRL (www.esrl.noaa.gov/gmd/ccgg/trends/) and Scripps Institution of Oceanography (scrippsco2.ucsd.edu/).
65. Y.A. Vetter, J.W. Deming, P.A. Jumars, and B.B. Krieger-Brockett, *A predictive model of bacterial foraging by means of freely released extracellular enzymes*, Microbial Ecology **36** (1998), 75–92.
66. M. O. Vlad, D. L. Huber, and J. Ross, *Rate statistics and thermodynamic analogies for relaxation processes in systems with static disorder: Application to stretched exponential*, J. Chem. Phys. **106** (1997), 4157–4167.
67. O. H. Walliser (ed.), *Global events and event stratigraphy in the Phanerozoic*, Springer, Berlin, 1996.
68. C. Wunsch, *The spectral description of climate change including the 100 ky energy*, Climate Dynamics **20** (2003), 353–363.
69. S. A. Zimov, E. A. G. Schuur, and F. S. Chapin III, *Permafrost and the global carbon budget*, Science **312** (2006), no. 5780, 1612–1613.

Lorenz Center, Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 U.S.A

E-mail address: dhr@mit.edu

THE GEOMETRY OF OUTER SPACE

KAREN VOGTMANN

Abstract. Outer space is a space of graphs used to study the group $\mathrm{Out}(F_n)$ of outer automorphisms of a finitely-generated free group. We discuss an emerging metric theory for Outer space and some applications to $\mathrm{Out}(F_n)$.

1. Introduction

Outer space was introduced in the early 1980's as a tool for studying the group $\mathrm{Out}(F_n)$ of outer automorphisms of a finitely-generated free group **[15]**. It is a contractible space on which $\mathrm{Out}(F_n)$ acts with finite stabilizers, and should be thought of as analogous to a symmetric space with the action of a non-uniform lattice, or to the Teichmüller space of a surface with the action of the mapping class group of the surface. Outer space also has close connections to other areas of mathematics, including tropical geometry, Kontsevich's graph homology theory and the mathematics of phylogenetic trees.

Outer space is a parameter space for certain metric objects, but historically the space itself and its quotient by $\mathrm{Out}(F_n)$ have been studied mostly by topological and combinatorial methods. These methods have yielded (and continue to yield) a wealth of information about $\mathrm{Out}(F_n)$, including information about its finiteness properties, subgroup structure, algorithmic properties and cohomology. However the determination of whether $\mathrm{Out}(F_n)$ has other basic properties enjoyed by lattices and mapping class groups has eluded these methods. An example particularly relevant to geometric group theory is the question of whether a group which has has (approximately) the same geometry as $\mathrm{Out}(F_n)$ must be (almost) isomorphic to $\mathrm{Out}(F_n)$; the technical term is whether $\mathrm{Out}(F_n)$ is *quasi-isometrically rigid.* In the cases of lattices and mapping class groups many of these elusive features were established using the geometry of symmetric spaces and Teichmüller spaces in essential ways.

Outer space is not a manifold, so some care is required when attempting to use geometric tools to study it. In the past few years a metric theory of Outer space has begun to emerge based on a natural non-symmetric metric. The resulting geometric point of view is yielding new information about $\mathrm{Out}(F_n)$ as well as elegant new proofs and better understanding of older results, and is strengthening the analogy between Outer space and the classical theories of symmetric spaces and Teichmüller spaces.

Many people are now contributing to developing a geometric theory of Outer space, and I have provided a cursory guide to some of the literature in the last section. In this short article I have chosen to sketch primarily work of Yael Algom-Kfir, Mladen Bestvina and Mark Feighn. For the interested reader, a more thorough introduction to their work can be found in Bestvina's lecture notes from his Summer

2010 *Mathematics Subject Classification.* 20F65 .

2012 course at the Park City Mathematics Institute, which will be published by the A.M.S. in the near future.

Acknowledgements. I would like to thank Mladen Bestvina, Mark Feighn, Saul Schleimer and Sam Taylor for very helpful conversations during the preparation of this article.

2. Definitions of Outer space

For the most part we will think of Outer space as a space of *marked metric graphs*, where the metric on a graph is specified by giving each edge a positive real length, and the marking identifies the fundamental group of the graph with the free group F_n. However there are several equivalent ways to define Outer space, and we will begin with the one which is quickest to state. This is in terms of actions of F_n on *metric simplicial trees*, i.e. contractible one-dimensional simplicial complexes metrized so that each edge is isometric to an interval of the real line. An action is called *minimal* if it has no invariant subtree. Using this concept we can give a succinct definition of Outer space, though we will temporarily duck the question of what topology to put on it.

Definition 2.1. *(Unprojectivized) Outer space in rank n* is the space of free minimal actions of F_n by isometries on metric simplicial trees.

Here two actions are considered the same if there is an equivariant isometry between the trees. The notation cv_n has become standard for this space. The projectivized version is denoted CV_n, i.e. in CV_n two actions are considered equivalent if they differ only by scaling the tree's metric. Instead of taking equivalence classes one can also think of a point in CV_n as normalized so that the “volume” of the quotient graph (i.e. the sum of its edge lengths) is one. In most of the remainder of this article this is the convention we will adopt.

The action of $\text{Out}(F_n)$ on cv_n or CV_n is also easily described using this definition. Given a point $\rho\colon F_n \to \text{Isom}(T)$ and an automorphism ϕ, we get a new point by composing $\rho \circ \phi\colon F_n \to \text{Isom}(T)$. This descends to an action of $\text{Out}(F_n)$ since inner automorphisms give equivariantly isometric actions, i.e. they fix all of cv_n. Note that this is a *right* action, and does not affect the metric on the tree.

2.1. **Definition in terms of graphs.**

Here's how to translate the definition of CV_n given above into terms of graphs. The graph corresponding to an action of F_n on a tree T is the quotient of T by the action. The fact that we are only considering minimal actions implies that these quotient graphs are compact and have no univalent or bivalent vertices. If we choose a basis for F_n and a basepoint for T, then arcs joining the basepoint to its images under the generators of F_n descend to an immersion of a rose R_n into the quotient graph Γ; this map $g\colon R_n \to \Gamma$ is a *marking* which identifies $\pi_1(R_n) \equiv F_n$ with $\pi_1(\Gamma)$. The fact that equivariantly isometric trees are the same point in CV_n can be stated in terms of marked graphs by defining (g, Γ) and (g', Γ') to be equivalent if there is an isometry $f\colon \Gamma \to \Gamma'$ with $f \circ g$ homotopic to g'. Since we do not want this definition to depend on the choice of basepoint, we do not require that this homotopy preserve basepoints.

An element $\phi \in \text{Out}(F_n)$ acts by changing the marking, i.e. if we represent ϕ by a homotopy equivalence $f\colon R_n \to R_n$, then $(g, \Gamma) \cdot \phi = (g \circ f, \Gamma)$.

2.2. **Simplicial decomposition and topology.** The description of CV_n in terms of graphs makes it easy to decompose CV_n into a union of open simplices. The simplex containing (g, Γ) is formed by assigning all possible positive edge lengths to the edges of Γ, subject to the condition that the volume must be one. The simplex containing (g', Γ') is a face of the one containing (g, Γ) if (g', Γ') can be obtained from (g, Γ) by collapsing some edges of Γ to points; the face of the simplex is obtained by assigning those edges zero length. Some faces of each simplex are missing, since you cannot collapse a loop without changing the rank of the fundamental group. If we formally add all of the "missing faces" we obtain a simplicial complex called the *simplicial completion* of Outer space. The simplices which are not in CV_n are said to be *at infinity.* (we remark that the term "simplicial completion" is used slightly differently in **[1]**, though the paper proves that her notion agrees with the one used here).

There are several natural ways of topologizing CV_n, but they are all equivalent to giving it the quotient topology which arises from its description as the disjoint union of open simplices modulo the above face relations.

2.3. **Definition in terms of sphere systems.** Outer space CV_n can also be described using spheres embedded in a doubled handlebody M_n (i.e. M_n is the connected sum of n copies of $S^1 \times S^2$). Since $\pi_1(M_n) = F_n$, any diffeomorphism of M induces an (outer) automorphism of F_n. Laudenbach showed that the kernel of the resulting map from $\pi_0(Diff(M_n))$ to $Out(F_n)$ is a finite 2-group which acts trivially on isotopy classes of embedded spheres. Thus one may build a simplicial complex with an $\mathrm{Out}(F_n)$-action whose vertices are isotopy classes of embedded spheres. A set of k such vertices spans a k-simplex if spheres representing the vertices can be embedded disjointly into M_n; this is called the *sphere complex* $\mathcal{S}_n$.

Remark 2.2. Replacing "a doubled handlebody M_n" by "a surface S" and "embedded spheres" by "simple closed curves" transforms the above definition into the usual definition of the curve complex $\mathcal{C}(S)$ associated to a surface.

Hatcher shows that CV_n embeds into the the sphere complex $\mathcal{S}_n$ as a union of open simplices, corresponding to sphere systems with simply-connected complementary pieces **[22]**. The marked graph corresponding to a sphere system is simply the graph dual to the spheres, with one vertex in each complementary piece and one edge intersecting each sphere. Barycentric coordinates on the simplex corresponding to the sphere system determine edge-lengths for the graph, and the marking comes from the embedding of the graph into M_n. The open simplices coincide with those we described in the last section.

An advantage of this approach is that all points of CV_n (i.e. weighted sphere systems) are contained in the same object (M_n), so there are natural ways of comparing two points such as looking at their intersection.

3. The Lipschitz metric on Outer space

Thurston studied a non-symmetric metric on Teichmüller space which measures the infimum of Lipschitz constants for a homotopy class of homeomorphisms from one hyperbolic surface to another **[33]**. He proved basic properties of this metric, showed one can measure distance by looking at how much curves are stretched, described geodesics, and used the metric to give coordinates for Teichmüller space. Bestvina was perhaps the first to suggest transporting this idea to Outer space, and a nice account of the basic properties of this metric can be found in his 2012 PCMI lecture notes, where he attributes some of these as unpublished results of his former student Tad White. Francaviglia and Martino published the first systematic study of this metric in [**16**].

3.1. **Definition of the metric.** Let $f\colon X \to X'$ be a map of metric spaces. Recall that the *Lipschitz constant* (or, more informally, the *maximal stretch*) of f is

$$L(f) = \sup_{x,y\in X} \frac{d_{X'}(f(x), f(y))}{d_X(x,y)}.$$

If X is compact, then $L(f)$ is always finite, and we may replace sup with max in this definition.

Definition 3.1. The *Lipschitz distance* between two points (g,Γ) and (g',Γ') of CV_n is

$$d((g,\Gamma),(g',\Gamma')) = \inf_f \log(L(f))$$

where the infimum is taken over all $f\colon \Gamma \to \Gamma'$ with $f \circ g \simeq g'$.

Since Γ and Γ' are compact, the Arzela-Ascoli theorem says that the limit of a minimizing sequence exists, so we may replace inf with min in this definition.

We can decrease the Lipschitz constant of a map $f\colon \Gamma \to \Gamma'$ by "tightening" to make the restriction of f to each edge of Γ *a linear immersion*, i.e. any lift $\tilde f\colon \widetilde\Gamma \to \widetilde\Gamma'$ of f to universal covers maps edges of $\widetilde\Gamma$ linearly to arcs in $\widetilde\Gamma'$. Thus for the purposes of computing the Lipschitz distance between two points of CV_n, we need only consider maps with this property. Such a map f is called a *difference of markings* from (g,Γ) to (g',Γ') if $f\circ g \simeq g'$. Note that if two points of Outer space are in the same open simplex, then there is a difference of markings which is combinatorially the identity map.

In the rest of this article we will usually omit the marking when denoting a point of CV_n, referring to (g,Γ) simply as Γ unless it is strictly necessary to specify the marking.

3.2. **Calculating the Lipschitz distance.** Although the definition of Lipschitz distance involves taking the infimum over an infinite family of maps, it turns out that in practice it is very easy to calculate the distance by a finite process using any map in the family. To explain this we first need a few definitions.

If $f\colon \Gamma \to \Gamma'$ minimizes the stretch in its homotopy class, i.e. $L(f) = \inf_{f'\simeq f} L(f')$, then f is called an *optimal* map. Since Γ and Γ' are compact, optimal maps exist by the Arzela-Ascoli theorem. If f is optimal, the *tension subgraph* $\Delta = \Delta(f)$ is the subgraph spanned by edges which are stretched by exactly $L(f)$.

For any graph Γ, a *direction* at a point $x \in \Gamma$ is a germ of geodesic paths starting at x. Thus there are two directions at most points, but the number of directions at a vertex is equal to the valence of the vertex.

Proposition 3.2. [**7, 16**] *Let Γ and Γ' be two points in CV_n, and let $f\colon \Gamma \to \Gamma'$ be an optimal difference of markings with the additional condition that $\Delta = \Delta(f)$ is minimal, i.e. there is no optimal f' homotopic to f such that $\Delta(f')$ is a proper subgraph of Δ. Then*

(1) *Δ is a core graph, i.e. has no univalent vertices.*

(2) Δ *has no vertex v where all directions at v map to the same direction at $f(v)$, and*

(3) $d(\Gamma, \Gamma')$ *is equal to the supremum, over all loops α immersed in Γ, of*

$$\log\left(\frac{\text{length}(f(\alpha))}{\text{length}(\alpha)}\right).$$

It is easy to see that the supremum in statement (3) is realized on some loop which crosses each edge at most twice. With a little more thought one can see that it must be realized on either an embedded loop or an immersed loop which travels once around an embedded figure-eight or an embedded barbell. Such loops are called *candidates*, and we collect these observations in the following corollary.

Corollary 3.3. *The distance $d(\Gamma, \Gamma')$ can be computed from any difference of markings by looking at each loop in the finite collection of "candidate loops" in Γ and measuring the length of the shortest loop in the homotopy class of its image in Γ'.*

3.3. **Peculiarities of the Lipschitz distance.** It is not difficult to show that the Lipschitz distance obeys two of the axioms for a metric. The triangle inequality follows because Lipschitz constants sub-multiply i.e. $L(g \circ f) \leq L(g)L(f)$. The fact that $d(\Gamma, \Gamma') = 0$ if and only if $\Gamma = \Gamma'$ follows because a surjective map between two volume 1 graphs which stretches nothing must be an isometry.

This distance function fails dramatically to be symmetric, however. A simple example is given by two 2-petaled roses ρ_1 and ρ_2 in the same simplex of CV_2. If ρ_1 has edge lengths $\{1/2, 1/2\}$, and ρ_2 has edge-lengths $\{\epsilon, 1-\epsilon\}$ for some small ϵ, then the combinatorial identity map ρ_1 to ρ_2 stretches each edge by at most 2, while any map homotopic to the identity in the opposite direction stretches the loop of length ϵ by at least $1/2\epsilon$. So this distance is not only asymmetric, there is no bound to the difference between $d(x, y)$ and $d(y, x)$. Note that the first rose is at the barycenter of its simplex, while the second is close to a missing face. This is symptomatic of the following general phenomenon:

> *You can get to the edge of Outer space very fast, but it will take you a long, long time to get back.*

4. Using the Lipschitz metric to classify elements of $\text{Out}(F_n)$

The Lipschitz distance is invariant under the action of $\text{Out}(F_n)$ since the distance depends only on the *difference* of markings, i.e. $\text{Out}(F_n)$ acts on CV_n by isometries. In symmetric spaces such as real hyperbolic space, isometries can be divided into three classes: those which are elliptic (which fix a point in the space), hyperbolic (translate a geodesic and act with North-South dynamics on the boundary of the space), and parabolic (neither of the above, but always fix a point on the boundary). Isometries of Teichmüller space exhibit the same three behaviors, and Bers used this fact to give a new proof of Thurston's classification of elements of the mapping class group [**4**]. In particular, hyperbolic mapping classes are *pseudo-Anosov*, meaning they stretch the metric on the surface in one direction and shrink it in a complementary direction at all but finitely many points. This behavior had been codified by Thurston using a pair of complementary *train tracks* for a pseudo-Anosov homeomorphism, which give a powerful tool for working with these mapping classes.

Bestvina and Handel proved the existence of analogous structures on graphs, which they also called train tracks, for *fully irreducible* automorphisms of F_n, i.e.

automorphisms such that no power of the automorphism preserves the conjugacy class of any proper free factor of F_n [**13**]. The proof was combinatorial and quite intricate. Bestvina was motivated to revisit the Lipschitz metric when he realized it could be used to give a conceptual geometric proof of the existence of these train tracks along the lines of Bers' proof [**6**]. Here is a very rough sketch of the proof.

4.1. **The basic trichotomy.** We first divide elements of $\phi \in \mathrm{Out}(F_n)$ into classes by looking at the smallest distance they can move points in CV_n. Specifically, let $D = \inf_{\Gamma \in \mathrm{CV}_n} d(\Gamma, \Gamma\phi)$. Then ϕ is

- *elliptic* if $D = 0$ and is realized
- *hyperbolic* if $D > 0$ and is realized,
- *parabolic* if D is not realized.

Elliptic elements are easiest to understand, since ϕ is elliptic if and only if it fixes a point $\Gamma \in \mathrm{CV}_n$, in which case any difference of markings from Γ to $\Gamma\phi$ is homotopic to an isometry, so has finite order.

If ϕ is parabolic, consider any sequence of optimal maps $f_k \colon \Gamma_k \to \Gamma_k$ with $d(\Gamma_k, \Gamma_k\phi)$ converging to $\inf d(\Gamma, \Gamma\phi)$. Bestvina shows that all but finitely many of these Γ_k have a subgraph whose edges are very small compared to the other edges. Since edges can't get stretched by more than $\lambda = e^D$ by an optimal map, for k large the small subgraph is invariant under the difference of markings, so determines a proper free factor of F_n which is invariant (up to conjugacy) under some power of ϕ. We can conclude that if ϕ is fully irreducible it must be hyperbolic.

The key to understanding hyperbolic isometries is to find a point Γ in CV_n which is moved minimal distance and for which there is a particularly nice difference of markings map from Γ to $\Gamma\phi$. Here "particularly nice" means $f \colon \Gamma \to \Gamma$ is an optimal map, the tension subgraph Δ is mapped into itself and the restriction of f to Δ is an extremely efficient type of map called a *train track map.* To explain what this means, we introduce a little terminology.

- A *train track structure* on a graph Γ is an equivalence relation on the directions at each vertex.
- Equivalence classes of directions at a vertex v are called *gates.*
- A pair of directions at a vertex is called a *turn.*
- A turn is *illegal* if the directions are equivalent, i.e. belong to the same gate, and *legal* otherwise.
- A path in Γ is *legal* if it does not take any illegal turns

A map $f \colon \Delta \to \Delta$ is a *train track map* if there is a train track structure on Δ with the following properties:

- there are at least two gates at every vertex of Δ,
- the image of each edge of Δ is a legal path, and
- the image of each legal turn in Δ is a legal turn.

If ϕ is a hyperbolic automorphism, there is a very elegant proof that such a "particularly nice" difference of markings exists, which we sketch below. The tension graph Δ may well be a proper subgraph of Γ, in which case some power of ϕ is reducible. But if ϕ is fully irreducible, Δ must be all of Γ and, after adjusting f slightly to make vertices go to vertices we have the following theorem, originally proved by Bestvina and Handel:

Theorem 4.1. **[13]** *Let $\phi \in \mathrm{Out}(F_n)$ be fully irreducible. Then there is a graph Γ and train track map $f\colon \Gamma \to \Gamma$ taking vertices to vertices and inducing ϕ on $\pi_1(\Gamma) \cong F_n$.*

This theorem is the basis of a great deal of subsequent work on $\mathrm{Out}(F_n)$. It might be considered the "golden spike" uniting Thurston's theory of train tracks for surface automorphisms **[32]** with Stallings' theory of folding maps for automorphisms of free groups **[30]**.

4.1.1. *Hyperbolic automorphisms and train tracks.* Let ϕ be a hyperbolic automorphism and Γ a point in CV_n realizing $\inf_{\Gamma \in \mathrm{CV}_n} d(\Gamma, \Gamma\phi) = D > 0$. Choose an optimal difference of markings $f\colon \Gamma \to \Gamma$, i.e. a difference of markings with stretch factor $\lambda = e^D > 1$. By Proposition 3.2 we may assume the tension subgraph $\Delta = \Delta(f)$ is a core graph with at least two gates at every vertex. We define the *complexity* of such an f in terms of Δ, namely

$$c(f) = (\text{rank of } \Delta, -\text{number of components of } \Delta).$$

Since Δ has no univalent vertices, removing an edge from Δ either reduces rank or increases the number of connected components, so decreases complexity.

Take an f as above with minimal possible complexity. We want to prove that the restriction of f to Δ is a train track map. We have to show that f takes Δ to Δ, maps each edge to a legal path, and takes legal turns to legal turns.

Step 1. Suppose $f(\Delta) \not\subset \Delta$. Perturb Γ by uniformly expanding Δ while shrinking the complement $\Gamma - \Delta$. Do this a very small amount, so that no edges are added to Δ. Any edge of Δ which was mapped into Δ is still stretched by λ. But an edge of Δ whose image wanders outside of Δ will now be stretched by less than λ, so will disappear from Δ, contradicting minimality of $c(f)$. Note that Δ cannot disappear completely under our deformation: some edge must still be stretched by λ since λ is minimal for maps in the homotopy class of f. Thus $f(\Delta) \subset \Delta$.

Step 2. This step involves the notion of *folding* an illegal turn. If $\{d_1, d_2\}$ is a turn, folding by ϵ means identifying initial segments in the directions d_1 and d_2 of length ϵ, so that the "V" formed by d_1 and d_2 becomes a "Y" with a very short stem. If $\{d_1, d_2\}$ is illegal, then the directions $f(d_1)$ and $f(d_2)$ agree, so that f induces a map on the folded graph if ϵ is small enough. We call the induced map an ϵ-fold of f.

Suppose there is an edge e such that $f(e)$ makes an illegal turn. Fold that turn slightly, but not enough to add edges to Δ. In the folded map, the image of e is homotopic to a shorter loop so e "drops out of Δ" (actually, $c(f)$ decreases), contradicting minimality.

Step 3. Suppose there is a legal turn $\{d_1, d_2\}$ that gets mapped to an illegal turn $\{f(d_1), f(d_2)\}$. Fold $\{f(d_1), f(d_2)\}$ slightly (without adding edges to Δ). Then $\{d_1, d_2\}$ becomes illegal and we don't have to worry about where it gets mapped. But we do have to worry about the fact that formerly legal turns may have become illegal, and we may have decreased the number of gates at other vertices of Δ, maybe down to a single gate at some vertex v. If we've done that, though, we could fix it with a homotopy of f which moves the image of v slightly into the image of its adjacent edges. This makes Δ smaller (decreases the complexity of f), again contradicting minimality.

And that's it.

The nicest ways to get from A to B are those that involve minimal tension.

5. Geodesics in the Lipschitz metric

Although the Lipschitz metric is not symmetric, one could obtain a genuine metric by simply symmetrizing it. Francoviglia and Martino studied basic properties of both the Lipschitz metric and its symmetized version, and found that the unsymmetrized version has a number of advantages over the symmetrized one. Perhaps the greatest of these is that CV_n is geodesically complete in the unsymmetrized version, but is not in the symmetrized one. In fact we can exhibit and analyze specific geodesics between any two points, and understanding the behavior of these geodesics is the key to many applications of the metric.

5.1. **Recognizing geodesics.** A path $\gamma(t)$ is a geodesic if and only if the triangle inequality is an equality for any three points along the path, i.e. for any $t_0 \leq t_1 \leq t_2$,

$$d(\gamma(t_0), \gamma(t_1)) + d(\gamma(t_1), \gamma(t_2)) = d(\gamma(t_0), \gamma(t_2)).$$

In terms of the Lipschitz distance this translates to: If Γ_t is a path in CV_n such that the "same" loop is maximally stretched from each point on the path to any point further along, then Γ_t is a geodesic. Here loops are the same if they represent the same conjugacy class of F_n.

5.2. **Example: straight line in a simplex.** Suppose Γ and Γ' are points in CV_n which are in the same simplex, i.e. they differ only by the lengths of their edges Then we can define a path between them by simply scaling all edge lengths linearly. Lengths of edges are stretched (or shrunk) at a constant rate all along this path, so the same loop is maximally stretched all along the path. Therefore by the criterion stated in the last subsection, the path is a geodesic.

5.3. **Example: folding path.** Let $f\colon \Gamma \to \Gamma'$ be an optimal difference of markings such that the tension subgraph is all of Γ. Then there is a special class of geodesics from Γ to Γ' called *folding paths*. We've already seen tiny folds in the proof of the classification of automorphisms, when we identified initial segments in the directions of an illegal turn.

The map f induces a train track structure on Γ which puts directions at v into the same gate if they map to the same direction at $f(v) \in \Gamma'$ (note that this does *not* mean that f is a train track map!). If f is not an isometry there must be at least one illegal turn, and as before we can identify initial segments in the directions of this illegal turn to get an induced a "folded" map. The folded map is still optimal, and the tension subgraph is still the entire domain, so unless the folded map is an isometry we may continue the path by folding some more. If we do this long enough we will eventually arrive at Γ'. We think of this as a continuous process, and call this a *folding path* from Γ to Γ'.

If the tension subgraph is not all of Γ we may need to change the edge lengths of Γ (contracting the edges in Δ and expanding the others) to make f into an optimal map with $\Delta = \Gamma$. Since this only moves us within a simplex, we may accomplish this by traveling along a straight line, as in the last example. The composition of this straight line with a folding path as above is a geodesic in CV_n, and we will abuse terminology by calling this a folding path as well.

5.4. **So many roads.** There are many geodesics between two given points. If the points are in the same simplex you don't have to take the straight line between them, as long as you keep the same loop maximally stretched while you travel. If the two points are related by an optimal difference of markings whose tension subgraph is the entire graph, you can start by folding at any one of the illegal turns to get a folding path. To get a more canonical geodesic between two points you could take the straight line until the tension subgraph is the whole graph, then take the "greedy" folding path which folds all illegal turns simultaneously at the same rate. But that still doesn't give unique paths, because there can be different optimal maps between two points.

5.5. **Asymmetry of geodesics.** Since there are so many geodesics between points, one might hope that there is some path which is a geodesic in both directions. Coulbois and Weist demolished this hope by a simple example in rank 2 (see [**16**]). Take two marked theta-graphs Γ_1 and Γ_2 in adjacent triangles, with edge lengths $\{1/3, 1/2, 1/6\}$ and $\{1/6, 1/2, 1/3\}$ respectively. Any geodesic from Γ_1 to Γ_2 must contain a rose R in the common face of the triangles, with some edge lengths $(\ell, 1-\ell)$. Since this is a geodesic, we must have $d(G_1, R) + d(R, G_2) = d(G_1, G_2)$. Calculating these distances (using candidate loops) subject to this constraint shows that ℓ must be equal to 5/8. Calculating in the other direction gives $\ell = 3/8$, i.e. no geodesic from Γ_1 to Γ_2 is equal to any geodesic from Γ_2 to Γ_1.

> *There are many ways to travel from here to there, and the way back might avoid them all.*

5.6. **Geodesics in the thick part.** Traversing a geodesic from x to y backwards gives a path which needn't be a geodesic, and in fact may be arbitrarily far from any geodesic from y to x. However, this can't happen if the geodesic stays away from the "thin part" of CV_n. A marked graph Γ is in the *ϵ-thin part* if some embedded loop in Γ has length at most ϵ. Not surprisingly, the complement of the ϵ-thin part is called the *ϵ-thick part.* In a paper which analyzes the asymmetry of the Lipschitz metric quite precisely, Algom-Kfir and Bestvina prove the following statement.

Theorem 5.1 ([**2**]). *If a geodesic γ from x to y stays in the ϵ-thick part of* CV_n*, then the same path traversed backwards is a quasi-geodesic, i.e. stays a uniformly bounded distance from some geodesic, where the bound depends on ϵ.*

6. Hyperbolicity

Gromov introduced a notion of negative curvature for metric spaces which is now known as *Gromov hyperbolicity*, or simply *hyperbolicity.* Hyperbolicity is a coarse invariant, meaning that a metric space which is "close" to a hyperbolic metric space must also be hyperbolic. The technically correct notion here is *quasi-isometry.* Metric spaces X and Y are quasi-isometric if there is a map $f\colon X \to Y$ which distorts distances by a bounded amount and is coarsely surjective, i.e. every point of Y is within bounded distance of some $f(x)$. If a metric space is quasi-isometric to a hyperbolic space, then it itself is hyperbolic.

Gromov showed that acting properly and cocompactly on a hyperbolic metric space puts strong algebraic constraints on a group. He also pointed out that the Cayley graph of a group with respect to any finite generating set can be regarded as a metric space for the purpose of deciding whether the group so acts, and that Cayley graphs associated to different generating sets are quasi-isometric. In particular, hyperbolicity is a group invariant of finitely-generated groups.

Outer space with the Lipschitz metric is not hyperbolic, and the action of $\mathrm{Out}(F_n)$ is not cocompact. The fact that the action is not cocompact is easily fixed: replace CV_n by its simplicial closure $\overline{\mathrm{CV}}_n$. In addition to cocompactness we have (amazingly) gained hyperbolicity, by a deep theorem of Handel and Mosher:

Theorem 6.1 ([**21**]). *The simplicial closure $\overline{\mathrm{CV}}_n$ of Outer space is hyperbolic.*

Remark 6.2. $\overline{\mathrm{CV}}_n$ is also known as the *free splitting complex*, since vertices can be re-interpreted as actions on trees with one edge-orbit and trivial edge stabilizer, and by Bass-Serre theory such an action corresponds to a splitting of F_n as a free product or HNN extension. As we remarked in Section 2.3 , $\overline{\mathrm{CV}}_n$ is also the same as the sphere complex $\mathcal{S}_n$.

One algebraic consequence of a group being hyperbolic is that it cannot contain any free abelian subgroups of rank two. Since it is easy to find large abelian subgroups in $\mathrm{Out}(F_n)$ we know that $\mathrm{Out}(F_n)$ is not hyperbolic. In particular the action of $\mathrm{Out}(F_n)$ on $\overline{CV}_n$ could not be proper. In fact the stabilizer of any simplex at infinity is infinite.

This mirrors the situation for Teichmüller space of a surface S with its action by the mapping class group $\mathrm{Mod}(S)$. Teichmüller space is not hyperbolic with any $\mathrm{Mod}(S)$-invariant metric, but there is a related hyperbolic complex with simplices "at infinity" called the *curve complex* $\mathcal{C}(S)$. The action of the mapping class group on $\mathcal{C}(S)$ is not proper, but one can nevertheless use hyperbolicity to establish properties of the mapping class group. The key to these proofs lies in the fact, due to Masur and Minsky, that distances in (the non-hyperbolic group) $\mathrm{Mod}(S)$ can be approximately measured by adding up distances in the (hyperbolic!) curve complexes associated to all sub-surfaces of S [**26, 27**].

Masur and Minsky used the Teichmüller metric on Teichmüller space in an essential way in all of their work, and most of the current work on the geometry of Outer space is inspired by attempts to adapt their methods to the context of Outer space and the Lipschitz metric. In the next sections we will describe some of this work.

7. Contracting axes for iwips

Although Teichmüller space is not hyperbolic, geodesics in certain directions behave like geodesics in a hyperbolic space. Consider a geodesic in the hyperbolic plane and a ball of any radius disjoint from the geodesic. The projection of the ball onto the geodesic has uniformly bounded diameter. (This is in marked contrast with the behavior of balls and geodesics in the Euclidean plane!) A geodesic with this property in any metric space is called a *contracting* geodesic, and all geodesics in a hyperbolic space are contracting. Masur and Minsky showed that in Teichmüller space with the Teichmüller metric, the axis of any pseudo-Anosov mapping class is a contracting geodesic.

One property of pseudo-Anosov mapping classes is that they are *irreducible*, i.e. no proper sub-surface is preserved. An outer automorphism of F_n is similarly called *irreducible* if no proper free factor of F_n is preserved (i.e. sent to a conjugate of itself ... remember we are talking about *outer* automorphisms). If ϕ is irreducible and all of its powers are irreducible it is called an *iwip* ("irreducible with irreducible powers," also called a *fully irreducible* automorphism). Algom-Kfir showed that geodesics associated to iwips are coarsely contracting in the Lipschitz metric: one can define a notion of projection of all of Outer space onto the geodesic, and the image of a ball sufficiently far away from the axis has uniformly bounded diameter [**1**]. In order to make sense of this all notions must be translated into the language of coarse geometry: the "geodesic" associated to an iwip is not unique, the projection of a point is a bounded set, not a point, etc. But hyperbolicity is only a coarse notion anyway, so all of the benefits it confers are still available.

8. Using the Lipschitz metric to prove hyperbolicity of the free factor complex (Bestvina-Feighn)

A foundational result in all of Masur and Minsky's work is the fact that the curve complex is hyperbolic. For $\mathrm{Out}(F_n)$ there are actually several reasonable candidates for an analog of the curve complex. We have already mentioned the simplicial closure $\overline{\mathrm{CV}}_n$ but another natural choice is the complex of (conjugacy classes of) free factors $\mathcal{FF}_n$. A vertex of $\mathcal{FF}_n$ is a free factor of F_n, i.e. a subgroup A generated by part of a basis for F_n. Two free factors A and B are connected by an edge if some conjugate of A is a subgroup of B.

Bestvina and Feighn proved that $\mathcal{FF}_n$ is hyperbolic by adapting Masur and Minsky's methods to Outer space with the Lipschitz metric [**10**]. This was followed quite soon by Handel and Mosher's proof that $\overline{\mathrm{CV}}_n$ is hyperbolic, by much more combinatorial methods [**21**]. Kapovich and Rafi showed that in fact hyperbolicity of $\mathcal{FF}_n$ can be derived from hyperbolicity of $\overline{\mathrm{CV}}_n$ [**25**]. But in their work on subfactor projections Bestvina and Feighn needed stronger results than were available from Handel and Mosher's proof, so they showed that the Lipschitz metric and folding paths could be used to streamline the Handel-Mosher proof and in the process find more quasi-geodesics [**11**]. In the next few paragraphs I will attempt to give some of the ideas involved in their proof for $\mathcal{FF}_n$.

The proof depends on a criterion developed by Brian Bowditch for proving that a space X is hyperbolic [**14**]. He shows that it is sufficient to find a constant C and a family of paths in X satisfying the following conditions:

(1) Any two points $x, y \in X$ are coarsely joined by one of the paths, i.e. there is a path in the family with one endpoint within C of x and the other within C of y;
(2) The paths are unparamaterized quasi-geodesics, i.e. they stay inside a C-neighborhood of an actual geodesic;
(3) The paths satisfy the "thin triangles" condition, i.e. for any three points x, y, and z, a path in the family from x to z is the C-neighborhood of the union of any paths from x to y and y to z.

So to use Bowditch's criterion, we need to find a family of paths in $\mathcal{FF}_n$. We do have a family of preferred paths in CV_n, namely folding paths. And it is easy to define a projection from CV_n to $\mathcal{FF}_n$: for any marked graph (g, Γ), consider the free factors of F_n determined by proper subgraphs of Γ. We can define the projection by taking any one of them; this is coarsely well-defined since any two are within distance 4 in the free factor complex, and is coarsely Lipschitz. We now have a set of vertices of $\mathcal{FF}_n$ which changes at discrete times along the folding path, and we can connect the dots to form a path in $\mathcal{FF}_n$.

To check the first condition, note that we can coarsely connect any two free factors by projecting folding paths between marked graphs in CV_n such that the first free factor is realized as a subgraph of the first marked graph, and the second as a subgraph of the second.

We next have to show these paths are quasi-geodesics. To do this, Bestvina and Feighn define a projection in the other direction, from $\mathcal{FF}_n$ onto a folding path. Morally, the idea is to project a free factor A to the segment of the folding path in which A is "smallest," and show that that segment is of uniformly bounded size.

One way of measuring the size of a free factor A in a marked graph (g, Γ) is to "lift" Γ to the (unprojectivized) Outer space of A by taking the core of the cover corresponding to the subgroup $A < F_n = \pi_1(G)$. We can then measure the volume of the lift. Another way to describe this lift is to think of a point in CV_n as an action of F_n on a metric tree. The subgroup $A < F_n$ also acts on this tree, so there is a minimal subtree (consisting of the axes of all elements of A) which is a point in the Outer space for A. The volume of the lift is the volume of a fundamental domain for the action of A on this minimal subtree.

Unfortunately minimizing volume in this naive way doesn't work, and it is not so easy to define a projection of $\mathcal{FF}_n$ to a folding path with good control. The definition that Bestvina and Feighn come up with does begin by lifting each graph in the folding path to a path in the Outer space of A. The turns of the A-lift of Γ project to turns in Γ, so we can pull back the train-track structure on Γ to define a train track structure on its A-lift. It turns out that in the forward direction along the folding path these A-lifts have longer and longer immersed legal paths, and in the backwards direction they have only very short immersed legal paths. "Right" and "left" projections of A to the folding path can now be defined by specifying the times at which the lifts develop long legal paths or at which they have long almost totally illegal paths. Bestvina and Feighn show that using these projections one gets a coarse Lipschitz retraction from all of $\mathcal{FF}_n$ to the image of the folding path in $\mathcal{FF}_n$, which implies that this image is an unparamaterized quasi-geodesic.

Finally, they prove the thin triangles condition to complete the proof that $\mathcal{FF}_n$ is hyperbolic.

> *The proof that this actually works is a technical tour-de-force relying on an intimate understanding of the evolution of legal systems in folding paths.*

9. Subfactor projections (Bestvina-Feighn)

A second critical element of Masur-Minsky's theory relating the geometry of curve complexes to that of Teichmüller space is the notion of *subsurface projections.* For a subsurface A, the subsurface projection π_A is a map from subsurfaces of S to the curve complex of A. Recall that there is a natural projection of the Teichmüller space of S to the curve complex $\mathcal{C}(S)$ which picks out a shortest curve. As one travels along a Teichüller geodesic one may not be progressing at all in $C(S)$ (e.g. if a single curve α remains shortest), but there has to be some subsurface (not containing α) in which progress is being made. Masur and Minsky prove that in fact distance in Teichmüller space can be approximately measured by adding up distances in the subsurface projections. This has myriad applications, one of which is an application to the *dimension* of the mapping class group.

9.1. **Is the asymptotic dimension of** $\mathrm{Out}(F_n)$ **finite?** The usual (covering) dimension of a metric space is definitely not a quasi-isometry invariant (for example, all compact metric spaces are quasi-isometric!). Gromov defined a new type of dimension called *asymptotic dimension* which *is* invariant under quasi-isometry, so in particular defines an invariant for finitely-generated groups. The asymptotic dimension of any compact set is zero, the asymptotic dimension of $\mathbb{R}^n$ is equal to n and the asymptotic dimension of a tree is equal to one. Even these calculations are non-trivial, however, and in general it is very difficult to compute asymptotic dimension or even show that it is finite. Bestvina, Bromberg and Fujiwara did manage to show that $\mathrm{Mod}(S)$ has finite asymptotic dimension [**8**]. Their strategy was to use curve complexes and subsurface projections to construct a product of hyperbolic spaces on which each infinite order element of $\mathrm{Mod}(S)$ acts with positive translation length. Bell and Fujiwara had previously proved that curve complexes have finite asymptotic dimension [**31**], and this can be used to show that this product space does as well, which is enough to conclude that $\mathrm{Mod}(S)$ has finite asymptotic dimension.

Bestvina and Feighn begin to fill in the pieces of this scheme in the case of $\mathrm{Out}(F_n)$ by defining an analog of subsurface projections called *subfactor projections*, which project one free factor onto the free factor complex of another one [**11**] (technically, their projections land in the free splitting complex of the second free factor, but this can be followed by projection to the free factor complex, which is a uniformly Lipschitz map. Very recently, Sam Taylor has found an improved version of subfactor projection which projects one free factor directly onto the free factor complex of another [**31**].) These projections satisfy the conditions necessary for constructing a product of hyperbolic spaces with an $\mathrm{Out}(F_n)$-action, as in the case of the mapping class group. However, one can only conclude that exponentially growing automorphisms act with positive translation length, not that all infinite order ones do.

Another shortcoming of these subfactor projections is that they do not do a complete job of estimating distance in $\mathrm{Out}(F_n)$; adding up the distances in the free factor complexes of all subfactors gives a lower bound on the distance, but not an upper bound. One might say that not enough projections have yet been defined to measure progress along a geodesic in CV_n. This is a situation which will undoubtedly soon be remedied.

10. Afterword

Things are developing very rapidly in this subject. By the time you read this, more will be known and the theorems will have simpler proofs. Some things can be simplified by regarding $\overline{\mathrm{CV}}_n$ as the sphere complex $\mathcal{S}_n$, including the proof that $\overline{\mathrm{CV}}_n$ is hyperbolic [**23**]. Masur and Minsky's original theorem that the curve complex of a surface is hyperbolic now has a vastly simpler proof and the constant of hyperbolicity has been shown to be independent of the surface by several of groups of people, using new criteria for hyperbolicity and new combinatorial techniques; these simplifications may well be soon adapted to $\mathrm{Out}(F_n)$.

11. Guide to the references

There is quite a large literature by now concerning the geometry of CV_n. The references to this article contain a somewhat arbitrary selection of these papers, which break down roughly as follows.

11.1. **Classics and background.** [**4**], [**13**], [**14**], [**15**], [**22**], [**26**], [**27**], [**30**], [**32**], [**33**].

11.2. **Basics of the metric.** [**2**], [**3**], [**7**], [**16**].

11.3. **Lines in** CV_n**.** [**1**], [**17**], [**19**], [**24**].

11.4. **Geometry of** CV_n **from the point of view of sphere complexes.** [**17**, [**18**], [**24**], [**29**], [**28**].

11.5. **Hyperbolic complexes.** [**9**], [**10**], [**21**], [**23**], [**25**].

11.6. **Related work using more combinatorial methods.** [**19**], [**20**].

11.7. **Projections and asymptotic dimension.** [**5**], [**8**], [**10**], [**11**], [**12**], [**28**], [**31**].

References

1. Yael Algom-Kfir, *Strongly Contracting Geodesics in Outer Space*, Geom. Topol. **15** (2011), no. 4, 2181–2233.
2. Yael Algom-Kfir and Mladen Bestvina, *Asymmetry of Outer Space*, Geom. Dedicata **156** (2012), 81–92.
3. ______, *The Metric Completion of Outer Space*, arXiv:1202.6392.
4. Lipman Bers, *An extremal problem for quasiconformal mappings and a theorem by Thurston*, Acta Math. **141** (1978), 73–98.
5. Gregory C. Bell and Koji Fujiwara, *The asymptotic dimension of a curve graph is finite*, J. Lond. Math. Soc. (2) **77** (2008), no. 1, 33–50.
6. Mladen Bestvina, *A Bers-like proof of the existence of train tracks for free group automorphisms*, Fund. Math. **214** (2011), no. 1, 1–12.
7. ______, *PCMI Lectures on the geometry of Outer space*, (2013), 1–34.
8. Mladen Bestvina, Kenneth Bromberg, and Koji Fujiwara. *Constructing group actions on quasi-trees and applications to mapping class groups*, arXiv:1006.1939.
9. Mladen Bestvina and Mark Feighn, *A hyperbolic* $\mathrm{Out}(F_n)$*-complex*, Groups Geom. Dyn. **4** (2010), no. 1, 31–58.
10. ______, *Hyperbolicity of the complex of free factors*, arXiv:1107.3308. (2011).
11. ______, *Subfactor projections*, arXiv:1211.1730.
12. Mladen Bestvina and Koji Fujiwara, *Quasi-homomorphisms on mapping class groups*, Glas. Mat. Ser. III **42**(62) (2007), no. 1, 213–236.
13. Mladen Bestvina and Michael Handel, *Train Tracks and Automorphisms of Free Groups*, The Annals of Mathematics, Second Series **135** (1992), no. 1, 1–51.
14. Brian H Bowditch, *Intersection numbers and the hyperbolicity of the curve complex*, Journal für die Reine und Angewandte Mathematik. [Crelle's Journal] **598** (2006), 105–129.
15. Marc Culler and Karen Vogtmann, *Moduli of graphs and automorphisms of free groups*, Invent. Math. **84** (1986), no. 1, 91–119.
16. Stephano Francaviglia and Armando Martino, *Metric properties of Outer space*, Publ. Mat. **55** (2011), no. 2, 433–473.
17. Ursula Hamenstädt, *Lines of minima in Outer space*, arXiv:0911.3620 .
18. Ursula Hamenstädt and Sebastian Hensel, *Spheres and Projections for* $\mathrm{Out}(F_n)$, arXiv:1109.2687.
19. Michael Handel and Lee Mosher, *Axes in Outer Space*, Mem. Amer. Math. Soc. **213** (2011), no. 1004, vi+104 pp.

20. Michael Handel and Lee Mosher, *Lipschitz retraction and distortion for subgroups of* $\mathrm{Out}(F_n)$, Geom. Topol. **17** (2013), no. 3, 1535–1579.
21. Michael Handel and Lee Mosher, *The free splitting complex of a free group I: Hyperbolicity*, Geom. Topol. **17** (2013), no. 3, 1581–1672.
22. Allen Hatcher, *Homological stability for automorphism groups of free groups*, Comment. Math. Helv. **70** (1995), no. 1, 39–62.
23. Arnaud Hilion and Camille Horbez, *The hyperbolicity of the sphere complex via surgery paths*, arXiv:1210.6183.
24. Camille Horbez, *Sphere paths in outer space*, Algebr. Geom. Topol. **12** (2012), no. 4, 2493–2517.
25. Ilya Kapovich and Kasra Rafi, *On hyperbolicity of free splitting and free factor complexes*, arXiv:1206.3626.
26. Howard Masur and Yair Minsky, *Geometry of the complex of curves. I. Hyperbolicity*, Invent. Math. **138** (1999), no. 1, 103–149.
27. Howard Masur and Yair Minsky, *Geometry of the complex of curves. II. Hierarchical structure*, Geom. Funct. Anal. **10** (2000), no. 4, 902–974.
28. Lucas Sabalka and Dmitri Savchuk, *Submanifold projection*, arXiv:1211.3111.
29. Lucas Sabalka and Dmitri Savchuk, *On the geometry of a proposed curve complex analogue for* $Out(F_n)$, arXiv:1007.1998.
30. John Stallings, *Finite graphs and free groups.* Combinatorial methods in topology and algebraic geometry (Rochester, N.Y., 1982), 79–84, Contemp. Math., **44**, Amer. Math. Soc., Providence, RI, 1985.
31. Samuel J Taylor, *A note on subfactor projections*, to appear in Algebr. Geom. Topol.
32. William P. Thurston, The Geometry and topology of three-manifolds, Princeton Univ. Press (1978).
33. William P. Thurston, *Minimal stretch maps between hyperbolic surfaces*, arXiv:9801.039.

University of Warwick and Cornell University

RECENT ADVANCES IN SYMPLECTIC FLEXIBILITY

YAKOV ELIASHBERG

Abstract. Flexible and rigid methods coexisted in symplectic topology from its inception. While the rigid methods dominated the development of the subject during the last three decades, the balance has somewhat shifted to the flexible side in the last three years. In the talk we survey the flexibility symplectic advances in the work of E. Murphy, K. Cieliebak, T. Ekholm, I. Smith and the author.

1. The h-principle

Many problems in Mathematics and its applications deal with partial differential equations, partial differential inequalities, of more generally with *partial differential relations*, i.e any conditions imposed on partial derivatives of an unknown function. A solution of such a partial differential relation $\mathcal{R}$ is any function which satisfies this relation.

With any differential relation one can associate the underlying algebraic relation by substituting all the derivatives entering the relation with new independent functions. A solution of the corresponding algebraic relation, called a *formal* solution of the original differential relation $\mathcal{R}$, is a necessary condition for the solvability of $\mathcal{R}$. Though it seems that this necessary condition should be very far from being sufficient, it was a surprising discovery in the 1950s of geometrically interesting problems where existence of a formal solution is the only obstruction for the genuine solvability. One of the first such non-trivial examples were the C^1-isometric embedding theorem of J. Nash and N. Kuiper, [**38, 32**], and the immersion theory of S. Smale and M. Hirsch, [**45, 30**]. After Gromov's remarkable series of papers beginning with his paper [**27**] and culminating in his book [**28**] the area crystallized as an independent subject, called the *h*-principle.

Rigid and flexible results coexist in many areas of geometry, but nowhere else they come so close to each other, as in symplectic topology, which serves as a rich source of examples on both sides of the *flexible-rigid spectrum.* Flexible and rigid problems and the development of each side towards the other shaped and continues to shape the subject of symplectic topology from its inception.

2. Symplectic preliminaries

To set the stage we recall some basic notions of symplectic and contact geometry. Symplectic geometry was born as a geometric language of classical mechanics, and similarly contact geometry emerged as a natural set-up for geometric optics and mechanics with non-holonomic constraints.

The cotangent bundle T^*M of any smooth n-dimensional manifold M carries a canonical *Liouville* 1-form λ, usually denoted pdq, which in any local coordinates $(q_1,\dots,q_n)$ on M and dual coordinates $(p_1,\dots,p_n)$ on cotangent fibers can be written as $\lambda=\sum\limits_1^n p_i dq_i$. The differential $\omega := d\lambda = \sum\limits_1^n dp_i\wedge dq_i$ is called the *canonical* *symplectic structure* on the cotangent bundle T^*M. In the Hamiltonian formalism of classical mechanics the cotangent budle T^*M is viewed as the phase space of a mechanical system with the configuration space M, where the p-coordinates correspond to momenta. The full energy of the system expressed through coordinates and momenta, i.e. viewed as a function $H\colon T^*M \to \mathbb{R}$ on the cotangent bundle (or a time-dependent family of functions $H_t\colon T^*M \to \mathbb{R}$ if the system is not conservative) is called the *Hamiltonian* of the system. The dynamics is then defined by the Hamiltonian equations $\dot{z} = X_{H_t}(z), z \in T^*M$, where the Hamiltonian vector field X_{H_t} is determined by the equation $i(X_{H_t})\omega = dH_t$, which in the canonical (p,q)-coordinates has the form

$$X_{H_t} = \sum_1^n -\frac{\partial H_t}{\partial q_i}\frac{\partial}{\partial p_i} + \frac{\partial H_t}{\partial p_i}\frac{\partial}{\partial q_i}.$$

The flow of the vector field X_{H_t} preserves ω, i.e. $X_{H_t}^*\omega = \omega$. The isotopy generated by the vector field X_{H_t} is called *Hamiltonian.*

More generally, the Hamiltonian dynamics can be defined on any $2n$-dimensional manifold endowed with a *symplectic*, i.e. a closed and non-degenerate differential 2-form ω. According to a theorem of Darboux, any such form admits local *canonical coordinates* $p_1,\dots,p_n,q_1,\dots,q_n$ in which it can be written as $\omega = \sum\limits_1^n dp_i \wedge dq_i$. Diffeomorphisms preserving ω are called *symplectomorphisms* or, in the mechanical context, *canonical transformations.* Symplectomorphisms which can be included in a time dependent Hamiltonian flow are called *Hamiltonian.* When $n = 1$ a symplectic form is just an area form, and symplectomorphisms are area preserving transformations. Though in higher dimensions symplectomorphisms are also volume preserving but the subgroup of symplectomorphisms represents only a small part of the group of volume preserving diffeomorphisms.

The projectivized cotangent bundle PT^*M serves as the phase space in the *geometric optics.* It can be interpreted as the *space of contact elements* of the manifold M, i.e. the space of all tangent hyperplanes to M. The form pdq does not descend to PT^*M but its kernel does, and hence the space of contact elements carries a canonical field of tangent to it hyperplanes. This field turns out to be completely non-integrable. It is called a *contact structure.* More generally, a *contact structure* on a $(2n+1)$-dimensional manifold is a completely non-integrable field of tangent hyperplanes ξ, where the complete non-integrability can be expressed by the Frobenius condition $\alpha \wedge (d\alpha)^{\wedge n} \neq 0$ for a 1-form α (locally) defining ξ by the Pfaffian equation $\alpha = 0$. Though at first glance symplectic and contact geometries are quite different, they are in fact tightly interlinked and it is useful to study them in parallel.

An important property of symplectic and contact structures is the following stability theorem, which is due to Moser **[36]** in the symplectic case and to Gray **[25]** in the contact one: *Given a 1-parametric family of symplectic structures ω_t, or contact structures ξ_t on a manifold X, which coincide outside of a compact set and which, in the symplectic case, belong to the same cohomology class with compact support, then there exists an isotopy $h_t : X \to X$ with compact support which starts at the identity $h_0 = \mathrm{Id}$, and such that $h_t^*\omega_t = \omega_0$ or $h_t^*\xi_t = \xi_t$.*

Maximal integral (i.e. tangent to ξ) submanifolds of a $(2n+1)$-dimensional contact manifold (V,ξ) have dimension n and called *Legendrian*. Their symplectic counterparts are n-dimensional submanifolds L of a $2n$-dimensional symplectic manifold (W,ω) which are isotropic for ω, i.e. $\omega|_L = 0$. They are called *Lagrangian* submanifolds. Here are two important examples of Lagrangian submanifolds. A diffeomorphism $f\colon W \to W$ of a symplectic manifold (W,ω) is symplectic if and only if its graph $\Gamma_f = \{(x,f(x));\ x \in W\} \subset (W \times W, \omega \times (-\omega))$ is Lagrangian. A 1-form θ on a manifold M, viewed as a section of the cotangent bundle T^*M is Lagrangian, if and only if it is closed. In particular, if $H_1(M) = 0$ then Lagrangian sections are graphs of differentials of functions, and hence the intersection points of a Lagrangian with the the 0-section are critical points of the corresponding *generating* function. A general Lagrangian submanifold corresponds to a *multivalued function*, called the *front* of the Lagrangian manifold. Given a submanifold $N \subset M$ (of any codimension), the set of all tangent to N hyperplanes in TM is a Legendrian submanifold of the space of contact elements PT^*M.

3. Gromov's alternative and discovery of symplectic rigidity

It was an original idea of H. Poincaré that Hamiltonian systems should satisfy special qualitative properties. In particular, his study of periodic orbits in the so-called restricted 3-body problem led him to the following statement, now known as the "last geometric theorem of H. Poincaré": *any area preserving transformation of an annulus $S^1 \times [0,1]$ which rotates the boundary circles in opposite directions should have at least two fixed points.* Poincaré provided many convincing arguments why the statement should be true [**42**], but the actual proof was found by G.D. Birkhoff [**4**] in 1913, a few months after Poincaré's death. Birkhoff's proof was purely 2-dimensional and further development of Poincaré's dream into what is now called *symplectic topology* had to wait till 1960s when V. I. Arnold [**1**] formulated a number of conjectures formalizing this vision of Poincaré. In particular, one of Arnold's conjectures stated that the number of fixed points of a Hamiltonian diffeomorphism is bounded below by the minimal number of critical points of a function on the symplectic manifold.

At about the same time Gromov was proving his h-principle type results. He realized that symplectic problems exhibited some remarkable flexibility. This called into question whether Arnold's conjectures could be true in dimension > 2.

Among remarkable results pointing towards symplectic flexibility which were proven by Gromov at the end of 60s and the beginning of 70s were:

- h-principle for symplectic and contact structures on open manifolds: in any homotopy class of non-degenerate (not necessarily closed) 2-forms on an open manifold there is a symplectic form in any prescribed cohomology class. Moreover, any 2 such forms are homotopic as symplectic forms. Similarly any almost contact structure, i.e. a pair (λ,η) of 1- and 2-forms on a $(2k+1)$-dimensional open manifold which satisfies $\lambda \wedge \eta^k \neq 0$, is homotopic through almost contact structures to a pair $(\alpha, d\alpha)$.
- h-principle for Lagrangian immersions which asserts that the Lagrangian regular homotopy classes of Lagrangian immersions $L \to X$ are in 1-1 correspondence with homotopy classes of injective Lagrangian homomorphisms $TL \to TX$;

- h-principle for ϵ-Lagrangian *embeddings* (i.e. embeddings whose tangent planes deviate from Lagrangian directions by an angle $< \epsilon$).
- h-principle for the *iso-symplectic* and *iso-contact* embeddings. For instance, in the symplectic case, *if* (M, ω) *and* (N, η) *are two symplectic manifolds such that* $\dim N \geq \dim M + 4$ *then any smooth embedding* $f \colon M \to N$ *which pulls back the cohomology class of the form* η *to the cohomology class of* ω, *and whose differential* df *is homotopic to a symplectic bundle isomorphism, can be* C^0*-approximated by an iso-symplectic embedding* $\widetilde{f} \colon M \to N$, *i.e.* $\widetilde{f}^* \eta = \omega$. For iso-symplectic and iso-contact *immersions* the h-principle holds in codimension 2.

Gromov formulated (and proved) the following alternative: *either the group of symplectomorphisms (resp. contactomorphisms) is* C^0*-closed in the group of all diffeomorphisms, or its* C^0*-closure coincides with the group of volume preserving (resp. all) diffeomorphisms.* One of the corollaries of Gromov's convex integration method was that there are no additional lower bounds for the number of fixed points of a volume preserving diffeomorphism of a manifold of dimension ≥ 3. Clearly the bound on the number of fixed points is a C^0-property, and hence, if the second part of the alternative were true this would imply that Hamiltonian diffeomorphisms of symplectic manifolds of dimension > 2 have no special fixed point properties, and hence Poincaré's theorem and Arnold's conjectures reflected a pure 2-dimensional phenomenon. In fact, it was clear from this alternative, that all basic problems of symplectic topology are tightly interconnected.

Here are some of such problems, besides Gromov's alternative:

(i) Extension of symplectic and contact structures to the ball from a neighborhood of the boundary sphere.
(ii) 1-parametric version of the previous question: *is it true that two structures on the ball which coincide near the boundary and which are formally homotopic relative the boundary, are isotopic?*
(iii) *Fixed point problems for symplectomorphisms.* More generally, Lagrangian intersection problem: *Do Lagrangian manifolds under certain conditions have more intersection points than it is required by topology?*
(iv) Are there any non-formal obstructions to Legendrian isotopy?

Proving an h-principle type statement in one of these problems would imply that *all* symplectic problems have soft solutions, and hence a resolution of Gromov's alternative became a question about existence of symplectic topology as a subject.

At the beginning of 80s Gromov's alternative was resolved in favor of rigidity in the series of works **[3]**, **[7]**, **[14]**, **[15]**, culminating in Gromov's paper **[26]** where he introduced his method of (pseudo-)holomorphic curves in symplectic manifolds, which brought a genuine revolution into this subject. After Gromov's paper the rigid side of symplectic topology began enraveling with an exponentially increasing speed. We just mention here the discovery of Floer homology, Hofer's metric, Gromov-Witten invariants, Symplectic Field Theory, the link with mathematical theory of Mirror Symmetry, as well applications to lower dimensional topology such as Taubes's "Gromov-Witten = Seiberg-Witten" theorem, the Heegaard Floer homology of Ozsváth and Szabó, and the embedded contact homology of Hutchings an Taubes.

Applications of holomorphic curves in Hamiltonian Dynamics brought us closer to the realization of Poincaré's dream of establishing qualitative properties of mechanical systems (e.g. existence and the number of periodic trajectories) without actual solving the equations of motion. In particular, the *Weinstein conjecture* asserting existence of periodic trajectories of Reeb vector fields was proven in many cases, see **[50, 31]** , and in dimension 3 in full generality (see **[47]**.

4. Flexible milestones after the resolution of Gromov's alternative

Though in a shadow of successes on the rigid side, over the years the flexible side of symplectic topology had also a number of success stories. Here are examples of some interesting developments with a distinct flexible flavor.

Overtwisted contact structures. It was understood in 1989, see **[16]**, that in the world of 3-dimensional contact manifolds there is an important dichotomy: if a contact manifold contains the so-called *overtwisted disc*, i.e. an embedded disc which along its boundary is tangent to the contact structure, then the contact structure becomes very flexible and abides a certain h-principle: *two overtwisted contact structures which are homotopic as plane fields are homotopic as contact structures*, and hence in view of Gray's theorem are isotopic. Non-overtwisted contact manifolds are called *tight*, and that is where the rigid methods of symplectic topology are applicable.

The classification of overtwisted contact structures yields similar flexibility results for Legendrian knots in overtwisted contact 3-manifolds. Namely, Legenfdrian knots in the complement of an overtwisted disc, called *loose* in **[18]**, also satisfy an h-principle. The high-dimensional analog of loose knots is discussed in Section 5.1 below.

Despite a significant progress (see e.g. **[39, 40]**), the high-dimensional analogues of the overtwisting phenomenon is far from being understood.

Donaldson's almost holomorphic sections. We already mentioned above Gromov's h-principle for iso-symplectic embeddings in codimension > 2. Applying holomorphic curve technique it is not difficult to construct counter-examples to a similar h-principle in codimension 2. However, Simon Donaldson used his theory of *almost holomorphic sections* of complex line bundles over almost complex symplectic manifolds to prove, among other remarkable results, the following h-principle type theorem:

Theorem 4.1 (**[8]**). *For any closed $2n$-dimensional symplectic manifold (M, ω) with an integral cohomology class $[\omega] \in H^2(M)$ and a sufficiently large integer k there exists a codimension 2 symplectic submanifold $\Sigma \subset M$ which represents the homology class Poincaré dual to $k\omega$. Moreover, the complement $M \setminus \Sigma$ has a homotopy type of an n-dimensional cell complex* (as it is the case for complements of hyperplane sections in complex projective manifolds).

Furthermore, he proved the following *symplectic Lefschetz pencil theorem*:

Theorem 4.2 (**[9]**). *If (V, ω) is a symplectic manifold with integral cohomology class $[\omega] \in H^2(M)$. Then for sufficiently large integer k there exists a topological Lefshetz pencil in which the fibers are symplectic manifolds representing homology class dual to $k[\omega]$.*

By definition the topological Lefshetz pencil is equivalent to the complex algebraic one near all the singularities.

E. Giroux's theory **[23]** of contact book decompositions of contact manifolds can be viewed as an adaption of Donaldson's technique for the contact case.

Symplectic embeddings of polydiscs. Let us denote by $P(r_1,\dots,r_n)$ the polydisc $\{|z_1|\leq r_1,\dots,|z_n|\leq r_n\}\subset\mathbb{C}^n$, where we assume $r_1\leq r_2\leq\dots\leq r_n$. If $P(r_1,\dots,r_n)$ symplectically embeds into $P(R_1,\dots,R_n)$ then famous Gromov's non-squeezing theorem implies that $r_1\leq R_1$. We also have the volume constraint $r_1\dots r_n\leq R_1\dots R_n$.

It was a common believe that when $n>2$ there should be more constraints on the radii besides the Gromov width and volume constraints. However, Larry Guth proved the following remarkable result on the flexible side, which showed that the room for additional constraints is very limited.

Theorem 4.3 (**[29]**)**.** *There exists a constant $C(n)$ depending on the dimension n such that if $C(n)r_1\leq R_1$ and $C(n)r_1\dots r_n\leq R_1\dots R_n$ then a polydisc $P(r_1,\dots,r_n)$ symplectically embeds into $P(R_1,\dots,R_n)$.*

4.1. **Existence of Stein complex structure.** Stein manifolds are complex manifolds which admit proper holomorphic embeddings into $\mathbb{C}^N$. According to a theorem of H. Grauert a Stein manifold can also be characterized as a manifold which admits an exhausting strictly plurisubharmonic function. Here the word *exhausting* means *proper and bounded below*, while a real-valued function $\phi: V\to\mathbb{R}$ on a complex manifold V is called *strictly plurisubharmonic or i-convex* if the Hermitian form $-dd^{\mathbb{C}}\phi=2i\partial\overline{\partial}\phi$ which in local holomorphic coordinates is given by a matrix $\left(\frac{\partial^2\phi}{\partial z_i\partial\bar{z}_j}\right)$ is positive definite. For an arbitrary complex manifold with a complex structure J we will use the term J-convex, instead of strictly plurisubharmonic, to stress the dependence on the complex structure J, Here we denoted by $d^{\mathbb{C}}\phi(X):=d\phi(iX)$ the differential twisted by the operator of multiplication by $\sqrt{-1}$. It can be easily seen that critical points of a Morse strictly plurisubharmonic function on a complex n-dimensional manifold have index $\leq n$, and hence the Morse theory implies that a Stein manifold of complex dimension n has a homotopy type of a cell complex of real dimension n.

The following theorem is proved in **[17]**:

Theorem 4.4 (Existence of Stein structures)**.** *Let (V,J) be any manifold of dimension $2n>4$ and $\phi: V\to\mathbb{R}$ an exhausting Morse function without critical points of index $>n$. Then there exists an integrable complex structure $\widetilde{J}$ on V homotopic to J for which the function ϕ is target equivalent to a $\widetilde{J}$-convex function. In particular, $(V,\widetilde{J})$ is Stein.*

What is transpired from the proof of Theorem 4.4 that it is useful to define a symplectic analog of Stein manifold. The corresponding notion of *Weinstein* manifold, crucial for understanding of Morse theoretic properties of Stein structures, was introduced in **[19]**, formalizing the the Stein handlebody construction from **[17]** and symplectic handlebody construction from Alan Weinstein's paper **[52]**. We discuss this notion and related results in Section 5.5 below.

5. Renaissance of the h-principle in symplectic topology

The last two years witnessed a number of quite unexpected advances on the flexible side of symplectic geometry.

5.1. **Loose Legendrian knots.** It turns out that in contact manifolds of dimension > 3 there is a remarkable class of Legendrian knots, discovered by Emmy Murphy in **[37]**, which satisfies a certain form of an h-principle. These knots are called *loose* in analogy with loose knots in overtwisted contact manifolds. A remarkable fact about Murphy's loose knots is that in contrast with the 3-dimensional case they exist in *all* contact manifolds of dimension > 3.

Stabilization. The *stabilization construction* for Legendrian submanifolds, see **[17], [5], [37]**, can be defined as follows.

Consider standard contact $\mathbb{R}^{2n-1}$:

$$\mathbb{R}^{2n-1}_{\text{st}} = \left(\mathbb{R}^{2n-1}\ ,\ \xi_{\text{st}} = \ker\left(dz - \sum_1^{n-1} y_i dx_i\right)\right),$$

where $(x_1, y_1, \dots, x_{n-1}, y_{n-1}, z)$ are coordinates in $\mathbb{R}^{2n-1}$, and consider a diffeomorphic to $\mathbb{R}^{n-1}$ Legendrian submanifold

$$\Lambda_{\text{cu}} = \left\{(x_1, y_1, \dots, x_{n-1}, y_{n-1}, z)\colon x_1 = \frac{1}{2}y_1^2,\ y_2 = \dots = y_{n-1} = 0,\ z = \frac{1}{3}y_1^3\right\}$$

One can check that given any Legendrian $(n-1)$-submanifold $\Lambda \subset Y$ in a contact $(2n-1)$-manifold Y, any point $p \in \Lambda$ has a neighborhood $\Omega \subset Y$ that admits a map

$$\Phi\colon (\Omega, \Lambda \cap \Omega) \to (\mathbb{R}^{2n-1}_{\text{st}}, \Lambda_{\text{cu}}), \quad \Phi(p) = 0,$$

which is a contactomorphism onto a neighborhood of the origin.

The stabilization construction is a local modification of a Legendrian knot in a neighborhood of a point. It replaces the image of L_{cu} by an image of another Legendrian L^U_{cu}, which coincides with L_{cu} at infinity. We describe this modification below. The two branches of the front Γ_{cu} of the Legendrian Λ_{cu}, i.e. the projection to the $(x_1, \dots, x_{n-1}, z)$-coordinate subspace, are graphs of the functions $\pm h$, where

$$h(x) = h(x_1, \dots, x_{n-1}) = \tfrac{2\sqrt{2}}{3}\, x_1^{\frac{3}{2}},$$

defined on the half-space $\mathbb{R}^{n-1}_+ := \{x = (x_1, \dots, x_{n-1})\colon x_1 \geq 0\}$.

Let U be a domain with smooth boundary contained in the interior of $\mathbb{R}^{n-1}_+$, $U \subset \operatorname{Int}(\mathbb{R}^{n-1}_+)$. Pick a non-negative function $\phi\colon \mathbb{R}^{n+1}_+ \to \mathbb{R}$ with the following properties: ϕ has compact support in $\operatorname{Int}(\mathbb{R}^{n-1}_+)$, the function $\widetilde{\phi}(x) := \phi(x) - 2h(x)$ is Morse, $U = \widetilde{\phi}^{-1}([0, \infty))$, and 0 is a regular value of $\widetilde{\phi}$. Consider the front Γ^U_{cu} in $\mathbb{R}^{n-1} \times \mathbb{R}$ obtained from Γ_{cu} by replacing the lower branch of Γ_{cu}, i.e. the graph $z = -h(x)$, by the graph $z = \phi(x) - h(x)$. Since ϕ has compact support, the front Γ^U_{cu} coincides with Γ_{cu} outside a compact set. Consequently, the Legendrian embedding $\Lambda^U_{\text{cu}}\colon \mathbb{R}^{n-1} \to \mathbb{R}^{2n-1}$ defined by the front Γ^U_{cu} coincides with Λ_{cu} outside a compact set.

It turns out that if (and only if) the Euler characteristic of the domain U is equal to 0 then the Legendrian submanifolds Λ_{cu} and Λ^U_{cu} are *formally* Legendrian isotopic via a compactly supported Legendrian isotopy (however, they are never

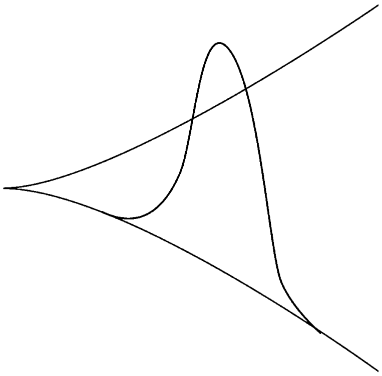


FIGURE 5.1. Stabilization

Legendrian isotopic if $U \neq \varnothing$). We recall that a formal Legendrian isotopy connecting Legendrian embeddings $f_0, f_1 : \Lambda \to (Y, \xi)$ is a pair f_t, Φ_t, $t \in [0,1]$, where f_t is a smooth isotopy and $\Phi_t : T\Lambda \to \xi$ is a family of Lagrangian homomorphisms connecting $\phi_0 = df_0$ and $\Phi_1 = df_1$, and such that the paths of homomorphisms df_t, Φ_t are homotopic with fixed end points as paths of injective homomorphisms $T\Lambda \to TY$. We also note that when $n = 1$ the Euler characteristic $\chi(U)$ is always positive, and hence the stabilization construction never preserves the formal isotopy class of a 1-dimensional Legendrian knot. This is the main point where the theory in high diimension deviates from the 1-dimensional case.

Now, given a Legendrian $(n-1)$-submanifold Λ of a contact $(2n-1)$-manifold Y and contactomorphism

$$\Phi \colon (\Omega, \Lambda \cap \Omega) \to (\mathbb{R}^{2n-1}_{\mathrm{st}}, \Lambda_0),$$

$\Omega \subset Y$ is a neighborhood of a point of $p \in \Lambda$, we replace $\Omega \cap \Lambda$ with $\Phi^{-1}(\Lambda_0^U)$. The resulted Legendrian embedding Λ^U which coincides with Λ outside of Ω is called the *U-stabilization* of Λ in Ω.

5.2. **Murphy's theorem.** A Legendrian embedding $\Lambda \to Y$ of a connected manifold Λ (which we also refer to as a *Legendrian knot*) is called *loose* if it is isotopic to the stabilization of another Legendrian knot. We point out that looseness depends on the ambient manifold. A loose Legendrian embedding Λ into a contact manifold Y need not be loose in a smaller neighborhood Y', $\Lambda \subset Y' \subset Y$.

The above construction shows that a Legendrian submanifold $\Lambda \subset Y$ can be made loose by stabilizing it in arbitrarily small neighborhood of a point, and even without changing its formal Legendrian isotopy class.

It was known since the early days of the h-principle that formally isotopic Legendrian knots become isotopic after a sufficiently many stabilizations. In dimension 3 the corresponding proof was carried out in [**23**]. Moreover, it was shown by J. Etnyre and K. Honda in [**21**] that no number a priopri given number of stabilizations of 1-dimensional Legendrian knots is sufficient.

It was an unexpected discovery of E. Murphy that in dimension > 1 one stabilization id enough. Namely, E. Murphy proved the following h-principle for loose Legendrian knots in contact manifolds of dimension $2n - 1 > 3$:

Theorem 5.1 ([**37**]). *Any two loose Legendrian embeddings which coincide outside a compact set and which can be connected by a formal compactly supported Legendrian isotopy can be connected by a genuine compactly supported Legendrian isotopy.*

5.3. **Lagrangian caps.** Murphy's discovery followed by a number of other results, which seemed to be out of reach before that. In particular, it turned out that Lagrangian embeddings with loose Legendrian boundaries also satisfy an h-principle.

The story begins with the following question: Let B be the round ball in the standard symplectic $\mathbb{R}^{2n}$. *Is there an embedded Lagrangian disc $\Delta \subset \mathbb{R}^{2n} \setminus \operatorname{Int} B$ with $\partial\Delta \subset \partial B$ such that $\partial\Delta$ is a Legendrian submanifold and Δ transversely intersects ∂B along its boundary?*

If $n = 2$ then such a Lagrangian disc does not exist: its existence contradicts the so-called *slice Bennequin inequality*, see [**43**]. Until recently no such examples were known in higher dimensions either. Surprisinhgly, it turns out that when $n > 2$ then Lagrangian discs with *loose* Legendrian boundary satisfy an h-principle, which, in particular, implies that such discs exist in abundance:

Theorem 5.2 ([**20**]). *Let L be a smooth manifold of dimension $n > 2$ with non-empty boundary such that its complexified tangent bundle $TL \otimes \mathbb{C}$ is trivial. Then there exists an exact Lagrangian embedding $f : (L, \partial L) \to (\mathbb{R}^{2n} \setminus \operatorname{Int} B, \partial B)$ with $f(\partial\Delta) \subset \partial B$ such that $f(\partial\Delta) \subset \partial B$ is a Legendrian submanifold and f transverse to ∂B along its boundary ∂L.*

Note that the triviality of the bundle $TL \otimes \mathbb{C}$ is a necessary (and according to Gromov's h-principle for Lagrangian immersions [**28**] sufficient) condition for existence of any Lagrangian *immersion* $L \to \mathbb{C}^n$.

Any Lagrangian embedding of a n-disc to the complement of a $2n$-ball with Legendrian boundary in the boundary sphere of the ball can be completed to a Lagrangian embedding of a n-sphere with a conical singular point. More precisely, given a symplectic manifold (X, ω) we say that $L \subset M$ is a *Lagrangian submanifold with an isolated conical point* if it is a Lagrangian submanifold away from a point $p \in L$, and there exists a symplectic embedding $f : B_\varepsilon \to X$ such that $f(0) = p$ and $f^{-1}(L) \subset B_\varepsilon$ is a Lagrangian cone. Here B_ε is the ball of radius ε in the standard symplectic $\mathbb{R}^{2n}$. Note that this cone is automatically a cone over a Legendrian sphere in the sphere ∂B_ε endowed with the standard contact structure given by the restriction to ∂B_ε of the Liouville form $\lambda_{\mathrm{st}} = \frac{1}{2}\sum\limits_{1}^{n}(p_i dq_i - q_i dp_i)$.

As a special case of Theorem 5.2 (when ∂L is a sphere) we get

Corollary 5.3. *Let L be an n-dimensional, $n > 2$, closed manifold such that the complexified tangent bundle $T^*(L \setminus p) \otimes \mathbb{C}$ is trivial. Then L admits an exact Lagrangian embedding into $\mathbb{R}^{2n}$ with exactly one conical point. In particularly an n-sphere admits a Lagrangian embedding to $\mathbb{R}^{2n}$ with one conical point for each $n > 2$.*

5.4. **Lagrangian non-intersections.** The conical singularity with an appropriate loose Legendrian asymptotics in Corollary 5.3 can be resolved into an immersion with 1 self-intersection point. This leads to surprising, constructions of Lagrangian immersions with minimal number of self-intersection points, which at first glance are going against popular Arnold type Lagrangian intersection conjectures. In particular, we get

Theorem 5.4 (**[13]**). *Let L be an n-dimensional closed manifold with trivial bundle $TL \otimes \mathbb{C}$. We denote by $s(L)$ the minimal number of double points of a Lagrangian immersion of L into the standard symplectic $\mathbb{R}^{2n}$. Then the following hold:*

(i) *If n is odd or if L is non-orientable, then $s(L) \in \{1,2\}$.*
(ii) *If $n = 3$ then $s(L) = 1$.*
(iii) *If n is even and L is orientable, then for $\chi(L) < 0$, $s(L,\sigma) = \frac{1}{2}|\chi(L)|$, and for $\chi(L) \geq 0$, either $s(L) = \frac{1}{2}\chi(L)$ or $s(L) = \frac{1}{2}\chi(L) + 2$.*

The case $n = 2$ is due to D. Sauvaget, **[46]**. It is interesting to compare Theorem 5.4 with the results of **[11, 12]**which show that for even n the standard n-sphere is the only homotopy n-sphere that admits a self-transverse Lagrangian immersion into Euclidean space with only one double point. This means in, particular, that in the case when $\dim(L)$ is even and $\chi(L) > 0$, $s(L)$ is generally not determined by the homotopy type of L. The following result constrains the homotopy type of a manifold for which this phenomenon may occur.

Theorem 5.5 (**[13]**). *Let L be an even dimensional spin manifold with $\chi(L) > 0$. If $s(L) = \frac{1}{2}\chi(L)$ then $\pi_1(L) = 1$ and $H_{2k+1}(L) = 0$ for all k. In particular if* $\dim L >$ *4 then L has the homotopy type of a CW-complex with $\chi(L)$ even-dimensional cells and no odd-dimensional cells.*

It is interesting to note that even for the standard odd-dimensional sphere S^{2k+1} the construction in Theorem 5.4 provides an immersion with a single double point of Maslov index 1, which is different from the standard Whitney Lagrangian immersion $S^{2k+1} \to \mathbb{R}^{4k+2}$, where the intesection point has index n. Using Polterovich's surgery **[48]** we then get

Corollary 5.6. *There exists a Lagrangian embedding $S^1 \times S^{2k} \to \mathbb{R}^{4k+2}_{\mathrm{st}}$ for which the generator of the first homology of positive action has non-positive Maslov index $2-2k$. In particular, there exists a Lagrangian embedding $S^1 \times S^2 \to \mathbb{R}^6_{\mathrm{st}}$ with zero Maslov class.*

Existence of a Lagrangian embedding $S^1 \times S^2 \to \mathbb{R}^6_{\mathrm{st}}$ with zero Maslov class was a well-known problem in symplectic topology.

5.5. **Flexible Stein and Weinstein manifolds.**

5.5.1. *Weinstein and Stein manifolds.* Symplectic topology of Weinstein and Stein manifolds is another playground where the theory of loose Legendrian knots yields interesting applications.

Definition. A *Weinstein structure* on an open manifold V is a triple (ω, X, ϕ), where

- ω is a symplectic form on V,
- $\phi : V \to \mathbb{R}$ is an exhausting Morse (or generalized Morse, i.e. having either non-degenerate or birth-death critical points) function,
- X is a complete vector field which is Liouville for ω (i.e. $L_X\omega = \omega$) and gradient-like for the function ϕ.

The quadruple (V, ω, X, ϕ) is then called a *Weinstein manifold.*

Though any Weinstein structure (ω, X, ϕ) can be perturbed to make the function ϕ Morse, in 1-parameter families of Weinstein structures birth-death zeroes are generically unavoidable.

We will also consider Weinstein *cobordism* structures. Let W be a compact manifold with boundary $\partial W = \partial_+ W \amalg \partial_- W$. A Morse (or generalized Morse) function $\phi : W \to \mathbb{R}$ is called *defining* if $\partial_\pm W$ are regular level sets of ϕ with $\phi|_{\partial_- W} = \min \phi$ and $\phi|_{\partial_+ W} = \max \phi$. The notion of a *Weinstein cobordism* (W, ω, X, ϕ) differs from that of a Weinstein manifold only in replacing the condition that ϕ is exhausting by the requirement that ϕ is a defining function, and replacing the completeness condition of X by the requirement that X points inward along $\partial_- W$ and outward along $\partial_+ W$. A Weinstein cobordism with $\partial_- W = \emptyset$ is called a *Weinstein domain.*

In the complex geometric context, let us recall that a J-convex function $\phi : V \to \mathbb{R}$ on a complex manifold (V, J) serves as a potential of a Kähler metric $H_{J,\phi}(X, Y) := g_{J,\phi}(X, Y) - i\omega_{J,\phi}(X, Y)$, where $\omega_{J,\phi} = -dd^{\mathbb{C}}\phi$ and $g_{J,\phi}(X, Y) = \omega_{J,\phi}(X, JY)$. The gradient $X_{J,\phi} := \nabla_{J,\phi}\phi$ of the function ϕ with respect to the metric $g_{J,\phi}$ is a Liouville field for $\omega_{J,\phi}$, i.e. $L_{X_{J,\phi}}\omega_{J,\phi} = \omega_{J,\phi}$. If (V, J) is Stein then for any exhausting J-convex function $\phi : V \to \mathbb{R}$ the vector field $X_{J,\phi}$ can be made complete by composing ϕ with any function $h : \mathbb{R} \to \mathbb{R}$ with positive first and second derivatives. Assuming that this is already done we associate with a Stein complex manifold (V, J) together with an exhausting J-convex (generalized) Morse function $\phi : V \to \mathbb{R}$ a Weinstein structure $\mathfrak{W}(V, J, \phi) = (V, \omega_{\phi,J}, X_{J,\phi}, \phi)$.

By a *Stein cobordism* structure on a cobordism W, we understand a pair (J, ϕ) where J is an integrable complex structure on W and $\phi : W \to \mathbb{R}$ a defining J-convex function. A Stein cobordism with empty $\partial_- W$ is called a *Stein domain.* As in the manifold case any Stein cobordism structure (J, ϕ) on W determines a Weinstein cobordism structure $\mathfrak{W}(J, \phi) = (W, \omega_{J,\phi}, X_{J,\phi}, \phi)$. The following result is an upgrade of Theorem 4.4.

Theorem 5.7 ([**5**])**.** (i) *Let $\mathfrak{W} = (V, \omega, X, \phi)$ be a Weinstein structure. Then there exists*

- *an integrable complex structure J on V for which ϕ is J-convex and*
- *a homotopy of Weinstein structures $\mathfrak{W}_t = (\omega_t, X_t, \phi)$ connecting $\mathfrak{W}_0 = \mathfrak{W}$ and $\mathfrak{W}_1 = \mathfrak{W}(V, J, \phi)$.*

(ii) *Let V be a manifold of dimension $2n \neq 4$, $\phi : V \to \mathbb{R}$ an exhausting Morse function without critical points of index $> n$, and η a non-degenerate 2-form. Then there exists a Weinstein structure (ω, X, ϕ) on V such that ω and η are homotopic as non-degenerate forms.*

A similar result holds in the cobordism case.

5.5.2. *Flexibility.* Each Weinstein manifold or cobordism can be cut along regular level sets of the function into Weinstein cobordisms that are *elementary* in the sense that there are no trajectories of the Liouville vector field connecting different critical points. An elementary $2n$-dimensional Weinstein cobordism (W, ω, X, ϕ), $n > 2$, is called *flexible* if the attaching spheres of all index n handles form in $\partial_- W$ a *loose Legendrian link.* i.e. each its component is loose in the complements of the others. A Weinstein cobordism or manifold structure (ω, X, ϕ) is called *flexible* if it can be decomposed into elementary flexible cobordisms.

A $2n$-dimensional Weinstein structure (ω, X, ϕ), $n \geq 2$, is called *subcritical* if all critical points of the function ϕ have index $< n$. Any subcritical Weinstein structure in dimension $2n > 4$ is by definition flexible.

Remark 5.8. The property of a Weinstein structure being subcritical is not preserved under Weinstein homotopies because one can always create a pair of critical points of index n and $n-1$. It is an open problem whether or not flexibility is preserved under Weinstein homotopies.

The following results are proven in **[5]**using in a crucial way the theory of loose Legendrian knots.

Theorem 5.9. (i) *Let $\mathfrak{W} = (V, \omega, X, \phi)$ be a flexible Weinstein manifold of dimension $2n > 4$. Then there exists a flexible Weinstein homotopy $\mathfrak{W} = (V, \omega_t, X_t, \phi_t)$, $t \in [0,1]$, with $\mathfrak{W}_0 = \mathfrak{W}$, which is fixed outside a compact set and such that the Morse function ϕ_1 has minimal number of critical points allowed by the Morse theory. If ϕ has finitely many critical points then the hmotopy can be made fixed at infinity.*

(ii) *Let $\mathfrak{W}_0 = (\omega_0, X_0, \phi_0)$ and $\mathfrak{W}_1 = (\omega_1, X_1, \phi_1)$ be two flexible Weinstein structures on a manifold V of dimension $2n$. Suppose that η_0 and η_1 are homotopic as non-degenerate (not necessarily closed) 2-forms. Then $\mathfrak{W}_0$ and $\mathfrak{W}_1$ can be connected by a homotopy $\mathfrak{W}_t = (\omega_t, X_t, \phi_t)$, $t \in [0,1]$, of flexible Weinstein structures.*

An analog of Theorem 5.9 also holds for Weinstein *cobordisms*. Combining with Theorem 5.7 we also get an analog of Theorem 5.9 in the Stein case. We will formulate it here only for Stein cobordisms.

Theorem 5.10 ([**5**]). (i) *Let $\mathfrak{W} = (W, J, \phi)$ be a flexible Stein cobordism of dimension $2n > 4$. Then there exists a homotopy of defining J-convex functions $\phi_t : W \to \mathbb{R}$, $t \in [0,1]$, with $\phi_0 = \phi$, such that the Morse function ϕ_1 has minimal number of critical points allowed by the Morse theory.*

(ii) *Any two flexible Stein cobordism structures (W, J_0) and (W, J_1) which are homotopic as almost complex structures are homotopic as flexible Stein structures.*

In particular, we have the following Stein-Weinstein version of the h-cobordism theorem.

Corollary 5.11 (Weinstein and Stein h-cobordism theorem)**.** *Any flexible Weinstein structure on a product cobordism $W = Y \times [0,1]$ of dimension $2n > 4$ is homotopic to a Weinstein structure (W, ω, X, ϕ), where $\phi : W \to [0,1]$ is a defining function without critical points. Similarly, any flexible Stein cobordism (W, J) which is diffeomorphic to $Y \times [0,1]$ admits a defining J-convex function without critical points .* □

We note that without the flexibility assumption the above claim is wrong, see [**49, 34**].

5.5.3. *Symplectomorphisms of flexible Weinstein manifolds.* Theorem 5.9 has the following consequence for symplectomorphisms of flexible Weinstein manifolds.

Corollary 5.12. *Let $\mathfrak{W} = (V, \omega, X, \phi)$ be a flexible Weinstein manifold of dimension $2n > 4$, and $f : V \to V$ a diffeomorphism such that $f^*\omega$ is homotopic to ω through nondegenerate 2-forms. Then there exists a diffeotopy $f_t : V \to V$, $t \in [0,1]$, such that $f_0 = f$, and f_1 is an exact symplectomorphism of (V, ω).*

Remark 5.13. Even if $\mathfrak{W}$ is of finite type, i.e. ϕ has finitely many critical points, and $f = \mathrm{id}$ outside a compact set, the diffeotopy f_t provided by Theorem 5.12 will be in general *not* equal to the identity outside a compact set.

5.5.4. *Equidimensional symplectic embeddings of flexible Weinstein manifolds.* The following result about equidimesional symplectic embeddings of flexible Weinstein domains is proven in **[20]** as an application of Lagrangian caps technique.

Theorem 5.14 ([**20**])**.** *Let (W, ω, X, ϕ) be a flexible Weinstein domain with Liouville form λ. Let Λ be any other Liouville form on W such that the symplectic forms ω and $\Omega := d\Lambda$ are homotopic as non-degenerate (not necessarily closed) 2-forms. Then there exists an isotopy $h_t : W \hookrightarrow W$ such that $h_0 = \mathrm{Id}$ and $h_1^*\Lambda = \varepsilon\lambda + dH$ for some small $\varepsilon > 0$ and some smooth function $H : W \to \mathbb{R}$. In particular, h_1 defines a symplectic embedding $(W, \varepsilon\omega) \hookrightarrow (W, \Omega)$.*

Corollary 5.15 ([**20**])**.** *Let (W, ω, X, ϕ) be a flexible Weinstein domain and (X, Ω) any symplectic manifold of the same dimension. Then any smooth embedding $f_0 : W \hookrightarrow X$ such that the form $f_0^*\Omega$ is exact and the differential $df : TW \to TX$ is homotopic to a symplectic homomorphism is isotopic to a symplectic embedding $f_1 : (W, \varepsilon\omega) \hookrightarrow (X, \Omega)$ for some small $\varepsilon > 0$. Moreover, if $\Omega = d\Lambda$, then the embedding f_1 can be chosen in such a way that the 1-form $f_1^*\Lambda - i_X\omega$ is exact. If, moreover, the Liouville vector field dual to Λ is complete, then the embedding f_1 exists for an arbitrarily large constant ε.* □

5.6. **Topology of polynomially and rationally convex domains.** We finish this section by implications of the discussed above flexibility results for a problem of a high-dimensional complex analysis concerning the topology of polynomially and rationally convex domains.

Polynomial, rational and holomorphic convexity. Recall the following complex analytic notions of convexity for domains in $\mathbb{C}^n$. For a compact set $K \subset \mathbb{C}^n$, one defines its *polynomial hull* as

$$\widehat{K}_{\mathcal{P}} := \{z \in \mathbb{C}^n \;\Big|\; |P(z)| \le \max_{u \in K} |P(u)| \text{ for all complex polynomials } P \text{ on } \mathbb{C}^n\},$$

and its *rational hull* as

$$\widehat{K}_{\mathcal{R}} := \{z \in \mathbb{C}^n \;\Big|\; |R(z)| \le \max_{u \in K} |R(u)| \text{ for all rational functions } R = \frac{P}{Q},\ Q|_K \ne 0\}.$$

Given an open set $U \supset K$, the *holomorphic hull of K in U* is defined as

$$\widehat{K}^U_{\mathcal{H}} := \{z \in U \;\Big|\; |f(z)| \le \max_{u \in K} |f(u)| \text{ for all holomorphic functions } f \text{ on } U\}.$$

A compact set $K \subset \mathbb{C}^n$ is called *rationally* (resp. *polynomially*) *convex* if $\widehat{K}_{\mathcal{R}} = K$ (resp. $\widehat{K}_{\mathcal{P}} = K$). An open set $U \subset \mathbb{C}^n$ is called *holomorphically convex* if $\widehat{K}^U_{\mathcal{H}}$ is compact for all compact sets $K \subset U$. A compact set $K \subset \mathbb{C}^n$ is called *holomorphically convex* if it is the intersection of its holomorphically convex open neighborhoods. We have

$$\text{polynomially convex} \Longrightarrow \text{rationally convex} \Longrightarrow \text{holomorphically convex.}$$

According to a theorem of E. Levi [**33**], any holomorphically convex domain $W \subset \mathbb{C}^n$ has *weakly* i-convex boundary ∂W. The converse statement that the interior of any domain in $\mathbb{C}^n$ with weakly i-convex boundary is holomorphically convex is known as the *Levi problem.* It was resolved in increasingly more general context in the series of papers begining from K. Oka's paper [**41**] to the paper [**10**] of F. Docquier and H. Grauert.

We call a domain $W \subset \mathbb{C}^n$ *i-convex* if its boundary is i-convex. Note that any weakly i-convex domain in $\mathbb{C}^n$ can be C^∞-approximated by a slightly smaller i-convex one.

Topology of polynomially and rationally convex domains. Any i-convex domain $W \subset \mathbb{C}^n$ admits a defining i-convex function, so in particular it admits a defining Morse function without critical points of index $> n$ (see e.g. [**5**]. It follows that any holomorphically, rationally or polynomially convex domain has the same property. We already stated above, see Theorem 4.4 , that for $n \geq 3$, any domain in $\mathbb{C}^n$ with such a Morse function is smoothly isotopic to an i-convex one.

It turns out, in the spirit of Theorem 4.4, that for $n \geq 3$ there are no additional constraints on the topology of rationally convex domains.

Theorem 5.16 ([**6**]). *A compact domain $W \subset \mathbb{C}^n$, $n \geq 3$, is smoothly isotopic to a rationally convex domain if and only if it admits a defining Morse function without critical points of index $> n$.*

The next result gives necessary and sufficient constraints on the topology of polynomially convex domains.

Theorem 5.17 ([**6**]). *A compact domain $W \subset \mathbb{C}^n$, $n \geq 3$, is smoothly isotopic to a polynomially convex domain if and only if it satisfies the following topological condition:*

(T) *W admits a defining Morse function without critical points of index $> n$, and $H_n(W; G) = 0$ for every abelian group G.*

The "only if" part is well known, see [**1**] (see also [**22**]). Note that, in view of the universal coefficient theorem, condition (T) is equivalent to the condition

(T') W admits a defining Morse function without critical points of index $> n$, $H_n(W) = 0$, and $H_{n-1}(W)$ has no torsion.

Further analysis of condition (T) yields

Proposition 5.18. *(a) If W is simply connected, then condition (T) is equivalent to the existence of a defining Morse function without critical points of index $\geq n$.*

(b) For any $n \geq 3$ there exists a (non-simply connected) domain W satisfying condition (T) with $\pi_n(W, \partial W) \neq 0$. In particular, W does not admit a defining function without critical points of index $\geq n$.

Theorems 5.16 and 5.17 are consequences of the following more precise result for flexible Stein domains:

Theorem 5.19 ([**6**]). *Let (W, J) be a* flexible *Stein domain of complex dimension $n \geq 3$, and $f : W \hookrightarrow \mathbb{C}^n$ a smooth embedding such that f^*i is homotopic to J through almost complex structures. Then (W, J) is deformation equivalent to a rationally convex domain in $\mathbb{C}^n$. More precisely, f is smoothly isotopic to an embedding $g : W \hookrightarrow \mathbb{C}^n$ such that $g(W) \subset \mathbb{C}^n$ is rationally convex, and g^*i is Stein homotopic to J. If in addition $H_n(W; G) = 0$ for every abelian group G, then $g(W)$ can be made polynomially convex.*

5.7. **Optimistic Principle.** The recent progress on the flexible side of symplectic topology described in Section 5 makes me somewhat optimistic that many old problems of symplectic topology have, in fact, flexible solutions. Let me, therefore formulate the following

Symplectic Optimism Principle: If one cannot disprove a flexible h-principle type conjecture using Floer homological methods, or any holomorphic curve related techniques, then it should be true.

Here are a couple of examples.

5.7.1. *Extension problem for symplectic and contact structures.* Let us recall problem (i) from Section 3: *When a germ of symplectic or contact structure along the boundary of an n-ball B extends to B?* There is a homotopical obstruction to such an extension: the extendability of the corresponding almost symplectic or almost contact structure. In the symplectic case there is an additional volume obstruction. The flexible h-principle type conjecture states that these conditions should be also sufficient.

However, it follows from an argument similar to Gromov's proof of the non-squeezing theorem that this conjecture is wrong in the symplectic case (in any dimension $n = 2k > 2$). Moreover, 1-parametric versions of the conjecture known to be wrong in both, symplectic and contact cases. On the other hand, the conjecture known to be true in the 3-dimensional contact case. As far as I know, currently there are no Floer homological or other holomorphic curve related tools which one could try to use to construct counterexamples to the flexible conjecture. Hence, the Symplectic Optimism Principle suggests that it could be true.

5.7.2. *Existence of symplectic and contact structures on closed manifolds.* The flexible extension conjecture from 5.7.1 together with Gromov's h-principle for contact structures on open manifolds would imply the flexible solution of the existence problem for contact structures on closed manifolds: *in any homotopy class of an almost symplectic structure on an odd-dimensional manifold could be realized by a contact structure.*

In the symplectic case a similar optimistic flexible existence conjecture is known to be wrong in dimension 4. However, in higher dimensions there are no known obstructions, and even no feasible approaches how such obstruction could be constructed. Could it be then that an optimistic existence conjecture holds for manifolds of higher dimension: *on any manifold M of dimension $n = 2k > 4$ with a cohomology class $\eta \in H^2(M)$ with $\eta^k \neq 0$ and a non-degenerate 2-form ω_0 there exists a symplectic form ω homotopic to ω_0 through non-degenerate forms, whose cohomology class $[\omega]$ can be deformed to η keeping its k-th power non-vanishing.*

As we already pointed out above the analogous statement is wrong in the 4-dimensional case. However, could it be that it is still true for manifolds of "general type": *Given any manifold M with a cohomology class $\eta \in H^2(M;\mathbb{Z})$ with $\eta^2 \neq 0$ and a non-degenerate 2-form ω_0 one can find an orientable surface $F \subset M$ which realizes a homology class dual to $N\eta$ for a sufficiently large N, and such that a certain branch cover of M along F admits a symplectic structure for which the branching locus is symplectic, whose cohomology class is the pull-back of η and which is homotopic to the pull-back of the 2-form ω_0 through non-degenerate forms.*

References

[1] 1. A. Andreotti and R. Narasimhan, *A topological property of Runge pairs*, Ann. of Math. **76** (1962), 499–509.

[2] V.I. Arnold, *Sur une propriété topologique des applications globalement canoniques de la méchanique classique*, C. R. Acad. Paris, **261** (1965), 3719–3722.

[3] D. Bennequin, *Entrelacements et équations de Pfaff*, Astérisque, **107** (1983), 87–161.

[4] G.D. Birkhoff, *Proof of Poincaré's geometric theorem*, Trans. Amer. Math. Soc., **14** (1913), 14–22.

[5] K. Cieliebak and Y. Eliashberg, *From Stein to Weinstein and Back – Symplectic Geometry of Affine Complex Manifolds*, Colloquium Publications Vol. 59, Amer. Math. Soc. (2012).

[6] K. Cieliebak and Y. Eliashberg, *The topology of rationally and polynomially convex domains*, preprint, arXiv:1305.1614.

[7] C.C Conley and E. Zehnder, *The Birkhoff-Lewis fixed point theorem and a conjecture of V. I. Arnold*, Invent. Math., **73** (1983), 33–49.

[8] S.K Donaldson, *Symplectic submanifolds and almost-complex geometry*, J. Differential Geom., **44** (1996), 666–705.

[9] S.K Donaldson, *Lefschetz pencils on symplectic manifolds*, J. Diff. Geom. **53** (1999), 205–236.

[10] F. Docquier and H. Grauert, *Levisches Problem und Rungescher Satz für Teilgebiete Steinscher Mannigfaltigkeiten*, Math. Ann. **140** (1960), 94–123.

[11] T. Ekholm and I. Smith, *Exact Lagrangian immersions with a single double point,* preprint, arXiv:1111.5932.

[12] T. Ekholm and I. Smith, *Exact Lagrangian immersions with one double point revisited,* preprint, arXiv:1211.1715.

[13] T. Ekholm, Y. Eliashberg, E. Murphy and I. Smith *Constructing exact Lagrangian immersions with few double points*, preprint, arXiv:1303.0588.

[14] Y. Eliashberg, *Rigidity of symplectic structures*, preprint 1981.

[15] Y. Eliashberg, *The wave fronts structure theorem and its applications to symplectic topology*, Funct. Anal. i Pril., **21** (1987), 65–72.

[16] Y. Eliashberg, *Classification of overtwisted contact structures on* 3*-manifolds*, Invent. Math. **98**, no. 3, 623–637 (1989).

[17] Y. Eliashberg, *Topological characterization of Stein manifolds of dimension* > 2, Internat. J. Math. **1**, no. 1, 29-46 (1990).

[18] Y. Eliashberg and M. Fraser, *Topologically trivial Legendrian knots*, J. Symp. Geom. **7**, no. 2, 77–127 (2009).

[19] Y. Eliashberg and M. Gromov, *Convex Symplectic Manifolds*, Proceedings of Symposia in Pure Mathematics, vol. 52, Part 2, 135–162 (1991).

[20] Y. Eliashberg and E. Murphy, *Lagrangian caps*, arXiv:1303.0586, to appear in Geom. and Funct. Anal.

[21] J. Etnyre and K. Honda, *On Connected Sums and Legendrian Knots*, Adv. Math., **179** (2003), 59–74.

[22] F. Forstnerič, *Complements of Runge domains and holomorphic hulls*, Mich. Math. J. **41** (1994), 297–308.

[23] D. Fuchs and S. Tabachnikov, *Invariants of Legendrian and transverse knots in the standard contact space*, Topology **36** (1997), 1025–1053.

[24] E. Giroux, *Géométrie de contact: de la dimension trois vers les dimensions supérieures*, Proc. ICM02 (Beijing, 2002), vol. II, 405–414.

[25] J. Gray, *Some global properties of contact structures*, Ann. of Math., **69**(1959), 421–450.

[26] M. Gromov, *Pseudoholomorphic curves in symplectic manifolds*, Invent. Math. **82**, no. 2, 307–347 (1985).

[27] M. Gromov, *Stable mappings of foliations into manifolds*, Izv. Akad. Nauk SSSR Ser. Mat., **33**(1969), 707–734.

[28] M. Gromov, *Partial Differential Relations*, Springer-Verlag, Berlin, 1986.

[29] L. Guth, Symplectic embeddings of polydisks, Invent. Math., **172**(2008), 477–489.

[30] M. Hirsch, *Immersions of manifolds*, Trans. Amer. Math. Soc., **93**(1959), 242–276.

[31] H. Hofer, *Pseudoholomorphic curves in symplectizations with applications to the Weinstein conjecture in dimension three*, Invent. Math., **114**(1993), 515–563.

[32] N.H. Kuiper, *On C^1-isometric imbeddings. I, II*, Nederl. Akad. Wetensch. Proc. Ser. A. **17**(1955), 545-556, 683–689.

[33] E. Levi, *Studii sui punti singolari essenziali delle funzioni analitiche di due o più variabili complesse*, Annali di Mat oura ed appl. **17**, 61–87 (1910).

[34] M. McLean, *Lefschetz fibrations and symplectic homology*, Geom. Topol. **13**, no. 4, 1877–1944 (2009).

[35] J. Milnor, *Whitehead torsion*, Bull. Amer. Math. Soc. **72**, 358–426 (1966).

[36] J. Moser, *On the volume elements on a manifold*, Trans. Amer. Math. Soc., **120**(1965), 286–294.

[37] E. Murphy, *Loose Legendrian embeddings in high dimensional contact manifolds*, arXiv:1201.2245.

[38] J. Nash, C^1*-isometric imbeddings*, Ann. of Math. **60**(1954), 383–396.

[39] K. Niederkruger, *The plastikstufe – a generalization of the overtwisted disk to higher dimensions*, Algebr. Geom. Topol. **6**(2006), 2473–2508.

[40] K. Niederkruger and O. van Koert, *Every Contact Manifolds can be given a Nonfillable Contact Structure*, Int. Math. Res. Notices, 2009, 4463–4479.

[41] K. Oka, *Sur les fonctions analytiques de plusieurs variables IX: Domaines finis sans point critique intérieur*, Jap. J. Math. **23**, 97–155 (1953).

[42] H. Poincaré, *Sur une théorème de géométrie*, Rend. Circ. Mat. Palermo **33**(1912), 375–507.

[43] L. Rudolph, *An obstruction to sliceness via contact geometry and classical gauge theory,* Invent. Math. **119**(1995), 155–163.

[44] S. Smale, *On the structure of manifolds*, Amer. J. of Math. **84**, 387–399 (1962).

[45] S. Smale, *The classification of immersions of spheres in Euclidean spaces*, Ann. of Math., **69**(1959), 327–344.

[46] Sauvaget, D. *Curiosités lagrangiennes en dimension 4.* Ann. Inst. Fourier (Grenoble) **54** (2004) no. 6, 1997–2020.

[47] C.H. Taubes, *The Seiberg-Witten equations and the Weinstein conjecture*, Geom. Topol., **11**(2007), 2117–2202.

[48] L. Polterovich. *The surgery of Lagrange submanifolds.* Geom. Func. Anal. **2** (1991), 198–210.

[49] P. Seidel and I. Smith, *The symplectic topology of Ramanujam's surface*, Comment. Math. Helv. **80**, no. 4, 859–881 (2005).

[50] C. Viterbo, *A proof of Weinstein's conjecture in* $\mathbf{R}^{2n}$, Ann. Inst. H. Poincaré Anal. Non Linéaire, **4**(19870, 337–356.

[51] C.T.C. Wall, *Geometrical connectivity I*, J. London Math. Soc. (2) **3**, 597–604 (1971).

[52] A. Weinstein, *Contact surgery and symplectic handlebodies*, Hokkaido Math. J. **20**(1991), 241–251.

Department of Mathematics, Stanford University, Stanford, CA 94305

PRIMES IN INTERVALS OF BOUNDED LENGTH

ANDREW GRANVILLE

Abstract. In April 2013, Yitang Zhang proved the existence of a finite bound B such that there are infinitely many pairs of distinct primes which differ by no more than B. This is a massive breakthrough, makes the twin prime conjecture look highly plausible (which can be re-interpreted as the conjecture that one can take $B = 2$) and his work helps us to better understand other delicate questions about prime numbers that had previously seemed intractable. The original purpose of this talk was to discuss Zhang's extraordinary work, putting it in its context in analytic number theory, and to sketch a proof of his theorem.

Zhang had even proved the result with $B = 70\,000\,000$. Moreover, a co-operative team, *polymath8*, collaborating only on-line, had been able to lower the value of B to 4680. Not only had they been more careful in several difficult arguments in Zhang's original paper, they had also developed Zhang's techniques to be both more powerful and to allow a much simpler proof. Indeed the proof of Zhang's Theorem, that will be given in the write-up of this talk, is based on these developments.

In November, inspired by Zhang's extraordinary breakthrough, James Maynard dramatically slashed this bound to 600, by a substantially easier method. Both Maynard, and Terry Tao who had independently developed the same idea, were able to extend their proofs to show that for any given integer $m \geqslant 1$ there exists a bound B_m such that there are infinitely many intervals of length B_m containing at least m distinct primes. We will also prove this much stronger result herein, even showing that one can take $B_m = e^{8m+5}$.

If Zhang's method is combined with the Maynard-Tao set up then it appears that the bound can be further reduced to 576. If all of these techniques could be pushed to their limit then we would obtain $B(= B_2)= 12$, so new ideas are still needed to have a feasible plan for proving the twin prime conjecture.

The article will be split into two parts. The first half, which appears here, we will introduce the work of Zhang, Polymath8, Maynard and Tao, and explain their arguments that allow them to prove their spectacular results. As we will discuss, Zhang's main novel contribution is an estimate for primes in relatively short arithmetic progressions. The second half of this article sketches a proof of this result; the Bulletin article will contain full details of this extraordinary work.

Part 1. Primes in short intervals

1. Introduction

1.1. **Intriguing questions about primes.** Early on in our mathematical education we get used to the two basic rules of arithmetic, addition and multiplication.

2010 *Mathematics Subject Classification.* 11P32.

To Yitang Zhang, for showing that one can, no matter what.

When we define a prime number, simply in terms of the number's multiplicative properties, we discover a sequence of numbers, which is easily defined, yet difficult to gain a firm grasp of, perhaps since the primes are defined in terms of what they are not (i.e. that they *cannot* be factored into two smaller integers)).

When one writes down the sequence of prime numbers:

$$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, \dots$$

one sees that they occur frequently, but it took a rather clever construction of the ancient Greeks to even establish that there really are infinitely many. Looking further at a list of primes, some patterns begin to emerge; for example, one sees that they often come in pairs:

$$3 \text{ and } 5;\ 5 \text{ and } 7;\ 11 \text{ and } 13;\ 17 \text{ and } 19;\ 29 \text{ and } 31;\ 41 \text{ and } 43;\ 59 \text{ and } 61, \dots$$

One might guess that there are infinitely many such prime pairs. But this is an open, elusive question, the *twin prime conjecture*. Until recently there was little theoretical evidence for it. All that one could say was that there was an enormous amount of computational evidence that these pairs never quit; and that this conjecture (and various more refined versions) fit into an enormous network of *conjecture*, which build a beautiful elegant structure of all sorts of prime patterns. If the twin prime conjecture were false then the whole edifice would crumble.

The twin prime conjecture is certainly intriguing to both amateur and professional mathematicians alike, though one might argue that it is an artificial question, since it asks for a very delicate additive property of a sequence defined by its multiplicative properties. Indeed, number theorists had struggled, until very recently, to identify an approach to this question that seemed likely to make any significant headway. In this article we will discuss these latest shocking developments. In the first few sections we will take a leisurely stroll through the historical and mathematical background, so as to give the reader a sense of the great theorems that have been recently proved, from a perspective that will prepare the reader for the details of the proof.

1.2. **Other patterns.** Looking at the list of primes above we see other patterns that begin to emerge, for example, one can find four primes which have all the same digits, except the last one:

$$11, 13, 17 \text{ and } 19; \text{ which is repeated with } 101, 103, 107 \text{ and } 109;$$

and one can find many more such examples – are there infinitely many? More simply how about prime pairs with difference 4:

$$3 \text{ and } 7;\ 7 \text{ and } 11;\ 13 \text{ and } 17;\ 19 \text{ and } 23;\ 37 \text{ and } 41;\ 43 \text{ and } 47;\ 67 \text{ and } 71, \dots;$$

or difference 10:

$$3 \text{ and } 13;\ 7 \text{ and } 17;\ 13 \text{ and } 23;\ 19 \text{ and } 29;\ 31 \text{ and } 41;\ 37 \text{ and } 47;\ 43 \text{ and } 53, \dots?$$

Are there infinitely many such pairs? Such questions were probably asked back to antiquity, but the first clear mention of twin primes in the literature appears in a paper of de Polignac from 1849. In his honour we now call any integer h, for which there are infinitely many prime pairs $p, p+h$, a *de Polignac number*.[1]

Then there are the *Sophie Germain pairs*, primes p and $q := 2p+1$, which prove useful in several simple algebraic constructions:[2]

$$2 \text{ and } 5;\ 3 \text{ and } 7;\ 5 \text{ and } 11;\ 11 \text{ and } 23;\ 23 \text{ and } 47;\ 29 \text{ and } 59;\ 41 \text{ and } 83; \dots;$$

Now we have spotted all sorts of patterns, we need to ask ourselves whether there is a way of predicting which patterns can occur and which do not. Let's start by looking at the possible differences between primes: It is obvious that there are not infinitely many prime pairs of difference 1, because one of any two consecutive integers must be even, and hence can only be prime if it equals 2. Thus there is just the one pair, 2 and 3, of primes with difference 1. One can make a similar argument for prime pairs with odd difference. Hence if h is an integer for which there are infinitely many prime pairs of the form $p,\ q = p+h$ then h must be even. We have seen many examples, above, for each of $h = 2,\ h = 4$ and $h = 10$, and the reader can similarly construct lists of examples for $h = 6$ and for $h = 8$, and indeed for any other even h that takes her or his fancy. This leads us to bet on the *generalized twin prime conjecture*, which states that for any even integer $2k$ there are infinitely many prime pairs $p,\ q = p + 2k$.

What about prime triples? or quadruples? We saw two examples of prime quadruples of the form $10n+1,\ 10n+3,\ 10n+7,\ 10n+9$, and believe that there are infinitely many. What about other patterns? Evidently any pattern that includes an odd difference cannot succeed. Are there any other obstructions? The simplest pattern that avoids an odd difference is $n, n+2, n+4$. One finds the one example $3,\ 5,\ 7$ of such a prime triple, but no others. Further examination makes it clear why not: One of the three numbers is always divisible by 3. This is very similar to what happened with $n, n+1$; and one can verify that, similarly, one of $n, n+6, n+12, n+18, n+24$ is always divisible by 5. The general obstruction can be described as follows:

For a given set of distinct integers $a_1 < a_2 < \ldots < a_k$ we say that prime p is an *obstruction* if p divides at least one of $n+a_1, \ldots, n+a_k$, for every integer n. In other words, p divides

$$\mathcal{P}(n) = (n+a_1)(n+a_2)\ldots(n+a_k)$$

for every integer n; which can be classified by the condition that the set $a_1, a_2, \ldots, a_k \pmod p$ includes all of the residue classes mod p. If no prime is an obstruction then we say that $x+a_1, \ldots, x+a_k$ is an *admissible* set of forms.[3].

[1] Some authors make a slightly different definition: That p and $p+h$ should also be consecutive primes.

[2] The group of reduced residues mod q is a cyclic group of order $q-1 = 2p$, and therefore isomorphic to $C_2 \times C_p$ if $p > 2$. Hence the order of each element in the group is either 1 (that is, $1 \pmod q$), 2 (that is, $-1 \pmod q$), p (the squares mod q) or $2p = q-1$. Hence $g \pmod q$ generates the group of reduced residues if and only if g is not a square mod q and $g \not\equiv -1 \pmod q$.

[3] Notice that $a_1, a_2, \ldots, a_k \pmod p$ can occupy no more than k residue classes mod p and so, if $p > k$ then p cannot be an obstruction. Hence, to check whether a given set A of k integers is

In 1904 Dickson made the optimistic conjecture that if there is no such "obvious" obstruction to a set of linear forms being infinitely often prime, then they are infinitely often simultaneously prime. That is:

Conjecture: *If $x + a_1, \dots, x + a_k$ is an admissible set of forms then there are infinitely many integers n such that $n + a_1, \dots, n + a_k$ are all prime numbers.*

In this case, we call $n + a_1, \dots, n + a_k$ a *k-tuple* of prime numbers.

To date, this has not been proven for any $k > 1$ though, following Zhang's work, we begin to get close for $k = 2$. Indeed, Zhang has proved a weak variant of this conjecture for $k = 2$, as we shall see. Moreover Maynard [**29**] , and Tao [**39**] , have gone on to prove a weak variant *for any $k \geqslant 2$.*

The above conjecture can be extended, as is, to all sets of k linear forms with integer coefficients in one variable (for example, the triple $n, 2n + 1,\ 3n - 2$), so long as we extend the notion of admissibility to also exclude the possible obstruction that two of the linear forms have different signs for all but finitely many n, (since, for example, n and $2 - n$, can never be simultaneously prime); some people call this the "obstruction at the 'prime', -1". We can also extend the conjecture to more than one variable (for example the set of forms $m, m + n, m + 4n$):

The prime k-tuplets conjecture: *If a set of k linear forms in n variables is admissible then there are infinitely many sets of n integers such that when we substitute these integers into the forms we get a k-tuple of prime numbers.*

There has been substantial recent progress on this conjecture. The famous breakthrough was Green and Tao's theorem [**19**] for the k-tuple of linear forms in the two variables a and d:

$$a,\ a + d,\ a + 2d, \dots,\ a + (k - 1)d$$

(in other words, there are infinitely many k-term arithmetic progressions of primes.) Along with Ziegler, they went on to prove the prime k-tuplets conjecture for *any* admissible set of linear forms, provided no two satisfy a linear equation over the integers, [**20**]. What a remarkable theorem! Unfortunately these exceptions include many of the questions we are most interested in; for example, $p,\ q = p+2$ satisfy the linear equation $q - p = 2$; and $p,\ q = 2p + 1$ satisfy the linear equation $q - 2p = 1$).

Finally, we also believe that the conjecture holds if we consider any admissible set of k irreducible polynomials with integer coefficients, with any number of variables. For example we believe that $n^2 + 1$ is infinitely often prime, and that there are infinitely many prime triples $m,\ n,\ m^2 - 2n^2$.

admissible, one needs only find one residue class $b_p \pmod p$, for each prime $p \leqslant k$, which does not contain any element of A.

1.3. **The new results; primes in bounded intervals.** In this section we state Zhang's main theorem, as well as the improvement of Maynard and Tao, and discuss a few of the more beguiling consequences:

Zhang's main theorem: *There exists an integer k such that the following is true: If $x + a_1, \ldots, x + a_k$ is an admissible set of forms then there are infinitely many integers n such that* at least two of $n + a_1, \ldots, n + a_k$ *are prime numbers.*

Note that the result states that only two of the $n + a_i$ are prime, not all (as would be required in the prime k-tuplets conjecture). Zhang proved this result for a fairly large value of k, that is $k = 3500000$, which has been reduced to $k = 105$ by Maynard. Of course if one could take $k = 2$ then we would have the twin prime conjecture,[4] but the most optimistic plan at the moment, along the lines of Zhang's proof, would yield $k = 5$.

To deduce that there are bounded gaps between primes from Zhang's Theorem we need only show the existence of an admissible set with k elements. This is not difficult, simply by letting the a_i be the first k primes $> k$.[5] Hence we have proved:

Corollary 1.1 (Bounded gaps between primes)**.** *There exists a bound B such that there are infinitely many integers pairs of prime numbers $p < q < p + B$.*

Finding the smallest B for a given k is a challenging question. The prime number theorem together with our construction above suggests that $B \leqslant k(\log k + C)$ for some constant C, but it is interesting to get better bounds. For Maynard's $k = 105$, Engelsma showed that one can take $B = 600$,[6] and that this is best possible.

Our Corollary further implies

Corollary 1.2. *There is an integer $h, 0 < h \leqslant B$ such that there are infinitely many pairs of primes $p, p + h$.*

That is, some positive integer $\leqslant B$ is a de Polignac number. In fact one can go a little further using Zhang's main theorem, and deduce that if A is *any* admissible set of k integers then there is an integer $h \in (A - A)^+ := \{a - b : \ a > b \in A\}$ such that there are infinitely many pairs of primes $p, p + h$. One can find many beautiful consequences of this; for example, that a positive proportion of even integers are de Polignac numbers.

[4] And the generalized twin prime conjecture, and that there are infinitely many Sophie Germain pairs, and ...

[5] This is admissible since none of the a_i is $0 \pmod p$ for any $p \leqslant k$, and the $p > k$ were handled in the previous footnote.

[6] Sutherland's website `http://math.mit.edu/~primegaps/` gives Engelsma's admissible 105-tuple: $0, 10, 12, 24, 28, 30, 34, 42, 48, 52, 54, 64, 70, 72, 78, 82, 90, 94, 100, 112, 114, 118, 120, 124, 132, 138, 148, 154, 168, 174, 178, 180, 184, 190, 192, 202, 204, 208, 220, 222, 232, 234, 250, 252, 258, 262, 264, 268, 280, 288, 294, 300, 310, 322, 324, 328, 330, 334, 342, 352, 358, 360, 364, 372, 378, 384, 390, 394, 400, 402, 408, 412, 418, 420, 430, 432, 442, 444, 450, 454, 462, 468, 472, 478, 484, 490, 492, 498, 504, 510, 528, 532, 534, 538, 544, 558, 562, 570, 574, 580, 582, 588, 594, 598, 600.$

Next we state the Theorem of Maynard and of Tao:

The Maynard-Tao theorem: *For any given integer $m \geqslant 2$, there exists an integer k such that the following is true: If $x + a_1, \dots, x + a_k$ is an admissible set of forms then there are infinitely many integers n such that* at least m of $n + a_1, \dots, n + a_k$ *are prime numbers.*

This includes and extends Zhang's Theorem (which is the case $k = 2$). The proof even allows one make this explicit (we will obtain $k \leqslant e^{8m+4}$, and Maynard improves this to $k \leqslant cm^2e^{4m}$ for some constant $c > 0$).

Corollary 1.3 (Bounded intervals with m primes)**.** *For any given integer $m \geqslant 2$, there exists a bound B_m such that there are infinitely many intervals $[x, x + B_m]$ (with $x \in \mathbb{Z}$) which contain m prime numbers.*

We will prove that one can take $B_m = e^{8m+5}$ (and Maynard improves this to $B_m = cm^3e^{4m}$ for some constant $c > 0$).

A *Dickson k-tuple* is a set of integers $a_1 < \dots < a_k$ such that there are infinitely many integers for which $n + a_1, n + a_2, \dots, n + a_k$ are each prime

Corollary 1.4. *A positive proportion of m-tuples of integers are Dickson m-tuples.*

Proof. With the notation as in the Maynard-Tao theorem let $R = \prod_{p \leqslant k} p$, select x to be a large integer multiple of R and let $\mathcal{N} := \{n \leqslant x : (n, R) = 1\}$ so that $|\mathcal{N}| = \frac{\phi(R)}{R}x$. Any subset of k elements of $\mathcal{N}$ is admissible, since it does not contain any integer $\equiv 0 \pmod p$ for each prime $p \leqslant k$. There are $\binom{|\mathcal{N}|}{k}$ such k-tuples. Each contains a Dickson m-tuple by the Maynard-Tao theorem.

Now suppose that are $T(x)$ Dickson m-tuples with $1 \leqslant a_1 < \dots < a_m \leqslant x$. Any such m-tuple is a subset of exactly $\binom{|\mathcal{N}|-m}{k-m}$ of the k-subsets of $\mathcal{N}$, and hence

$$T(x) \cdot \binom{|\mathcal{N}| - m}{k - m} \geqslant \binom{|\mathcal{N}|}{k},$$

and therefore $T(x) \geqslant (|\mathcal{N}|/k)^m = (\frac{\phi(R)}{R}/k)^m \cdot x^m$ as desired. □

This proof yields that, as a proportion of the m-tuples in $\mathcal{N}$,

$$T(x)/\binom{|\mathcal{N}|}{m} \geqslant 1/\binom{k}{m}.$$

The $m = 2$ case implies that at least $\frac{1}{5460}$th of the even integers are de Polignac numbers.

Zhang's Theorem and the Maynard-Tao theorem each hold for any admissible k-tuple of linear forms (not just those of the form $x + a$). With this we can prove several other amusing consequences:

• The last Corollary holds if we insist that the primes in the Dickson k-tuples are consecutive primes.

• There are infinitely many m-tuples of consecutive primes such that each pair in the m-tuple differ from one another by just two digits when written in base 10.

• For any $m \geqslant 2$ and coprime integers a and q, there are infinitely many intervals $[x, x + qB_m]$ (with $x \in \mathbb{Z}$) which contain exactly m prime numbers, each $\equiv a \pmod{q}$.[7]

• Let $d_n = p_{n+1} - p_n$ where p_n is the nth smallest prime. Fix $m \geqslant 1$. There are infinitely many n for which $d_n < d_{n+1} < \ldots < d_{n+m}$. There are also infinitely many n for which $d_n > d_{n+1} > \ldots > d_{n+m}$. This was a favourite problem of Paul Erdős, though we do not see how to deduce such a result for other orderings of the d_n.

1.4. **Bounding the gaps between primes. A brief history.** The young Gauss, examining Chernac's table of primes up to one million, surmised that "the density of primes at around x is roughly $1/\log x$". This was subsequently verified, as a consequence of the *prime number theorem*. Therefore we are guaranteed that there are infinitely many pairs of primes $p < q$ with $q - p \leqslant (1 + \epsilon) \log p$ for any fixed $\epsilon > 0$, which is not quite as small a gap as we are hoping for! Nonetheless this raises the question: Fix $c > 0$. Can we even prove that

There are infinitely many pairs of primes $p < q$ with $q < p + c \log p$?

This follows for all $c > 1$ by the prime number theorem, but it is not easy to prove such a result for any particular value of $c \leqslant 1$. The first such results were proved conditionally assuming the Generalized Riemann Hypothesis. This is surprising since the Generalized Riemann Hypothesis was formulated to better understand the distribution of primes in arithmetic progressions, so why would it appear in an argument about short gaps between primes? It is far from obvious by the argument used, and yet this connection deepened and broadened as the literature developed. We will discuss primes in arithmetic progressions in detail in the next section.

The first unconditional (though inexplicit) such result, bounding gaps between primes, was proved by Erdős in 1940 using the small sieve. In 1966, Bombieri and Davenport [**2**] substituted the Bombieri-Vinogradov theorem for the Generalized Riemann Hypothesis in earlier, conditional arguments, to prove this unconditionally for any $c \geqslant \frac{1}{2}$. The Bombieri-Vinogradov Theorem is also a result about primes in arithmetic progressions (as we will discuss later). In 1988 Maier [**28**] observed that one can easily modify this to obtain any $c \geqslant \frac{1}{2}e^{-\gamma}$; and he further improved this, by combining the approaches of Erdős and of Bombieri and Davenport, to obtain some bound a little smaller than $\frac{1}{4}$, in a technical *tour-de-force*.

[7] Thanks to Tristan Freiberg for pointing this out to me.

The first big breakthrough occurred in 2005 when Goldston, Pintz and Yildirim **[15]** were able to show that there are infinitely many pairs of primes $p < q$ with $q < p + c \log p$, for *any* given $c > 0$. Indeed they extended their methods to show that, for any $\epsilon > 0$, there are infinitely many pairs of primes $p < q$ for which

$$q - p < (\log p)^{1/2+\epsilon}.$$

It is their method which forms the basis of the discussion in this paper. Like Bombieri and Davenport, they showed that one can better understand small gaps between primes by obtaining strong estimates on primes in arithmetic progressions, as in the Bombieri-Vinogradov Theorem. Even more, if one assumes a strong, but widely believed, conjecture about the equi-distribution of primes in arithmetic progressions, which extends the Bombieri-Vinogradov Theorem, then one can show that there are infinitely many pairs of primes $p < q$ which differ by no more than 12 (that is, $p < q \leqslant p + 12$)! In fact one can take $k = 5$ in Zhang's theorem, and then apply the result to the admissible 5-tuple, $\{0,\ 2,\ 6,\ 8,\ 12\}$ What an extraordinary statement! We know that if $p < q \leqslant p + 12$ then $q - p = 2,\ 4,\ 6,\ 8,\ 10$ or 12, and so at least one of these difference occurs infinitely often. That is, there exists a positive, even integer $2k \leqslant 12$ such that there are infinitely pairs of primes $p,\ p+2k$. It would be good to refine this further.

After Goldston, Pintz and Yildirim, most of the experts tried and failed to obtain enough of an improvement of the Bombieri-Vinogradov Theorem to deduce the existence of some finite bound B such that there are infinitely many pairs of primes that differ by no more than B. To improve the Bombieri-Vinogradov Theorem is no mean feat and people have longed discussed "barriers" to obtaining such improvements. In fact a technique had been developed by Fouvry **[10]**, and by Bombieri, Friedlander and Iwaniec **[3]**, but this was neither powerful enough nor general enough to work in this circumstance.

Enter Yitang Zhang, an unlikely figure to go so much further than the experts, and to find exactly the right improvement and refinement of the Bombieri-Vinogradov Theorem to establish the existence of the elusive bound B such that there are infinitely many pairs of primes that differ by no more than B. By all accounts, Zhang was a brilliant student in Beijing from 1978 to the mid-80s, finishing with a master's degree, and then working on the Jacobian conjecture for his Ph.D. at Purdue, graduating in 1992. He did not proceed to a job in academia, working in odd jobs, such as in a sandwich shop, at a motel and as a delivery worker. Finally in 1999 he got a job at the University of New Hampshire as a lecturer (though with the same teaching load as tenure-track faculty). From time-to-time a lecturer devotes their energy to working on proving great results, but few have done so with such aplomb as Zhang. Not only did he prove a great result, but he did so by improving *technically* on the experts, having important key ideas that they missed and developing a highly ingenious and elegant construction concerning exponential sums. Then, so as not to be rejected out of hand, he wrote his difficult paper up in such a clear manner that it could not be denied. Albert Einstein worked in a patent office, Yitang Zhang in a Subway sandwich shop; both found time, despite the unrelated calls on their time and energy, to think the deepest thoughts in science. Moreover Zhang's breakthrough came at the relatively advanced age of 50 (or more). Truly *extraordinary.*

After Zhang, a group of researchers decided to team up online to push the techniques, created by Zhang, to their limit. This was the eighth incarnation of the *polymath* project, which is an experiment to see whether this sort of collaboration can help research develop beyond the traditional boundaries set by our academic culture. The original bound of $70,000,000$ was quickly reduced, and seemingly every few weeks, different parts of Zhang's argument could be improved, so that the bound came down in to the thousands. Moreover the polymath8 researchers found variants on Zhang's argument about the distribution of primes in arithmetic progressions, that allow one to avoid some of the deeper ideas that Zhang used. These modifications enabled your author to give an accessible complete proof in this article.

After these clarifications of Zhang's work, two researchers asked themselves whether the original "set-up" of Goldston, Pintz and Yildirim could be modified to get better results. James Maynard obtained his Ph.D. this summer at Oxford, writing one of the finest theses in sieve theory of recent years. His thesis work equipped him perfectly to question whether the basic structure of the proof could be improved. Unbeknownst to Maynard, at much the same time (late October), one of the world's greatest living mathematicians, Terry Tao, asked himself the same question. Both found, to their surprise, that a relatively minor variant made an enormous difference, and that it was suddenly much easier to prove Zhang's Main Theorem and to go far beyond, because one can avoid having to prove any difficult new results about primes in arithmetic progressions. Moreover it is now not difficult to prove results about m primes in a bounded interval, rather than just two.

2. The distribution of primes, divisors and prime k-tuplets

2.1. **The prime number theorem.** As we mentioned in the previous section, Gauss observed, at the age of 16, that "the density of primes at around x is roughly $1/\log x$", which leads quite naturally to the conjecture that

$$\#\{\text{primes } p \leqslant x\} \approx \int_2^x \frac{dt}{\log t} \sim \frac{x}{\log x} \quad \text{as } x \to \infty.$$

(We use the symbol $A(x) \sim B(x)$ for two functions A and B of x, to mean that $A(x)/B(x) \to 1$ as $x \to \infty$.) This was proved in 1896, the *prime number theorem*, and the integral provides a considerably more precise approximation to the number of primes $\leqslant x$, than $x/\log x$. However, this integral is rather cumbersome to work with, and so it is natural to instead weight each prime with $\log p$; that is we work with

$$\Theta(x) := \sum_{\substack{p \text{ prime} \\ p \leqslant x}} \log p$$

and the prime number theorem is equivalent to

$$\Theta(x) \sim x \quad \text{as } x \to \infty. \tag{2.1}$$

2.2. **The prime number theorem for arithmetic progressions, I.** Any prime divisor of (a, q) is an obstruction to the primality of values of the polynomial $qx+a$, and these are the only such obstructions. The prime k-tuplets conjecture therefore implies that if $(a, q) = 1$ then there are infinitely many primes of the form $qn + a$. This was first proved by Dirichlet in 1837. Once proved, one might ask for a more quantitative result. If we look at the primes in the arithmetic progressions (mod 10):

$$\begin{aligned}&11,\ 31,\ 41,\ 61,\ 71,\ 101\\&3,\ 13,\ 23,\ 43,\ 53,\ 73,\ 83,\ 103\\&7,\ 17,\ 37,\ 47,\ 67,\ 97,\ 107\\&19,\ 29,\ 59,\ 79,\ 89,\ 109\end{aligned}$$

then there seem to be roughly equal numbers in each, and this pattern persists as we look further out. Let $\phi(q)$ denote the number of $a \pmod q$ for which $(a, q) = 1$, and so we expect that

$$\Theta(x; q, a) := \sum_{\substack{p \text{ prime} \\ p\leqslant x \\ p\equiv a \pmod q}} \log p \sim \frac{x}{\phi(q)} \quad \text{as } x \to \infty.$$

This is the *prime number theorem for arithmetic progressions* and was first proved by suitably modifying the proof of the prime number theorem.

The function $\phi(q)$ was studied by Euler, who showed that it is *multiplicative*, that is

$$\phi(q) = \prod_{p^e \| q} \phi(p^e)$$

(where $p^e \| q$ means that p^e is the highest power of prime p dividing q) and that $\phi(p^e) = p^e - p^{e-1}$ for all $e \geqslant 1$.

2.3. **The prime number theorem and the Möbius function.** Multiplicative functions lie at the heart of much of the theory of the distribution of prime numbers. One, in particular, the Möbius function, $\mu(n)$, plays a prominent role. It is defined as $\mu(p) = -1$ for every prime p, and $\mu(p^m) = 0$ for every prime p and exponent $m \geqslant 2$; the value at any given integer n is then deduced from the values at the prime powers, by multiplicativity: If n is squarefree then $\mu(n)$ equals 1 or -1 depending on whether n has an even or odd number of prime factors, respectively. One might guess that there are roughly equal numbers of each, which one can phrase as the conjecture that

$$\frac{1}{x}\sum_{n\leqslant x} \mu(n) \to 0 \quad \text{as} \quad n \to \infty.$$

This is a little more difficult to prove than it looks; indeed it is also equivalent to (2.1). That equivalence is proved using the remarkable identity

$$\sum_{ab=n} \mu(a) \log b \ = \begin{cases} \log p & \text{if } n = p^m, \text{ where } p \text{ is prime}, m \geqslant 1; \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

For more on this connection see the forthcoming book **[18]**.

2.4. **A quantitative prime k-tuplets conjecture.** We are going to develop a heuristic to guesstimate the number of pairs of twin primes $p, p+2$ up to x. We start with Gauss's statement that "the density of primes at around x is roughly $1/\log x$. Hence the probability that p is prime is $1/\log x$, and the probability that $p+2$ is prime is $1/\log x$ so, assuming that these events are independent, the probability that p and $p+2$ are simultaneously prime is

$$\frac{1}{\log x} \cdot \frac{1}{\log x} = \frac{1}{(\log x)^2};$$

and so we might expect about $x/(\log x)^2$ pairs of twin primes $p, p+2 \leqslant x$. However there is a problem with this reasoning, since we are implicitly assuming that the events "p is prime for an arbitrary integer $p \leqslant x$", and "$p+2$ is prime for an arbitrary integer $p \leqslant x$", can be considered to be independent. This is obviously false since, for example, if p is even then $p+2$ must also be.[8] So, to correct for the non-independence modulo small primes q, we determine the ratio of the probability that both p and $p+2$ are not divisible by q, to the probabiliity that p and p' are not divisible by q.

Now the probability that q divides an arbitrary integer p is $1/q$; and hence the probability that p is not divisible by q is $1-1/q$. Therefore the probability that both of two independently chosen integers are not divisible by q, is $(1-1/q)^2$.

The probability that q does not divide either p or $p+2$, equals the probability that $p \not\equiv 0$ or $-2 \pmod q$. If $q > 2$ then p can be in any one of $q-2$ residue classes mod q, which occurs, for a randomly chosen $p \pmod q$, with probability $1-2/q$. If $q=2$ then p can be in any just one residue class mod 2, which occurs with probability $1/2$. Hence the "correction factor" for divisibility by 2 is

$$\frac{(1-\frac{1}{2})}{(1-\frac{1}{2})^2} = 2,$$

whereas the "correction factor" for divisibility by any prime $q > 2$ is

$$\frac{(1-\frac{2}{q})}{(1-\frac{1}{q})^2}.$$

Now divisibility by different small primes is independent, as we vary over values of n, by the Chinese Remainder Theorem, and so we might expect to multiply together all of these correction factors, corresponding to each "small" prime q. The question then becomes, what does "small" mean? In fact, it doesn't matter much because the product of the correction factors over larger primes is very close to 1, and hence we can simply extend the correction to be a product over all primes q. (More precisely, the infinite product over all q, converges.) Hence we define the *twin prime constant* to be

$$C := 2 \prod_{\substack{q \text{ prime} \\ q \geqslant 3}} \frac{(1-\frac{2}{q})}{(1-\frac{1}{q})^2} \approx 1.3203236316,$$

[8] Also note that the same reasoning would tell us that there are $\sim x/(\log x)^2$ prime pairs $p, p+1 \leqslant x$.

and we conjecture that the number of prime pairs $p, p+2 \leqslant x$ is

$$\sim C\frac{x}{(\log x)^2}.$$

Computational evidence suggests that this is a pretty good guess. The analogous argument implies the conjecture that the number of prime pairs $p, p+2k \leqslant x$ is

$$\sim C\prod_{\substack{p|k\\ p\geqslant 3}}\left(\frac{p-1}{p-2}\right)\ \frac{x}{(\log x)^2}.$$

This argument is easily modified to make an analogous prediction for any k-tuple: Given $a_1, \dots, a_k$, let $\Omega(p)$ be the set of distinct residues given by $a_1, \dots, a_k \pmod p$, and then let $\omega(p) = |\Omega(p)|$. None of the $n + a_i$ is divisible by p if and only if n is in any one of $p - \omega(p)$ residue classes mod p, and therefore the correction factor for prime p is

$$\frac{(1-\frac{\omega(p)}{p})}{(1-\frac{1}{p})^k}.$$

Hence we predict that the number of prime k-tuplets $n + a_1, \dots, n + a_k \leqslant x$ is,

$$\sim C(a)\frac{x}{(\log x)^k} \quad \text{where} \quad C(a) := \prod_p \frac{(1-\frac{\omega(p)}{p})}{(1-\frac{1}{p})^k}.$$

An analogous conjecture, via similar reasoning, can be made for the frequency of prime k-tuplets of polynomial values in several variables. What is remarkable is that computational evidence suggests that these conjectures do approach the truth, though this rests on the rather shaky theoretical framework given here. A more convincing theoretical framework based on the *circle method* (so rather more difficult) was given by Hardy and Littlewood **[21]**, which we will discuss in the extended (Bulletin) article.

3. Uniformity in arithmetic progressions

3.1. **When primes are first equi-distributed in arithmetic progressions.** There is an important further issue when considering primes in arithmetic progressions: In many applications it is important to know when we are first guaranteed that the primes are more-or-less equi-distributed amongst the arithmetic progressions $a \pmod q$ with $(a, q) = 1$; that is

$$\Theta(x; q, a) \sim \frac{x}{\phi(q)} \text{ for all } (a, q) = 1. \tag{3.1}$$

To be clear, here we want this to hold when x is a function of q, as $q \to \infty$.

Extensive calculations give evidence that, for any $\epsilon > 0$, if q is sufficiently large and $x \geqslant q^{1+\epsilon}$ then the primes up to x are equi-distributed amongst the arithmetic progressions $a \pmod q$ with $(a, q) = 1$, that is (3.1) holds. This is not only unproved at the moment, also no one really has a plausible plan of how to show such a result. However the slightly weaker statement that (3.1) holds for any $x \geqslant q^{2+\epsilon}$, can be shown to be true, assuming the Generalized Riemann Hypothesis. This gives us a clear plan for proving such a result, but one which has seen little progress in the last century!

The best unconditional results known are much weaker than we have hoped for, equidistribution only being proved once $x \geqslant e^{q^{\epsilon}}$. This is the *Siegel-Walfisz Theorem*, and it can be stated in several (equivalent) ways with an error term: For any $B > 0$ we have

$$\Theta(x; q, a) = \frac{x}{\phi(q)} + O\left(\frac{x}{(\log x)^B}\right) \text{ for all } (a, q) = 1. \tag{3.2}$$

Or: for any $A > 0$ there exists $B > 0$ such that if $q < (\log x)^A$ then

$$\Theta(x; q, a) = \frac{x}{\phi(q)} \left\{1 + O\left(\frac{1}{(\log x)^B}\right)\right\} \text{ for all } (a, q) = 1. \tag{3.3}$$

That x needs to be so large compared to q limited the number of applications of this result.

The great breakthough of the second-half of the twentieth century came in appreciating that for many applications, it is not so important that we know that equidistribution holds for *every* a with $(a, q) = 1$, and *every* q up to some Q, but rather that this holds for *most* such q (with $Q = x^{1/2-\epsilon}$). It takes some juggling of variables to state the Bombieri-Vinogradov Theorem: We are interested, for each modulus q, in the size of the largest error term

$$\max_{\substack{a \mod q \\ (a,q)=1}} \left|\Theta(x; q, a) - \frac{x}{\phi(q)}\right|,$$

or even

$$\max_{y \leqslant x} \max_{\substack{a \mod q \\ (a,q)=1}} \left|\Theta(y; q, a) - \frac{y}{\phi(q)}\right|.$$

The bounds $0 \leqslant \Theta(x; q, a) \ll \frac{x}{q} \log x$ are trivial, the upper bound obtained by bounding the possible contribution from each term of the arithmetic progression. (Throughout, the symbol "$\ll$", as in "$f(x) \ll g(x)$" means "there exists a constant $c > 0$ such that $f(x) \leqslant cg(x)$.") We would like to improve on the "trivial" upper bound, perhaps by a power of $\log x$, but we are unable to do so for all q. However, what we can prove is that *exceptional q are few and far between*, and the Bombieri-Vinogradov Theorem expresses this in a useful form. The first thing we do is add up the above quantities over all $q \leqslant Q < x$. The "trivial" upper bound is then

$$\ll \sum_{q \leqslant Q} \frac{x}{q} \log x \ll x(\log x)^2.$$

The Bombieri-Vinogradov states that we can beat this trivial bound by an arbitrary power of $\log x$, provided Q is a little smaller than $\sqrt{x}$:

The Bombieri-Vinogradov Theorem. *For any given $A > 0$ there exists a constant $B = B(A)$, such that*

$$\sum_{q\leqslant Q} \max_{\substack{a \mod q \\ (a,q)=1}} \left|\Theta(x;q,a) - \frac{x}{\phi(q)}\right| \ll_A \frac{x}{(\log x)^A}$$

where $Q = x^{1/2}/(\log x)^B$.

In fact one can take $B = 2A + 5$; and one can also replace the summand here by the expression above with the maximum over y (though we will not need to do this here).

3.2. **Breaking the $x^{1/2}$-barrier.** It is believed that estimates like that in the Bombieri-Vinogradov Theorem hold with Q significantly larger than $\sqrt{x}$; indeed Elliott and Halberstam conjectured [**8**] that one can take $Q = x^c$ for any constant $c < 1$:

The Elliott-Halberstam conjecture *For any given $A > 0$ and η, $0 < \eta < \frac{1}{2}$, we have*

$$\sum_{q\leqslant Q} \max_{\substack{a \mod q \\ (a,q)=1}} \left|\Theta(x;q,a) - \frac{x}{\phi(q)}\right| \ll \frac{x}{(\log x)^A}$$

where $Q = x^{1/2+\eta}$.

However, it was shown in [**13**] that one *cannot* go so far as to take $Q = x/(\log x)^B$.

This conjecture was the starting point for the work of Goldston, Pintz and Yıldırım [**15**], that was used by Zhang [**43**] (which we give in detail in the next section). It can be applied to obtain the following result, which we will prove.

Theorem 3.1 (Goldston-Pintz-Yıldırım). [**15**] *Let $k \geqslant 2$, $l \geqslant 1$ be integers, and $0 < \eta < 1/2$, such that*

$$1 + 2\eta > \left(1 + \frac{1}{2l+1}\right)\left(1 + \frac{2l+1}{k}\right). \tag{3.4}$$

Assume that the Elliott-Halberstam conjecture holds with $Q = x^{1/2+\eta}$. If $x + a_1, \dots, x + a_k$ is an admissible set of forms then there are infinitely many integers n such that at least two of $n + a_1, \dots, n + a_k$ are prime numbers.

The conclusion here is exactly the statement of Zhang's main theorem.

If the Elliott-Halberstam conjecture conjecture holds for some $\eta > 0$ then select l to be an integer so large that $\left(1 + \frac{1}{2l+1}\right) < \sqrt{1 + 2\eta}$. Theorem (3.1) then implies Zhang's theorem for $k = (2l + 1)^2$.

The Elliott-Halberstam conjecture seems to be too difficult to prove for now, but progress has been made when restricting to one particular residue class: Fix integer

$a \neq 0$. We believe that for any fixed $\eta,\ 0 < \eta < \frac{1}{2}$, one has

$$\sum_{\substack{q \leqslant Q \\ (q,a)=1}} \left| \Theta(x;q,a) - \frac{x}{\phi(q)} \right| \ll \frac{x}{(\log x)^A}$$

where $Q = x^{1/2+\eta}$, which follows from the Elliott-Halberstam conjecture (but is weaker).

The key to progress has been to notice that if one can"factor" the key terms here then the extra flexibility allows one to make headway. For example by factoring the modulus q as, say, dr where d and r are roughly some pre-specified sizes. The simplest class of integers q for which this can be done is the *y-smooth integers*, those integers whose prime factors are all $\leqslant y$. For example if we are given a y-smooth integer q and we want $q = dr$ with d not much smaller than D, then we select d to be the largest divisor of q that is $\leqslant D$ and we see that $D/y < d \leqslant D$. This is precisely the class of moduli that Zhang considered.

The other "factorization" concerns the sum $\Theta(x;q,a)$. The terms of this sum can be written as a sum of products, as we saw in (2.2); in fact we will decompose this further, partitioning the values of a and b into different ranges. This will be discussed in full detail in the accompanying article.

Theorem 3.2 (Yitang Zhang's Theorem)**.** *There exist constants $\eta, \delta > 0$ such that for any given integer a, we have*

$$\sum_{\substack{q \leqslant Q \\ (q,a)=1 \\ q \text{ is } y-smooth \\ q \text{ squarefree}}} \left| \Theta(x;q,a) - \frac{x}{\phi(q)} \right| \ll_A \frac{x}{(\log x)^A} \tag{3.5}$$

where $Q = x^{1/2+\eta}$ and $y = x^{\delta}$.

Zhang [**43**] proved his Theorem for $\eta/2 = \delta = \frac{1}{1168}$, and his argument works provided $414\eta + 172\delta < 1$. We will prove this result, by a somewhat simpler proof, provided $162\eta + 90\delta < 1$, and the more sophisticated proof of [**34**] gives (3.5) provided $43\eta + 27\delta < 1$. We expect that this estimate holds for every $\eta \in [0, 1/2)$ and every $\delta \in (0, 1]$, but just proving it for any positive pair $\eta, \delta > 0$ is an extraordinary breakthrough that has an enormous effect on number theory, since it is such an applicable result (and technique). This is the technical result that truly lies at the heart of Zhang's result about bounded gaps between primes, and sketching a proof of this is the focus of the second half of the complete paper (we will give a brief sketch at the end of this article).

4. Goldston-Pintz-Yildirim's argument

We now give a version of the combinatorial argument of Goldston-Pintz-Yıldırım [**15**], which lies at the heart of the proof that there are bounded gaps between primes. (Henceforth we will call it "the GPY argument".)

4.1. **The set up.** Let $\mathcal{H} = (a_1 < a_2 < \ldots < a_k)$ be an admissible k-tuple, and take $x > a_k$. Our goal is to select a weight for which weight$(n) \geqslant 0$ for all n, such that

$$\sum_{x<n\leqslant 2x} \text{weight}(n) \left(\sum_{i=1}^{k} \theta(n+a_i) - \log 3x \right) > 0, \tag{4.1}$$

where $\theta(m) = \log m$ if $m = p$ is prime, and $\theta(m) = 0$ otherwise. If we can do this then there must exist an integer n such that

$$\text{weight}(n) \left(\sum_{i=1}^{k} \theta(n+a_i) - \log 3x \right) > 0.$$

In that case weight$(n) \neq 0$ so that weight$(n) > 0$, and therefore

$$\sum_{i=1}^{k} \theta(n+a_i) > \log 3x.$$

However each $n + a_i \leqslant 2x + a_k < 2x + x$ and so each $\theta(n + a_i) < \log 3x$. This implies that at least two of the $\theta(n + a_i)$ are non-zero, that is, at least two of $n + a_1, \ldots, n + a_k$ are prime.

A simple idea, but the difficulty comes in selecting the function weight(n) with these properties in such a way that we can evaluate the sum. Moreover in **[15]** they also require that weight(n) is sensitive to when each $n + a_i$ is "almost prime". All of these properties can be acquired by using a construction championed by Selberg. In order that weight$(n) \geqslant 0$ one can simply take it to be a square. Hence we select

$$\text{weight}(n) := \left(\sum_{\substack{d|\mathcal{P}(n)\\ d\leqslant R}} \lambda(d) \right)^2,$$

where the sum is over the positive integers d that divide $\mathcal{P}(n)$, and

$$\lambda(d) := \mu(d) G\left(\frac{\log d}{\log R}\right),$$

where $G(.)$ is a measurable, bounded function, supported only on $[0, 1]$.[9], and μ is the Möbius function. Therefore $\lambda(d)$ is supported only on squarefree, positive integers, that are $\leqslant R$.

We can select $G(t) = (1 - t)^m/m!$ to obtain the results of this section but it will pay, for our understanding of the Maynard-Tao construction, if we prove the GPY result for more general $G(.)$.

[9] By *supported only on* we mean "can be non-zero only on".

4.2. **Evaluating the sums over n.** Now, expanding the above sum gives

$$\sum_{\substack{d_1,d_2\leqslant R\\ D:=[d_1,d_2]}} \lambda(d_1)\lambda(d_2)\left(\sum_{i=1}^k \sum_{\substack{x<n\leqslant 2x\\ D|\mathcal{P}(n)}} \theta(n+a_i) - \log 3x \sum_{\substack{x<n\leqslant 2x\\ D|\mathcal{P}(n)}} 1\right). \tag{4.2}$$

Let $\Omega(D)$ be the set of congruence classes $m \pmod D$ for which $D|P(m)$; and let $\Omega_i(D)$ be the set of congruence classes $m \in \Omega(D)$ with $(D, m+a_i) = 1$. Hence the parentheses in the above line equals

$$\sum_{i=1}^k \sum_{m\in\Omega_i(D)} \sum_{\substack{x<n\leqslant 2x\\ n\equiv m \pmod D}} \theta(n+a_i) - \log 3x \sum_{m\in\Omega(D)} \sum_{\substack{x<n\leqslant 2x\\ n\equiv m \pmod D}} 1. \tag{4.3}$$

Our first goal is to evaluate the sums over n. The final sum is easy; there are $x/D + O(1)$ integers in a given arithmetic progression with difference D, in an interval of length x. The error term here is much smaller than the main term, and is easily shown to be irrelevant to the subsequent calculations.

Counting the number of primes in a given arithmetic progression with difference D, in an interval of length x. is much more difficult. We expect that (3.1) holds, so that each

$$\Theta(2x; D, m+a_i) - \Theta(x; D, m+a_i) \sim \frac{x}{\phi(D)}.$$

Here the error terms are larger and more care is needed. They can be handled by standard techniques, provided that the error terms are smaller than the main terms by an arbitrarily large power of $\log x$, at least on average. This shows why the Bombieri-Vinogradov Theorem is so useful, since it implies the needed estimate provided $R < x^{1/4-o(1)}$ so that each $D < x^{1/2-o(1)}$. Going any further is difficult, so that the $\frac{1}{4}$ is an important barrier. Goldston, Pintz and Yıldırım showed that if one can go just beyond $\frac{1}{4}$ then one can prove that there are bounded gaps between primes, but there did not seem to be any techniques available to them to do so.

For the purposes of the next part of this discussion let us not worry about the range in which such an estimate holds, nor about the size of the accumulated error terms, but rather make the substitution and see where it leads. First, though, we need to better understand the sets $\Omega(D)$ and $\Omega_i(D)$. Since they may be constructed using the Chinese Remainder Theorem from the sets with D prime, therefore if $\omega(D) := |\Omega(D)|$ then $\omega(.)$ is a multiplicative function. Moreover each $|\Omega_i(p)| = \omega(p) - 1$, which we denote by $\omega^*(p)$, and each $|\Omega_i(D)| = \omega^*(D)$, extending ω^* to be a multiplicative function. Putting this altogether we obtain in (4.3) a main term of

$$k\omega^*(D)\frac{x}{\phi(D)} - (\log 3x)\omega(D)\frac{x}{D} = x\left(k\frac{\omega^*(D)}{\phi(D)} - (\log 3x)\frac{\omega(D)}{D}\right).$$

This is typically negative which explains why we cannot simply take our weights, $\lambda(d)$, to all be positive. Substituting this in to (4.2) we obtain, in total, the sums

$$x\left(k\sum_{\substack{d_1,d_2\leqslant R\\ D:=[d_1,d_2]}}\lambda(d_1)\lambda(d_2)\frac{\omega^*(D)}{\phi(D)}-(\log 3x)\sum_{\substack{d_1,d_2\leqslant R\\ D:=[d_1,d_2]}}\lambda(d_1)\lambda(d_2)\frac{\omega(D)}{D}\right). \tag{4.4}$$

The two sums over d_1 and d_2 in (4.4) are not easy to evaluate: The use of the Möbius function leads to many terms being positive, and many negative, so that there is, in fact, a lot of cancelation. There are two techniques in analytic number theory that allow one to get accurate estimates for such sums, when there is a lot of cancelation, one more analytic (**[15]**), the other more combinatorial (**[38]**, **[16]**). We will discuss them both, but only fully develop the latter.

4.3. **Evaluating the sums using Perron's formula.** Perron's formula allows one to study inequalities using complex analysis:

$$\frac{1}{2i\pi}\int_{\mathrm{Re}(s)=2}\frac{y^s}{s}\,ds=\begin{cases}1 & \text{if } y>1;\\ 1/2 & \text{if } y=1;\\ 0 & \text{if } 0<y<1.\end{cases}$$

(Here the subscript "$\mathrm{Re}(s)=2$" means that we integrate along the line $s:\ \mathrm{Re}(s)=2$; that is $s=2+it$, with $-\infty<t<\infty$.) So to determine whether $d<R$ we simply compute this integral with $y=R/d$. (The special case, $d=R$, has a negligible effect on our sums, and can be avoided by selecting $R\notin\mathbb{Z}$). Hence the second sum in (4.4) equals

$$\sum_{\substack{d_1,d_2\geqslant 1\\ D:=[d_1,d_2]}}\lambda(d_1)\lambda(d_2)\frac{\omega(D)}{D}\cdot\frac{1}{2i\pi}\int_{\mathrm{Re}(s_1)=2}\frac{(R/d_1)^{s_1}}{s_1}\,ds_1\cdot\frac{1}{2i\pi}\int_{\mathrm{Re}(s_2)=2}\frac{(R/d_2)^{s_2}}{s_2}\,ds_2.$$

Re-organizing this we obtain

$$\frac{1}{(2i\pi)^2}\int_{\substack{\mathrm{Re}(s_1)=2\\ \mathrm{Re}(s_2)=2}}\left(\sum_{\substack{d_1,d_2\geqslant 1\\ D:=[d_1,d_2]}}\frac{\lambda(d_1)\lambda(d_2)}{d_1^{s_1}d_2^{s_2}}\frac{\omega(D)}{D}\right)R^{s_1+s_2}\frac{ds_2}{s_2}\cdot\frac{ds_1}{s_1} \tag{4.5}$$

We will compute the sum in the middle in the special case that $\lambda(d)=\mu(d)$, the more general case following from a variant of this argument. Hence we have

$$\sum_{d_1,d_2\geqslant 1}\frac{\mu(d_1)\mu(d_2)}{d_1^{s_1}d_2^{s_2}}\frac{\omega([d_1,d_2])}{[d_1,d_2]}. \tag{4.6}$$

The summand is a multiplicative function, which means that we can evaluate it prime-by-prime. For any given prime p, the summand is 0 if p^2 divides d_1 or d_2 (since then $\mu(d_1)=0$ or $\mu(d_2)=0$). Therefore we have only four cases to consider: $p\nmid d_1,d_2$; $p|d_1, p\nmid d_2$; $p\nmid d_1, p|d_2$; $p|d_1,p|d_2$, so the pth factor is

$$1-\frac{1}{p^{s_1}}\cdot\frac{\omega(p)}{p}-\frac{1}{p^{s_2}}\cdot\frac{\omega(p)}{p}+\frac{1}{p^{s_1+s_2}}\cdot\frac{\omega(p)}{p}.$$

We have seen that $\omega(p) = k$ for all sufficiently large p so, in that case, the above becomes

$$1 - \frac{k}{p^{1+s_1}} - \frac{k}{p^{1+s_2}} + \frac{k}{p^{1+s_1+s_2}}.$$

In the analytic approach, we compare the integrand to a (carefully selected) power of the Riemann-zeta function, $\zeta(s)$. The pth factor of $\zeta(s)$ is $\left(1 - \frac{1}{p^s}\right)^{-1}$ so, as a first approximation, the last line is roughly

$$\left(1 - \frac{1}{p^{1+s_1+s_2}}\right)^{-k} \left(1 - \frac{1}{p^{1+s_1}}\right)^{k} \left(1 - \frac{1}{p^{1+s_2}}\right)^{k}.$$

Substituting this back into (4.5) we obtain

$$\frac{1}{(2i\pi)^2} \int\int_{\substack{\mathrm{Re}(s_1)=2\\ \mathrm{Re}(s_2)=2}} \frac{\zeta(1+s_1+s_2)^k}{\zeta(1+s_1)^k \zeta(1+s_2)^k} G(s_1, s_2) \ \ R^{s_1+s_2} \frac{ds_2}{s_2} \cdot \frac{ds_1}{s_1}.$$

where

$$G(s_1, s_2) := \prod_{p \text{ prime}} \left(1 - \frac{1}{p^{1+s_1+s_2}}\right)^{k} \left(1 - \frac{1}{p^{1+s_1}}\right)^{-k} \left(1 - \frac{1}{p^{1+s_2}}\right)^{-k} \times \left(1 - \frac{\omega(p)}{p^{1+s_1}} - \frac{\omega(p)}{p^{1+s_2}} + \frac{\omega(p)}{p^{1+s_1+s_2}}\right).$$

The idea is to move both contours in the integral slightly to the left of $\mathrm{Re}(s_1) = \mathrm{Re}(s_2) = 0$, and show that the main contribution comes, via Cauchy's Theorem, from the pole at $s_1 = s_2 = 0$. This can be achieved using our understanding of the Riemann-zeta function, and noting that

$$G(0, 0) := \prod_{p \text{ prime}} \left(1 - \frac{\omega(p)}{p}\right) \left(1 - \frac{1}{p}\right)^{-k} = C(a) \neq 0.$$

Remarkably when one does the analogous calculation with the first sum in (4.4), one takes $k - 1$ in place of k, and then

$$G^*(0, 0) := \prod_{p \text{ prime}} \left(1 - \frac{\omega^*(p)}{p-1}\right) \left(1 - \frac{1}{p}\right)^{-(k-1)} = C(a),$$

also. Since it is so unlikely that these two quite different products give the same constant by co-incidence, one can feel sure that the method is correct!

This was the technique used in **[15]** and, although the outline of the method is quite compelling, the details of the contour shifting can be complicated.

4.4. **Evaluating the sums using Selberg's combinatorial approach, I.** As discussed, the difficulty in evaluating the sums in (4.4) is that there are many positive terms and many negative terms. In developing his upper bound sieve method, Selberg encountered a similar problem and dealt with it in a surprising way, using combinatorial identities to remove this issue. The method rests on a *reciprocity law*: Suppose that $L(d)$ and $Y(r)$ are sequences of numbers, supported only on the squarefree integers. If

$$Y(r) := \mu(r) \sum_{m:\ r|m}{}' L(m) \text{ for all } r \geqslant 1,$$

then

$$L(d) = \mu(d) \sum_{n:\ d|n}{}' Y(n) \text{ for all } d \geqslant 1$$

Here, and henceforth, $\sum'$ denotes the restriction to squarefree integers that are $\leqslant R$. Let ϕ_ω be the multiplicative function (defined here, only on squarefree integers) for which $\phi_\omega(p) = p - \omega(p)$. We apply the above reciprocity law with

$$L(d) := \frac{\lambda(d)\omega(d)}{d} \quad \text{and} \quad Y(r) := \frac{y(r)\omega(r)}{\phi_\omega(r)}.$$

Now since $d_1 d_2 = D(d_1, d_2)$ we have

$$\lambda(d_1)\lambda(d_2)\frac{\omega(D)}{D} = L(d_1)L(d_2)\ \frac{(d_2, d_2)}{\omega((d_2, d_2))}$$

and therefore

$$S_1 := \sum_{\substack{d_1, d_2 \\ D:=[d_1,d_2]}}{}' \lambda(d_1)\lambda(d_2)\frac{\omega(D)}{D} = \sum_{r,s} Y(r)Y(s) \sum_{\substack{d_1, d_2 \\ d_1|r,\ d_2|s}}{}' \mu(d_1)\mu(d_2)\frac{(d_1, d_2)}{\omega((d_1, d_2))}.$$

The summand (of the inner sum) is multiplicative and so we can work out its value, prime-by-prime. We see that if $p|r$ but $p \nmid s$ (or vice-versa) then the sum is $1-1=0$. Hence if the sum is non-zero then $r = s$ (as r and s are both squarefree). In that case, if $p|r$ then the sum is $1-1-1+p/\omega(p) = \phi_\omega(p)/\omega(p)$. Hence the sum becomes

$$S_1 = \sum_r Y(r)^2 \frac{\phi_\omega(r)}{\omega(r)} = \sum_r \frac{y(r)^2\omega(r)}{\phi_\omega(r)}. \tag{4.7}$$

We will select

$$y(r) := F\left(\frac{\log r}{\log R}\right)$$

when r is squarefree, where $F(t)$ is measurable and supported only on $[0,1]$; and $y(r) = 0$ otherwise. Hence we now have a sum with all positive terms so we do not have to fret about complicated cancelations.

4.5. **Sums of multiplicative functions.** An important theme in analytic number theory is to understand the behaviour of sums of multiplicative functions, some being easier than others. Multiplicative functions f for which the $f(p)$ are fixed, or almost fixed, were the first class of non-trivial sums to be determined. Indeed from the Selberg-Delange theorem,[10] one can deduce that

$$\sum_{n \leqslant x} \frac{g(n)}{n} \sim \kappa(g) \cdot \frac{(\log x)^k}{k!}, \tag{4.8}$$

[10] This also follows from the relatively easy proof of Theorem 1.1 of **[26]** .

where

$$\kappa(g) := \prod_{p \text{ prime}} \left(1 + \frac{g(p)}{p} + \frac{g(p^2)}{p^2} + \dots\right)\left(1 - \frac{1}{p}\right)^k$$

when $g(p)$ is typically "sufficiently close" to some given positive integer k that the Euler product converges. Moreover, by partial summation, one deduces that

$$\sum_{n \leqslant x} \frac{g(n)}{n} F\left(\frac{\log n}{\log x}\right) \sim \kappa(g)(\log x)^k \cdot \int_0^1 F(t)\frac{t^{k-1}}{(k-1)!}dt. \tag{4.9}$$

We apply this in the sum above, noting that here $\kappa(g) = 1/C(a)$, to obtain

$$C(a)S_1 = C(a) \sum_r \frac{\omega(r)}{\phi_\omega(r)} F\left(\frac{\log r}{\log R}\right)^2 \sim (\log R)^k \cdot \int_0^1 F(t)^2 \frac{t^{k-1}}{(k-1)!}dt.$$

A similar calculation reveals that

$$C(a)\lambda(d) \sim \mu(d) \cdot (1 - v_d)^k \int_{v_d}^1 F(t)\frac{t^{k-1}}{(k-1)!}dt \cdot (\log R)^k,$$

where $v_d := \frac{\log d}{\log R}$.

4.6. **Selberg's combinatorial approach, II.** A completely analogous calculation, but now applying the reciprocity law with

$$L(d) := \frac{\lambda(d)\omega^*(d)}{\phi(d)} \quad \text{and} \quad Y(r) := \frac{y^*(r)\omega^*(r)}{\phi_\omega(r)},$$

yields that

$$S_2 := \sideset{}{'}\sum_{\substack{d_1, d_2 \\ D := [d_1, d_2]}} \lambda(d_1)\lambda(d_2)\frac{\omega^*(D)}{\phi(D)} = \sum_r \frac{y^*(r)^2\omega^*(r)}{\phi_\omega(r)}. \tag{4.10}$$

We need to determine $y^*(r)$ in terms of the $y(r)$, which we achieve by applying the reciprocity law twice:

$$\begin{aligned} y^*(r) &= \mu(r)\frac{\phi_\omega(r)}{\omega^*(r)} \sum_{d:\ r|d} \frac{\omega^*(d)}{\phi(d)}\mu(d)\frac{d}{\omega(d)} \sum_{n:\ d|n} \frac{y(n)\omega(n)}{\phi_\omega(n)} \\ &= \frac{r}{\phi(r)} \sum_{n:\ r|n} \frac{y(n)}{\phi_\omega(n/r)} \sum_{d:\ d/r|n/r} \mu(d/r)\frac{\omega^*(d/r)d/r}{\phi(d/r)}\omega(n/d) \\ &= r \sideset{}{'}\sum_{n:\ r|n} \frac{y(n)}{\phi(n)} = \frac{r}{\phi(r)} \sideset{}{'}\sum_{m:\ (m,r)=1} \frac{y(mr)}{\phi(m)} \\ &\sim \int_{\frac{\log r}{\log R}}^1 F(t)dt \cdot \log R, \end{aligned}$$

where the last estimate was obtained by applying (4.9) with $k = 1$, and taking care with the Euler product.

We now can insert this into (4.10), and apply (4.9) with k replaced by $k-1$, noting that $\kappa(g^*) = 1/C(a)$, to obtain

$$C(a)S_2 = C(a)\sum_r \frac{y^*(r)^2\omega^*(r)}{\phi_\omega(r)} \sim (\log R)^{k+1}\cdot \int_0^1 \left(\int_t^1 F(u)du\right)^2 \frac{t^{k-2}}{(k-2)!}dt.$$

4.7. **Finding a positive difference; the proof of Theorem 3.1.** From these estimate, we deduce that $C(a)$ times (4.4) is asymptotic to $x(\log 3x)(\log R)^k$ times

$$k\,\frac{\log R}{\log 3x}\cdot \int_0^1 \left(\int_t^1 F(u)du\right)^2 \frac{t^{k-2}}{(k-2)!}dt - \int_0^1 F(t)^2\frac{t^{k-1}}{(k-1)!}dt. \tag{4.11}$$

Define

$$\rho_k(F) := k\int_0^1 \left(\int_t^1 F(u)du\right)^2 \frac{t^{k-2}}{(k-2)!}dt \Big/ \int_0^1 F(t)^2\frac{t^{k-1}}{(k-1)!}dt. \tag{4.12}$$

Assume that the Elliott-Halberstam conjecture holds with exponent $\frac{1}{2}+\eta$, so that we may take $R=\sqrt{Q}$. Hence we deduce that if

$$\frac{1}{2}\left(\frac{1}{2}+\eta\right)\ \rho_k(F) > 1$$

for some F that satisfies the above hypotheses, then (4.11) implies that (4.4), and so (4.1), is > 0

We now need to select a suitable function $F(t)$ to proceed. A good choice is $F(t) = \frac{(1-t)^\ell}{\ell!}$. Using the beta integral identity

$$\int_0^1 \frac{v^k}{k!}\frac{(1-v)^\ell}{\ell!}dv = \frac{1}{(k+\ell+1)!},$$

we obtain

$$\int_0^1 F(t)^2\frac{t^{k-1}}{(k-1)!}dt = \int_0^1 \frac{(1-t)^{2\ell}}{\ell!^2}\frac{t^{k-1}}{(k-1)!}dt = \frac{1}{(k+2\ell)!}\binom{2\ell}{\ell},$$

and

$$\int_0^1 \left(\int_t^1 F(u)du\right)^2 \frac{t^{k-2}}{(k-2)!}dt = \int_0^1 \left(\frac{(1-t)^{\ell+1}}{\ell+1}\right)^2 \frac{t^{k-2}}{(k-2)!}dt = \frac{1}{(k+2\ell+1)!}\binom{2\ell+2}{\ell+1}.$$

Therefore (4.12) is > 0 if (3.4) holds, and so we deduce Theorem 3.1.

In particular if the Elliott-Halberstam conjecture holds with exponent $\frac{1}{2}+\eta$, then we select ℓ to be a sufficiently large integer for which $1+2\eta > \left(1+\frac{1}{2\ell+1}\right)^2$. Selecting $k=(2\ell+1)^2$ we deduce that for every admissible k-tuple, there are infinitely many n for which the k-tuple, evaluated at n, contains two primes.

5. Zhang's modifications of GPY

At the end of the previous section we saw that if the Elliott-Halberstam conjecture holds with any exponent $> \frac{1}{2}$, then for every admissible k-tuple, there are infinitely many n for which the k-tuple contains two primes. However the Elliott-Halberstam conjecture remains unproven.

In (3.5) we stated Zhang's result, which breaks the $\sqrt{x}$-barrier in such results, but at the cost of restricting the moduli to being y-smooth, and restricting the arithmetic progressions $a \pmod q$ to having the same value of a as we vary over q . Can the Goldston-Pintz-Yıldırım argument be modified to handle these restrictions?

5.1. **Averaging over arithmetic progressions.**

In the GPY argument we need estimates for the number of primes in the arithmetic progressions $m + a_i \pmod D$ where $m \in \Omega_i(D)$. When using the Bombieri-Vinogradov Theorem, it does not matter that $m + a_i$ varies as we vary over D; but it does matter when employing Zhang's Theorem 3.2.

Zhang realized that one can exploit the structure of the set $O_i(D) = \Omega_i(D) + a_i$, since it is constructed from the $O_i(p)$ with $p|D$ using the Chinese Remainder Theorem, to get around this issue:

Let $\nu(D)$ denote the number of prime factors of (squarefree) D, so that $\tau(D) = 2^{\nu(D)}$. Any squarefree D can be written as $[d_1, d_2]$ for $3^{\nu(D)}$ pairs d_1, d_2, which means that we need an appropriate upper bound on

$$\leqslant \sum_{D \leqslant Q}{}' \, 3^{\nu(D)} \sum_{b \in O_i(D)} \left| \Theta(X; D, b) - \frac{X}{\phi(D)} \right|$$

where $Q = R^2$ and $X = x$ or $2x$, for each i.

Let L be the lcm of all of the D in our sum. Then the set, $O_i(L)$, reduced mod D, gives $|O_i(L)|/|O_i(D)|$ copies of $O_i(D)$ and so

$$\frac{1}{|O_i(D)|} \sum_{b \in O_i(D)} \left| \Theta(X; D, b) - \frac{X}{\phi(D)} \right| = \frac{1}{|O_i(L)|} \sum_{b \in O_i(L)} \left| \Theta(X; D, b) - \frac{X}{\phi(D)} \right|.$$

Hence we need to divide and multiply by $|O_i(D)|$ in each term of the above sum. Since $|O_i(D)| = \omega^*(D) \leqslant (k-1)^{\nu(D)}$, the above is therefore

$$\begin{aligned} &\leqslant \sum_{D \leqslant Q}{}' \, \tau(D)^A \cdot \frac{1}{|O_i(D)|} \sum_{b \in O_i(D)} \left| \Theta(X; D, b) - \frac{X}{\phi(D)} \right| \\ &= \frac{1}{|O_i(L)|} \sum_{b \in O_i(L)} \sum_{D \leqslant Q}{}' \, \tau(D)^A \cdot \left| \Theta(X; D, b) - \frac{X}{\phi(D)} \right| \\ &\leqslant \max_{a \in \mathbb{Z}} \sum_{\substack{D \leqslant Q \\ (D,a)=1}}{}' \, \tau(D)^A \cdot \left| \Theta(X; D, a) - \frac{X}{\phi(D)} \right| \end{aligned}$$

where $2^A = 3(k-1)$.

It now needs a standard technical argument to bound this using Theorem (3.2): By Cauchy's Theorem, the square of this is

$$\leqslant \sum_{D\leqslant Q} \frac{\tau(D)^{2A}}{D} \cdot \sum_{D\leqslant Q}{}' D\left|\Theta(X;D,b) - \frac{X}{\phi(D)}\right|^2 .$$

The first sum is bounded by $(c\log Q)^{9k^2}$, and we have $D|\Theta(X;D,b)| \leqslant (X + D)\log X$, trivially, and so

$$D\left|\Theta(X;D,b) - \frac{X}{\phi(D)}\right| \ll X\log X.$$

The result now follows by applying Theorem (3.2).

5.2. **Restricting the support to smooth integers.**

Zhang simply took the same coefficients $y(r)$ as above, but now restricted to y-smooth integers; and called this restricted class of coefficients, $z(r)$. Evidently the sum in (4.7) with $z(r)$ in place of $y(r)$, is bounded above by the sum in (4.7). The sum in (4.10) with $z(r)$ in place of $y(r)$, is a little more tricky, since we need a lower bound. Zhang proceeds by showing that if L is sufficiently large and δ sufficiently small, then the two sums differ by only a negligible amount. In particular we will prove Zhang's Theorem when

$$162\eta + 90\delta < 1.$$

Zhang's argument here holds when $L = 863, k = L^2$ and $\eta = 1/(L-1)$.

It should be noted that Motohashi and Pintz **[32]** had already given an argument to accomplish the goals of this section, in the hope that someone might prove an estimate like (3.5)!

6. Goldston-Pintz-Yildirim in higher dimensional analysis

In the set-up in the argument of Goldston, Pintz and Yıldırım, we saw that we study the divisors d of the product of the values of the k-tuple; that is

$$d|\mathcal{P}(n) = (n+a_1)\dots(n+a_k).$$

with $d \leqslant R$.

The latest breakthrough stems from the idea of instead studying the k-tuples of divisors $d_1, d_2, \dots, d_k$ of each individual element of the k-tuple; that is

$$d_1|n+a_1,\ d_2|n+a_2, \dots, d_k|n+a_k.$$

Now, instead of $d \leqslant R$, we take $d_1d_2\dots d_k \leqslant R$.

6.1. **The set up.** One can proceed much as in the previous section, though technically it is easier to restrict our attention to when n is an appropriate congruence class mod m where m is the product of the primes for which $\omega(p) < k$. (because, if $\omega(p) = k$ then p can only divide one $n + a_i$ at a time). Hence we study

$$S_0 := \sum_{r\in\Omega(m)} \sum_{\substack{n\sim x\\ n\equiv r \pmod m}} \left(\sum_{j=1}^{k} \theta(n+a_j) - h\log 3x\right)\left(\sum_{d_i|n+a_i \text{ for each } i} \lambda(d_1,\dots,d_k)\right)^2$$

which upon expanding, as $(d_i, m)|(n + a_i, m) = 1$, equals

$$\sum_{\substack{d_1,\dots,d_k\geqslant 1\\ e_1,\dots,e_k\geqslant 1\\ (d_ie_i,m)=1 \text{ for each } i}} \lambda(d_1,\dots,d_k)\lambda(e_1,\dots,e_k) \sum_{r\in\Omega(m)} \sum_{\substack{n\sim x\\ n\equiv r \pmod m\\ [d_i,e_i]|n+a_i \text{for each } i}} \left(\sum_{j=1}^{k} \theta(n+a_j) - h\log 3x\right).$$

Next notice that $[d_i, e_i]$ is coprime with $[d_j, e_j]$ whenever $i \neq j$, since their gcd divides $(n + a_j) - (n + a_i)$, which divides m, and so equals 1 as $(d_ie_i, m) = 1$. Hence, in our internal sum, the values of n belong to an arithmetic progression with modulus $m\prod_i[d_i, e_i]$. Also notice that if $n + a_j$ is prime then $d_j = e_j = 1$.

Therefore, ignoring error terms,

$$S_0 = \sum_{1\leqslant \ell\leqslant k} \frac{\omega(m)}{\phi(m)} S_{2,\ell}\cdot x - h\frac{\omega(m)}{m} S_1 \cdot x\log 3x$$

where

$$S_1 := \sum_{\substack{d_1,\dots,d_k\geqslant 1\\ e_1,\dots,e_k\geqslant 1\\ (d_i,e_j)=1 \text{ for } i\neq j}} \frac{\lambda(d_1,\dots,d_k)\lambda(e_1,\dots,e_k)}{\prod_i\ [d_i,e_i]}$$

and

$$S_{2,\ell} := \sum_{\substack{d_1,\dots,d_k\geqslant 1\\ e_1,\dots,e_k\geqslant 1\\ (d_i,e_j)=1 \text{ for } i\neq j\\ d_\ell=e_\ell=1}} \frac{\lambda(d_1,\dots,d_k)\lambda(e_1,\dots,e_k)}{\prod_i \phi([d_i,e_i])}.$$

6.2. **The combinatorics.** The reciprocity law generalizes quite beautifully to higher dimension: Suppose that $L(d)$ and $Y(r)$ are two sequences of complex numbers, indexed by $d, r \in \mathbb{Z}^k_{\geqslant 1}$, and non-zero only when each d_i (or r_i) is squarefree. Then

$$L(d_1,\dots,d_k) = \prod_{i=1}^{k} \mu(d_i) \sum_{\substack{r_1,\dots,r_k\geqslant 1\\ d_i|r_i \text{ for all } i}} Y(r_1,\dots,r_k)$$

if and only if

$$Y(r_1,\dots,r_k) = \prod_{i=1}^{k} \mu(r_i) \sum_{\substack{d_1,\dots,d_k\geqslant 1\\ r_i|d_i \text{ for all } i}} L(d_1,\dots,d_k).$$

We use this much as above, in the first instance with

$$L(d_1,\dots,d_k) = \frac{\lambda(d_1,\dots,d_k)}{d_1,\dots,d_k} \text{ and } Y(r_1,\dots,r_k) = \frac{y(r_1,\dots,r_k)}{\phi_k(r_1\dots r_k)}$$

where

$$y(r_1,\ldots,r_k)=F\left(\frac{\log r_1}{\log R},\ldots,\frac{\log r_k}{\log R}\right)$$

with $F\in\mathbb{C}[t_1,\ldots,t_k]$, such that that there is a uniform bound on all of the first order partial derivatives, and F is only supported on

$$T_k:=\{(t_1,\ldots,t_k):\ \text{Each } t_j\geqslant 1,\ \text{and } t_1+\ldots+t_k\leqslant 1\}.$$

Proceeding much as before we obtain

$$S_1\sim\sum_{r_1,\ldots,r_k\geqslant 1}\frac{y(r_1,\ldots,r_k)^2}{\phi_k(r_1\ldots r_k)}. \tag{6.1}$$

6.3. **Sums of multiplicative functions.** By (4.8) we have

$$\sum_{\substack{1\leqslant n\leqslant N\\(n,m)=1}}\frac{\mu^2(n)}{\phi_k(n)}=\prod_{p|m}\frac{p-1}{p}\prod_{p\nmid m}\frac{(p-1)\phi_{k-1}(p)}{p\ \phi_k(p)}\cdot(\log N+O(1)) \tag{6.2}$$

We apply this k times; firstly with m replaced by $mr_1\ldots r_{k-1}$ and n by r_k, then with m replaced by $mr_1\ldots r_{k-2}$, etc By the end we obtain

$$C_m(a)\sum_{\substack{1\leqslant r_1\leqslant R_1,\\ \ldots,\\ 1\leqslant r_k\leqslant R_k}}\frac{\mu^2(r_1\ldots r_k m)}{\phi_k(r_1,\ldots,r_k)}=\prod_i(\log R_i+O(1)), \tag{6.3}$$

where

$$C_m(a):=\prod_{p|m}\left(1-\frac{1}{p}\right)^{-k}\prod_{p\nmid m}\left(1-\frac{k}{p}\right)\left(1-\frac{1}{p}\right)^{-k}.$$

From this, and partial summation, we deduce from (6.1) , that

$$C_m(a)S_1\sim(\log R)^k\cdot\int_{t_1,\ldots,t_k\in T_k}F(t_1,\ldots,t_k)^2dt_k\ldots dt_1. \tag{6.4}$$

Had we stopped our calculation one step earlier we would have found

$$C_m(a)\sum_{\substack{1\leqslant r_1\leqslant R_1,\\ \ldots,\\ 1\leqslant r_{k-1}\leqslant R_{k-1}}}\frac{\mu^2(r_1\ldots r_{k-1}m)}{\phi_k(r_1,\ldots,r_{k-1})}=\frac{m}{\phi(m)}\cdot\prod_i(\log R_i+O(1)), \tag{6.5}$$

6.4. **The combinatorics, II.** We will deal only with the case $\ell=k$, the other cases being analogous. Now we use the higher dimensional reciprocity law with

$$L(d_1,\ldots,d_{k-1})=\frac{\lambda(d_1,\ldots,d_{k-1},1)}{\phi(d_1\ldots d_{k-1})}\ \text{ and }\ Y_k(r_1,\ldots,r_{k-1})=\frac{y_k(r_1,\ldots,r_{k-1})}{\phi_k(r_1\ldots r_{k-1})}$$

where $d_k=r_k=1$, so that, with the exactly analogous calculations as before,

$$S_{2,k}\sim\sum_{r_1,\ldots,r_{k-1}\geqslant 1}\frac{y_k(r_1,\ldots,r_{k-1})^2}{\phi_k(r_1\ldots r_{k-1})}.$$

Using the reciprocity law twice to determine the $y_k(r)$ in terms of the $y(n)$, we obtain that

$$y_k(r_1,\ldots,r_{k-1}) \sim \frac{\phi(m)}{m}\cdot\int_{t\geqslant 0} F(\rho_1,\ldots,\rho_{k-1},t)dt\cdot\log R$$

where each $r_i = N^{\rho_i}$. Therefore, using (6.5), we obtain
(6.6)

$$C_m(a)S_{2,k} \sim \int_{0\leqslant t_1,\ldots,t_{k-1}\leqslant 1}\left(\int_{t_k\geqslant 0} F(t_1,\ldots,t_{k-1},t_k)dt_k\right)^2 dt_{k-1}\ldots dt_1\cdot\frac{\phi(m)}{m}(\log R)^{k+1}.$$

6.5. **Finding a positive difference.** By the Bombieri-Vinogradov Theorem we can take $R = x^{1/4-o(1)}$, so that, by (6.4) and (6.6) , $C_m(a)S_0$ equals $\frac{\omega(m)}{m}x(\log 3x)(\log R)^k$ times

$$\frac{1}{4}\sum_{\ell=1}^{k}\int_{\substack{0\leqslant t_i\leqslant 1 \text{ for}\\ 1\leqslant i\leqslant k,\ \ i\neq\ell}}\left(\int_{t_\ell\geqslant 0} F(t_1,\ldots,t_k)dt_\ell\right)^2\prod_{\substack{1\leqslant j\leqslant k\\ i\neq\ell}} dt_j - h\int_{t_1,\ldots,t_k\in T_k} F(t_1,\ldots,t_k)^2 dt_k\ldots dt_1 + o(1).$$

One can show that the optimal choice for F must be symmetric. Hence $S_0 > 0$ follows if there exists a symmetric F (with the restrictions above) for which the ratio

$$\rho(F) := \frac{k\int_{t_1,\ldots,t_{k-1}\geqslant 0}\left(\int_{t_k\geqslant 0} F(t_1,\ldots,t_k)dt_k\right)^2 dt_{k-1}\ldots dt_1}{\int_{t_1,\ldots,t_k\geqslant 0} F(t_1,\ldots,t_k)^2 dt_k\ldots dt_1}.$$

satisfies $\rho(F) > 4h$.

Proposition 6.1. *Fix $h \geqslant 1$. Suppose that there exists $F \in \mathbb{C}(x_1,\ldots,x_k)$ which is measurable, supported on T_k, for which there is a uniform bound on the first order partial derivatives and such that $\rho(F) > 4h$. Then, for every admissible k-tuple of linear forms, there are infinitely many integers n such that there are $> h$ primes amongst the k linear forms when evaluated at n. If the Elliott-Halberstam conjecture holds then we only need that $\rho(F) > 2h$.*

6.6. **A special case.** If $F(t_1,\ldots,t_k) = f(t_1+\ldots+t_k)$ then since

$$\int_{\substack{t_1,\ldots,t_k\geqslant 0\\ t_1+\ldots+t_k=t}} dt_{k-1}\ldots dt_1 = \frac{t^{k-1}}{(k-1)!},$$

we deduce that

$$\rho(F) = \rho_k(f)$$

as defined in (4.12); that is, reduce to the original GPY argument.

We need to make some choices for F that do not lead back to the original GPY argument, in the hope that we can do better; evidently we should avoid selecting F to be a function of one variable. Since F is symmetric it makes sense to define the symmetric sums as $P_j = \sum_{i=1}^k t_i^j$; in the GPY argument F was a function of P_1. A first guess might be to work now with functions of P_1 and P_2, so as to consider functions F that do not appear in the GPY argument.

6.7. **Maynard's Fs, and gaps between primes.** For $k=5$ let

$$F(t_1,\dots,t_5)=70P_1P_2-49P_1^2-75P_2+83P_1-34.$$

A calculation yields that

$$\rho(F)=\frac{1417255}{708216}>2.$$

Therefore, by Proposition (6.1), if we assume the Elliott-Halberstam conjecture with $h=1$ then for every admissible 5-tuple of linear forms, there are infinitely many integers n such that there are at least two primes amongst the five linear forms when evaluated at n. In particular, from the admissible forms $\{x,x+2,x+6,x+8,x+12\}$ we deduce that there are infinitely many pairs of distinct primes that differ by no more than 12. Also from the admissible forms $\{x+1,2x+1,4x+1,8x+1,16x+1\}$ we deduce that there are infinitely many pairs of distinct primes, p,q for which $(p-1)/(q-1)=2^j$ for $j=0,1,2,3$ or 4.

Unconditionally, Maynard shows that there exists a polynomial of the form

$$\sum_{\substack{a,b\geqslant 0\\ a+2b\leqslant 11}} c_{a,b}(1-P_1)^aP_2^b$$

with $k=105$, for which

$$\rho(F)=4.0020697\dots$$

How does Maynard prove this? With F of the above form, one sees that both the numerator and denominator of $\rho(F)$ are quadratic forms in the variables $c_{a,b}$. There are 42 such coefficients, and we let a be the vector of c-values. Therefore there exist easily calculable matrices M_1 and M_2 for which the numerator of F is a^TM_2a, and the denominator is a^TM_1a. By the theory of Lagrangian multipliers, Maynard shows that

$$M_1^{-1}M_2a=\rho(F)a$$

so that $\rho(f)$ can be taken to be the largest eigenvalue of $M_1^{-1}M_2$, and a the corresponding eigenvector. These calculations are easily completed using a computer algebra package and yield the result above.

By Proposition 6.1 with $h=1$, we deduce that for every admissible 105-tuple of linear forms, there are infinitely many integers n such that there are at least two primes amongst the 105 linear forms when evaluated at n.

6.8. **F as a product of one dimensional functions.** We make the choice that

$$F(t_1,\dots t_k)=\begin{cases} g(kt_1)\dots g(kt_k) & \text{if } t_1+\dots+t_k\leqslant 1\\ 0 & \text{otherwise,}\end{cases}$$

where g is some integrable function supported only on $[0,T]$. Let $\gamma:=\int_{t\geqslant 0}g(t)^2dt$, so that the denominator of $\rho(F)$ is

$$I_k=\int_{t\in T_k} f(t_1,\dots t_k)^2dt_k\dots dt_1\leqslant\int_{t_1,\dots,t_k\geqslant 0}(g(kt_1)\dots g(kt_k))^2dt_k\dots dt_1=k^{-k}\gamma^k.$$

We rewrite the numerator of $\rho(F)$ as $L_k - M_k$ where

$$L_k := k\int_{t_1,\ldots,t_{k-1}\geqslant 0}\left(\int_{t_k\geqslant 0} g(kt_1)\ldots g(kt_k)dt_k\right)^2 dt_{k-1}\ldots dt_1 = k^{-k}\gamma^{k-1}\left(\int_{t\geqslant 0} g(t)dt\right)^2.$$

As $g(t)$ is only supported in $[0,T]$ we have, by Cauchying and letting $u_j = kt_j$,

$$\begin{aligned} M_k :&= \int_{t_1,\ldots,t_{k-1}\geqslant 0}\left(\int_{t_k\geqslant 1-t_1-\ldots-t_{k-1}} g(kt_1)\ldots g(kt_k)dt_k\right)^2 dt_{k-1}\ldots dt_1 \\ &\leqslant k^{-k}T\int_{\substack{u_1,\ldots,u_k\geqslant 0\\ u_1+\ldots+u_k\geqslant k}} g(u_1)^2\ldots g(u_k)^2 du_1\ldots du_k. \end{aligned}$$

Now assume that $\mu := \int_t tg(t)^2dt \leqslant (1-\eta)\int_t g(t)^2dt = (1-\eta)\gamma$ for some given $\eta > 0$; that is, that the "weight" of g^2 is centered around values of $t \leqslant 1-\eta$. We have

$$1 \leqslant \eta^{-2}\left(\frac{1}{k}(u_1+\ldots+u_k) - \mu/\gamma\right)^2$$

whenever $u_1+\ldots+u_k \geqslant k$. Therefore,

$$\begin{aligned} M_k &\leqslant \eta^{-2}k^{-k}T\int_{u_1,\ldots,u_k\geqslant 0} g(u_1)^2\ldots g(u_k)^2\left(\frac{1}{k}(u_1+\ldots+u_k) - \mu/\gamma\right)^2 du_1\ldots du_k \\ &= \eta^{-2}k^{-k-1}T\int_{u_1,\ldots,u_k\geqslant 0} g(u_1)^2\ldots g(u_k)^2(u_1^2 - \mu^2/\gamma^2)du_1\ldots du_k \\ &= \eta^{-2}k^{-k-1}\gamma^{k-1}T\left(\int_{u\geqslant 0} u^2g(u)^2du - \mu^2/\gamma\right) \leqslant \eta^{-2}k^{-k-1}\gamma^{k-1}T\int_{u\geqslant 0} u^2g(u)^2du, \end{aligned}$$

by symmetry. We deduce that

$$\rho(F) \geqslant \frac{\left(\int_{t\geqslant 0} g(t)dt\right)^2 - \frac{\eta^{-2}T}{k}\int_{u\geqslant 0} u^2g(u)^2du}{\int_{t\geqslant 0} g(t)^2dt}. \tag{6.7}$$

Notice that we can multiply g through by a scalar and not effect the value in (6.7).

6.9. **The optimal choice.** We wish to find the value of g that maximizes the right-hand side of (6.7). This can be viewed as an optimization problem:

Maximize $\int_{t\geqslant 0} g(t)dt$, subject to the constraints $\int_{t\geqslant 0} g(t)^2dt = \gamma$ and $\int_{t\geqslant 0} tg(t)^2dt = \mu$.

One can approach this using the calculus of variations or even by discretizing g and employing the technique of Lagrangian multipliers. The latter gives rise to (a discrete form of)

$$\int_{t\geqslant 0} g(t)dt - \alpha\left(\int_{t\geqslant 0} g(t)^2dt - \gamma\right) - \beta\left(\int_{t\geqslant 0} tg(t)^2dt - \mu\right),$$

for unknowns α and β. Differentiating with respect to $g(v)$ for each $v \in [0,T]$, we obtain

$$1 - 2\alpha g(v) - 2\beta vg(v) = 0;$$

that is, after re-scaling,

$$g(t) = \frac{1}{1 + At} \quad \text{for} \quad 0 \leqslant t \leqslant T,$$

for some real $A > 0$. We select T so that $1 + AT = e^A$, and let $A > 1$. We then calculate the integrals in (6.7):

$$\begin{aligned} \gamma = \int_t g(t)^2 dt \ &= \frac{1}{A}(1 - e^{-A}), \\ \int_t tg(t)^2 dt &= \frac{1}{A^2}\left(A - 1 + e^{-A}\right), \\ \int_t t^2 g(t)^2 dt &= \frac{1}{A^3}\left(e^A - 2A - e^{-A}\right), \\ \text{and} \qquad \int_t g(t) dt \ &= 1, \\ \text{so that} \qquad \eta \ &= \frac{1 - (A-1)e^{-A}}{A(1 - e^{-A})} \ > 0, \end{aligned}$$

which is necessary. (6.7) then becomes

$$(6.8) \ \ \rho(F) \geqslant \frac{A}{(1 - e^{-A})} - \frac{e^{2A}}{Ak}\left(1 - 2Ae^{-A} - e^{-2A}\right)\frac{(1 - e^{-A})^2}{(1 - (A-1)e^{-A})^2} \geqslant A - \frac{e^{2A}}{Ak}$$

Taking $A = \frac{1}{2}\log k + \frac{1}{2}\log\log k$, we deduce that

$$\rho(F) \geqslant \frac{1}{2}\log k + \frac{1}{2}\log\log k - 2.$$

Hence, for every $m \geqslant 1$ we find that $\rho(F) > 4m$ provided $e^{8m+4} < k \log k$.

This implies the following result:

Theorem 6.2. *For any given integer $m \geqslant 2$, let k be the smallest integer with $k \log k > e^{8m+4}$. For* any *admissible k-tuple of linear forms $L_1, \ldots, L_k$ there exists infinitely many integers n such that at least m of the $L_j(n)$, $1 \leqslant j \leqslant k$ are prime.*

For any $m \geqslant 1$, we let k be the smallest integer with $k \log k > e^{8m+4}$, so that $k > 10000$; in this range it is known that $\pi(k) \leqslant \frac{k}{\log k - 4}$. Next we let $x = 2k \log k > 10^5$ and, for this range it is known that $\pi(x) \geqslant \frac{x}{\log x}(1 + \frac{1}{\log x})$. Hence

$$\pi(2k \log k) - \pi(k) \geqslant \frac{2k \log k}{\log(2k \log k)}\left(1 + \frac{1}{\log(2k \log k)}\right) - \frac{k}{\log k - 4}$$

and this is $> k$ for $k \geqslant 311$ by an easy calculation. We therefore apply the theorem with the k smallest primes $> k$, which form an admissible set $\subset [1, 2k \log k]$, to obtain:

Corollary 6.3. *For any given integer $m \geqslant 2$, let $B_m = e^{8m+5}$. There are infinitely many integers x for which there are at least m distinct primes within the interval $[x, x + B_m]$.*

By a slight modification of this construction, Maynard obtains $B_m \ll m^3 e^{4m}$ in **[29]**.

Part 2. Primes in arithmetic progressions; breaking the $\sqrt{x}$-barrier

Our goal, in the rest of the article, is to sketch the ideas behind the proof of Yitang's extraordinary result, given in (3.5) , that primes are well-distributed on average in the arithmetic progressions $a \pmod q$ with q a little bigger than $\sqrt{x}$. We will see how this question fits into a more general framework, as developed by Bombieri, Friedlander and Iwaniec [**3**], so that Zhang's results should also allow us to deduce analogous results for interesting arithmetic sequences other than the primes.

To begin with we will need to discuss a key technique of analytic number theory, the idea of creating important sequences through convolutions:

7. Convolutions in number theory

The convolution of two functions f and g, written $f * g$, is defined by

$$(f * g)(n) := \sum_{ab=n} f(a)g(b),$$

for every integer $n \geqslant 1$, where the sum is over all pairs of positive integers a, b whose product is n. Hence if $\tau(n)$ counts the number of divisors of n then

$$\tau = 1 * 1,$$

where 1 is the function with $1(n) = 1$ for every $n \geqslant 1$. We already saw, in (2.2), that if $L(n) = \log n$ then $\mu * L = \Lambda$, where $\Lambda(n) = \log p$ if n is a power of prime p, and $\Lambda(n) = 0$ otherwise. In the GPY argument we used that $(1 * \mu)(n) = 0$ if $n > 1$.

There is no better way to understand why convolutions are useful than to present a famous argument of Dirichlet, estimating the average of $\tau(n)$. Now , if n is squarefree and has k prime factors then $\tau(n) = 2^k$, so we see that $\tau(n)$ varies greatly depending on the arithmetic structure of n, but the average is more stable:

$$\frac{1}{x}\sum_{n\leqslant x}\tau(n) = \frac{1}{x}\sum_{n\leqslant x}\sum_{d|n} 1 = \frac{1}{x}\sum_{d|n}\sum_{\substack{n\leqslant x\\ d|n}} 1 = \frac{1}{x}\sum_{d\leqslant x}\left[\frac{x}{d}\right]$$

$$= \frac{1}{x}\sum_{d\leqslant x}\Big(\frac{x}{d} + O(1)\Big) = \sum_{d\leqslant x}\frac{1}{d} + O\left(\frac{1}{x}\sum_{d\leqslant x} 1\right).$$

One can approximate $\sum_{d\leqslant x}\frac{1}{d}$ by $\int_1^x dt/t = \log x$. Indeed the difference tends to a limit, the Euler-Mascheroni constant $\gamma := \lim_{N\to\infty} \frac{1}{1} + \frac{1}{2} + \ldots + \frac{1}{N} - \log N$. Hence we have proved that the integers up to x have $\log x + O(1)$ divisors, on average, which is quite remarkable for such a wildly fluctuating function.

Dirichlet studied this argument and noticed that when we approximate $[x/d]$ by $x/d + O(1)$ for large d, say for those d in $(x/2, x]$, then this is not really a very good approximation, and gives a large cumulative error term, $O(x)$. However we know that $[x/d] = 1$ for each of these d, and so we can estimate this sum by $x/2 + O(1)$, which is much more precise. In general we write $n = dm$, where d and m are integers. When d is small then we should fix d, and count the number of such m, with $m \leqslant x/d$ (as we did above); but when m is small, then we should fix m, and count the number of d with $d \leqslant x/m$. In this way our sums are all over long intervals, which allows us to get an accurate approximation of their value:

$$\begin{aligned}\frac{1}{x}\sum_{n\leqslant x}\tau(n) &= \frac{1}{x}\sum_{n\leqslant x}\sum_{dm=n} 1 = \frac{1}{x}\sum_{d\leqslant\sqrt{x}}\sum_{\substack{n\leqslant x\\ d|n}} 1 + \frac{1}{x}\sum_{m<\sqrt{x}}\sum_{\substack{n\leqslant x\\ m|n}} 1 - \frac{1}{x}\sum_{d\leqslant\sqrt{x}}\sum_{m<\sqrt{x}} 1\\ &= \frac{1}{x}\sum_{d\leqslant\sqrt{x}}\Big(\frac{x}{d}+O(1)\Big) + \frac{1}{x}\sum_{m<\sqrt{x}}\Big(\frac{x}{m}+O(1)\Big) - 1 + O\left(\frac{1}{\sqrt{x}}\right)\\ &= \log x + 2\gamma - 1 + O\left(\frac{1}{\sqrt{x}}\right),\end{aligned}$$

since $\sum_{n\leqslant N} 1/n = \log N + \gamma + O(1/N)$, an extraordinary improvement upon the earlier error term.

7.1. **Vaughan's identity.**

We will need a more convoluted identity than (2.2) to prove our estimates for primes in arithmetic progressions. There are several possible suitable identities, the simplest of which is due to Vaughan [**40**]:

(7.1) *Vaughan's identity*: $\quad \Lambda_{\geqslant V} = \mu_{<U} * L - \mu_{<U} * \Lambda_{<V} * 1 + \mu_{\geqslant U} * \Lambda_{\geqslant V} * 1$

where $g_{>W}(n) = g(n)$ if $n > W$ and $g(n) = 0$ otherwise; and $g = g_{\leqslant W} + g_{>W}$. To verify this identity, we manipulate the algebra of convolutions:

$$\begin{aligned}\Lambda_{\geqslant V} &= \Lambda - \Lambda_{<V} = (\mu * L) - \Lambda_{<V} * (1 * \mu)\\ &= \mu_{<U} * L + \mu_{\geqslant U} * L - \mu_{<U} * \Lambda_{<V} * 1 - \mu_{\geqslant U} * \Lambda_{<V} * 1\\ &= \mu_{<U} * L - \mu_{<U} * \Lambda_{<V} * 1 + \mu_{\geqslant U} * (\Lambda * 1 - \Lambda_{<V} * 1),\end{aligned}$$

8. Distribution in arithmetic progressions

8.1. **General sequences in arithmetic progressions.**

One can ask whether *any* given sequence $(\beta(n))_{n\geqslant 1} \in \mathbb{C}$ is well-distributed in arithmetic progressions modulo q. We begin by formulating an appropriate analogy to (3.2), which should imply non-trivial estimates in the range $q \leqslant (\log x)^A$ for any fixed $A > 0$: We say that β satisfies a *Siegel-Walfisz condition* if, for any fixed $A > 0$, and whenever $(a, q) = 1$, we have

$$\left|\sum_{\substack{n\leqslant x\\ n\equiv a \pmod q}} \beta(n) - \frac{1}{\phi(q)}\sum_{\substack{n\leqslant x\\ (n,q)=1}} \beta(n)\right| \ll_A \frac{\|\beta\| x^{\frac{1}{2}}}{(\log x)^A},$$

with $\|\beta\| = \|\beta\|_2$ where, as usual,

$$\|\beta\|_2 := \left(\sum_{n\leqslant x} |\beta(n)|^2\right)^{\frac{1}{2}}.$$

Using Cauchy's inequality one can show that this assumption is "non-trivial" only for $q < (\log x)^{2A}$; that is, when x is very large compared to q.

Using the large sieve, Bombieri, Friedlander and Iwaniec **[3]** were able to prove two results that are very surprising, given the weakness of the hypotheses. In the first they showed that if β satisfies a Siegel-Walfisz condition,[11] then it is well-distributed for *almost all* arithmetic progressions $a \pmod q$, for *almost all* $q \leqslant x/(\log x)^B$:

Theorem 8.1. *Suppose that the sequence of complex numbers $\beta(n), n \leqslant x$ satisfies a Siegel-Walfisz condition. For any $A > 0$ there exists $B = B(A) > 0$ such that*

$$\sum_{q\leqslant Q} \ \sum_{a:\ (a,q)=1} \left| \sum_{n\equiv a \pmod q} \beta(n) - \frac{1}{\phi(q)} \sum_{(n,q)=1} \beta(n) \right|^2 \ll \|\beta\|^2 \frac{x}{(\log x)^A}$$

where $Q = x/(\log x)^B$.

The analogous result for $\Lambda(n)$ is known as the *Barban-Davenport-Halberstam theorem* and in that special case one can even obtain an asymptotic.

Before proceeding, let us assume, for the rest of this article, that we are given two sequences of complex numbers as follows:

- $\alpha(m),\ M < m \leqslant 2M$ and $\beta(n),\ N < n \leqslant 2N$, with $x^{1/3} < N \leqslant M \leqslant x^{2/3}$ and $MN \leqslant x$.
- $\beta(n)$ satisfies the Siegel-Walfisz condition.
- $\alpha(m) \ll \tau(m)^A(\log x)^B$ and $\beta(n) \ll \tau(n)^A(\log x)^B$ (there inequalities are satisfied by $\mu, 1, \Lambda, L$ and any convolutions of these sequences).

In their second result, Bombieri, Friedlander and Iwaniec, showed that rather general convolutions are well-distributed[12] for *all* arithmetic progressions $a \pmod q$, for *almost all* $q \leqslant x^{1/2}/(\log x)^B$.

Theorem 8.2. *Suppose that $\alpha(m)$ and $\beta(n)$ are as above. For any $A > 0$ there exists $B = B(A) > 0$ such that*

$$\sum_{q\leqslant Q} \max_{a:\ (a,q)=1} \left| \sum_{n\equiv a \pmod q} (\alpha * \beta)(n) - \frac{1}{\phi(q)} \sum_{(n,q)=1} (\alpha * \beta)(n) \right| \ll \|\alpha\|\|\beta\| \frac{x^{1/2}}{(\log x)^A}$$

where $Q = x^{1/2}/(\log x)^B$.

This allowed them to give a proof of the Bombieri-Vinogradov theorem for primes, using Vaughan's identity (7.1), that seems to be less dependent on very specific properties of the primes. The subject, though, has long been stuck with the bound $x^{1/2}$ on the moduli.[13]

[11] Their condition appears to be weaker than that assumed here, but it can be shown to be equivalent.

[12] This possibility has its roots in a paper of Motohashi **[31]**.

[13] There had been some partial progress with moduli $> x^{1/2}$, as in **[4]** , but no upper bounds which "win" by an arbitrary power of $\log x$ (which is what is essential to applications).

Bombieri, Friedlander and Iwaniec [**3**] made the following conjecture, and noted that in many applications, it suffices to work with a fixed (as is true in the application here).

Conjecture 8.3. *Suppose that $\alpha(m)$ and $\beta(n)$ are as above. For any $A, \epsilon > 0$, and every integer a, we have*

$$\sum_{\substack{q\leqslant Q\\ (q,a)=1}} \left| \sum_{n\equiv a \pmod q} (\alpha*\beta)(n) - \frac{1}{\phi(q)} \sum_{(n,q)=1} (\alpha*\beta)(n) \right| \ll \|\alpha\|\|\beta\| \frac{x^{1/2}}{(\log x)^A}$$

where $Q = x^{1-\epsilon}$.

The extraordinary work of Zhang breaks through the $\sqrt{x}$ barrier in some generality, working with moduli slightly larger than $x^{1/2}$, though his moduli are y-smooth, with $y = x^\delta$. The key result is as follows:

Theorem 8.4. *Suppose that $\alpha(m)$ and $\beta(n)$ are as above. There exist constants $\eta, \delta > 0$ such, for any $A > 0$, for any integer a,*

$$\sum_{\substack{q\leqslant Q\\ P(q)\leqslant x^\delta\\ (q,a)=1\\ q \text{ squarefree}}} \left| \sum_{\substack{n\leqslant x\\ n\equiv a \pmod q}} (\alpha*\beta)(n) - \frac{1}{\phi(q)} \sum_{\substack{n\leqslant x\\ (n,q)=1}} (\alpha*\beta)(n) \right| \ll_A \|\alpha\|\|\beta\| \frac{x^{1/2}}{(\log x)^A}$$

where $Q = x^{1/2+\eta}$.

We then deduce the same result but now for α and β supported in $x^{1/3} < m, n \leqslant x^{2/3}$ with $mn \leqslant x$, by dissecting this range up into into dyadic ranges (that is, $M < m \leqslant 2M$ and $N < n \leqslant 2N$) and smaller ranges, as well as possible, and then carefully accounting for the (m, n) pairs missed.

8.2. **The deduction of the main theorem for primes.** We will bound each term that arises from Vaughan's identity, (7.1) , with $U = V = x^{1/3}$, rewritten as

$$\Lambda = \Lambda_{<x^{1/3}} + \mu_{<x^{1/3}}*L - (\mu*\Lambda)_{<x^{1/3}}*1_{\geqslant x^{2/3}} - \mu_{<x^{1/3}}*\Lambda_{<x^{1/3}}*1_{<x^{2/3}} + \mu_{\geqslant x^{1/3}}*\Lambda_{\geqslant x^{1/3}}*1.$$

The first term is acceptably small, simply by taking absolute values. For the second term we write $(\mu_{<x^{1/3}} * L)(n) = \sum_{um=n,\ u<x^{1/3}} \mu(u) \log m$, to obtain the difference

$$\sum_{\substack{u<x^{1/3}\\ (u,q)=1}} \mu(u) \left(\sum_{\substack{x/u<m\leqslant 2x/u\\ m\equiv a/u \pmod q}} \log m - \frac{1}{\phi(q)} \sum_{\substack{x/u<m\leqslant 2x/u\\ (m,q)=1}} \log m \right)$$

Writing $M = x/u$, the inner sum is the difference between the sum of $\log m$ in $(M, 2M]$ over an arithmetic progression $b \pmod q$ with $(b, q) = 1$, minus the average of such sums. Now if $n_- = [M/q]$ and $n_+ = [2M/q]$, then, since $\log q[m/q] < \log m < \log q([m/q] + 1)$, such a sum is $> \sum_{n_-\leqslant n\leqslant n_+-1} \log qn$ and is $< \sum_{n_-+1\leqslant n\leqslant n_+1} \log qn$. The difference between these bounds in $\ll \log M$, and hence this is our bound on the term in parentheses. Summing over u yields a bound that is acceptably small.

We deal with the third term, by the same argument as just above, since we obtain an inner sum of 1, over the values of m in an interval of an arithmetic progression; and then we obtain a bound that is acceptably small.

We are left to work with two sums of convolutions:

$$\sum_{\substack{mn\asymp x\\ mn\equiv a \pmod q}} (\mu_{<x^{1/3}}*\Lambda_{<x^{1/3}})(m)1_{<x^{2/3}}(n) \text{ and } \sum_{\substack{mn\asymp x\\ mn\equiv a \pmod q}} (\Lambda_{\geqslant x^{1/3}}*1)(m)\mu_{\geqslant x^{1/3}}(n),$$

where $x^{1/3} \ll m, n \ll x^{2/3}$, and each convolution takes the form $\alpha(m)\beta(n)$ with $\alpha(m)$ and $\beta(n)$ as above. The result then follows from Zhang's result as discussed at the end of the last subsection.

8.3. **Further reductions.** We reduce Theorem (8.4) further. The first observation is that we can restrict our moduli to those with $< C\log\log x$ prime factors, for some large $C > 0$, since the moduli with more prime factors are rare and thus contribute little to the sum. Since the moduli are y-smooth, they can be factored as qr where $N/(yx^\epsilon) < r \leqslant N/x^\epsilon$. Since the modulus does not have a lot of prime factors, one can deduce that the smallest prime factor of q, denoted $p(q)$, is $\geqslant D_0 := x^{\epsilon/\log\log x}$. Hence we may also now assume

- $r \in (R, 2R]$ with $P(r) \leqslant y$ with $y := x^\delta$.
- $q \in (Q, 2Q]$ with $D_0 < p(q) \leqslant P(q) \leqslant y$.
- $N/(yx^\epsilon) < R \leqslant N/x^\epsilon$ and $x^{1/2}/(\log x)^B < QR \leqslant x^{1/2+\eta}$

In **[34]**, some gains are made by working instead with the full set of moduli that have this kind of convenient factorizations, rather than restrict attention just to those moduli which are y-smooth.

We begin by noting that

$$\sum_{n\equiv a \pmod{qr}} \gamma(n) - \frac{1}{\phi(qr)}\sum_{(n,qr)=1}\gamma(n) =$$

$$\sum_{n\equiv a \pmod{qr}} \gamma(n) - \frac{1}{\phi(q)}\sum_{\substack{(n,q)=1\\ n\equiv a \pmod r}}\gamma(n) + \frac{1}{\phi(q)}\left(\sum_{\substack{(n,q)=1\\ n\equiv a \pmod r}}\gamma(n) - \frac{1}{\phi(r)}\sum_{\substack{(n,q)=1\\ (n,r)=1}}\gamma(n)\right)$$

with $\gamma = \alpha * \beta$. We sum the absolute value of these terms, over the moduli $d \in [D, 2D]$, factored into qr as above. Since $\beta(n)$ satisfies the Siegel-Walfisz criterion, we can deduce that $\beta(n)1_{(n,q)=1}$ also satisfies it, and therefore Theorem (8.2) is applicable for $\alpha(m) * \beta(n)1_{(n,q)=1}$; this allows us to bound the sum of the second terms here, suitably. Hence it remains to prove

$$\sum_{\substack{q\in[Q,2Q]\\ D_0<p(q)\leqslant P(q)\leqslant y}} \sum_{\substack{r\in[R,2R],\\ P(r)\leqslant y\\ qr \text{ squarefree}}} \left| \sum_{\substack{n\equiv a \pmod r\\ n\equiv b \pmod q}} (\alpha*\beta)(n) - \sum_{\substack{n\equiv a \pmod r\\ n\equiv b' \pmod q}} (\alpha*\beta)(n) \right| \ll_A \|\alpha\|\|\beta\| \frac{x^{1/2}}{(\log x)^A}, \tag{8.1}$$

for any integers a, b, b' with $p(abb') > y$.

9. Removing the weights, and an unweighted arithmetic progression

9.1. **Removing the weights.** The sums in (8.1) are complicated, and the innermost sum is over an unknown function $\alpha * \beta$. In this section we use Cauchy's inequality to "unfold" the sum, so as to remove the weight from the innermost sum:

In the left-hand side of (8.1) we replace the absolute value in the (q, r) term by a complex number $c_{q,r}$ of absolute value 1, to obtain, after a little re-arranging:

$$\sum_r \sum_m \alpha(m) \left(\sum_q \sum_{n:\ mn\equiv a \pmod r} c_{q,r}\beta(n)(1_{mn\equiv b \pmod q} - 1_{mn\equiv b' \pmod q}) \right).$$

By the Cauchy-Schwarz inequality the square of this is

$$\leqslant \sum_r \sum_m |\alpha(m)|^2 \leqslant R\|\alpha\|^2$$

times

$$\sum_r \sum_m \left| \sum_q \sum_{n:\ mn\equiv a \pmod r} c_{q,r}\beta(n)(1_{mn\equiv b \pmod q} - 1_{mn\equiv b' \pmod q}) \right|^2. \tag{9.1}$$

When we expand the square, we obtain the sum of four terms of the form

$$\pm \sum_r \sum_m \sum_{q_1,q_2} \sum_{\substack{n_1,n_2\\ mn_1\equiv mn_2\equiv a \pmod r}} c_{q_1,r}\overline{c_{q_2,r}}\beta(n_1)\overline{\beta(n_2)} 1_{mn_1\equiv b_1 \pmod{q_1}} 1_{mn_2\equiv b_2 \pmod{q_2}}$$

$$= \pm \sum_r \sum_{q_1,q_2} \sum_{\substack{n_1,n_2\\ n_1\equiv n_2 \pmod r}} c_{q_1,r}\overline{c_{q_2,r}}\beta(n_1)\overline{\beta(n_2)} \cdot \sum_m 1_{\substack{m\equiv b_1/n_1 \pmod{q_1}\\ m\equiv b_2/n_2 \pmod{q_2}\\ m\equiv a/n_1 \pmod r}} \tag{9.2}$$

where we get "+" when $b_1 = b_2 = b$ or b', and "−" otherwise, since $(mn, qr) = 1$.

We have achieved our goal of having an unweighted innermost sum. Indeed, if it is non-zero,[14] then it is just the number of integers in an interval of an arithmetic progression with common difference $r[q_1, q_2]$.

[14] This sum cannot possibly contain any integers, and so is 0, if the congruences are incompatible. Since $(r, q_1q_2) = 1$ they are compatible unless $b_1/n_1 \equiv b_2/n_2 \pmod{(q_1, q_2)}$. Note that this criterion is irrelevant if $(q_1, q_2) = 1$.

9.2. **The main terms.** The number of integers in an interval of length M, from an arithmetic progression with common difference $r[q_1, q_2]$ is

$$\frac{M}{r[q_1,q_2]} + O(1).$$

We study now the sum of the "main terms", the $M/r[q_1, q_2]$. Firstly, for the terms with $(q_1, q_2) = 1$ the main terms sum to

$$\pm \sum_r \sum_{\substack{q_1,q_2 \\ (q_1,q_2)=1}} \sum_{\substack{n_1,n_2 \\ n_1 \equiv n_2 \pmod r}} c_{q_1,r}\overline{c_{q_2,r}}\beta(n_1)\overline{\beta(n_2)} \cdot \frac{M}{rq_1q_2},$$

which is independent of the values of b_1, b_2 and hence cancel, when we sum over the four terms (and the two '+', and two '−', signs). For the terms with $(q_1, q_2) \neq 1$ we have $(q_1, q_2) \geqslant D_0$ (since the prime factors of the q_i are all $\geqslant D_0$), and it is not difficult to show that these are $\ll x(\log x)^{O(1)}/RD_0$, which is acceptably small.

9.3. **The error terms and the advent of exponential sums.** The "$O(1)$"s in (9.2) can add up to a total that is far too large. One can show that in most of the terms of the sum, the common difference of the arithmetic progression is larger than the length of the interval, so the correct count is either 0 or 1: It is hardly surprising that an error term of "$O(1)$" is too insensitive to help us.

To proceed, instead of approximating, we will give a *precise* formula for the number of integers in an arithmetic progression in an interval, using a sum of exponentials. By the Chinese Remainder Theorem, we can rewrite our triple of congruence conditions

$$m \equiv b_1/n_1 \pmod{q_1},\ m \equiv b_2/n_2 \pmod{q_2},\ m \equiv a/n_1 \pmod r$$

as one,

$$m \equiv m_0(n_1, n_2) \pmod q$$

where $q = rg\ell_1\ell_2$, when there is a solution, which happens if and only if $b_1/n_1 \equiv b_2/n_2 \pmod g$, where $g = (q_1, q_2)$ and we now define $\ell_1 = q_1/g,\ \ell_2 = q_2/g$.

To identify whether m is in a given interval I, we use Fourier analysis. The *discrete Fourier transform* is defined by

$$\hat{f}(h) := \sum_{b \pmod q} f(b)e_q(hb),$$

for any function f of period q. If f is any such function and $I(.)$ is the characteristic function for the interval $(M, 2M]$, then

$$\sum_{m\in I} f(m) = \frac{1}{q} \sum_{h \pmod q} \hat{I}(h)\hat{f}(-h), \tag{9.3}$$

is an example of Plancherel's formula. This has a "main term" at $h = 0$ (which is the same as the main term we found above, in that special case). The coefficients $\hat{I}(h)$ are easily evaluated and bounded:

$$\hat{I}(h) = \sum_{m=M+1}^{2M} e_q(hm) = e_q(2hM) \cdot \frac{e_q(hM) - 1}{e_q(h) - 1}.$$

The numerator has absolute value $\leqslant 2$ and, using the Taylor expansion, the denominator has absolute value $\asymp |h|/q$. Hence

$$|\hat{I}(h)| \ll \min\{M, q/|h|\},$$

We apply (9.3) with $f = \sum_i c_i 1_{m\equiv a_i \pmod q}$, take absolute values, and use our bounds for $|\hat{I}(h)|$, to obtain

$$\left|\sum_i c_i \left(\sum_{\substack{m\asymp M \\ m\equiv a_i \pmod q}} 1 - \frac{M}{q} \right)\right| \ll \sum_{\substack{0\leqslant j\leqslant J \\ H_j:=2^j q/M}} \frac{1}{H_j} \sum_{1\leqslant |h|\leqslant H_j} \left|\sum_i c_i e_q(a_i h)\right| . \tag{9.4}$$

The error terms in (9.2) are bounded by

$$\sum_{r\asymp R}\sum_{g\leqslant G}\sum_{\substack{\ell_1,\ell_2\asymp Q/g \\ (\ell_1,\ell_2)=1}} \left| \sum_{\substack{n_1,n_2\asymp N \\ n_1\equiv n_2 \pmod r \\ b_1/n_1\equiv b_2/n_2 \pmod g}} \beta(n_1)\overline{\beta(n_2)} \cdot \left(\sum_{\substack{m\asymp M \\ m\equiv m_0(n_1,n_2) \pmod{rg\ell_1\ell_2}}} 1 - \frac{M}{rg\ell_1\ell_2} \right) \right|$$

which, by (9.4) , is

$$\ll \sum_{r\asymp R}\sum_{g\leqslant G}\sum_{\substack{\ell_1,\ell_2\asymp Q/g \\ (\ell_1,\ell_2)=1}} \sum_{\substack{0\leqslant j\leqslant J \\ H_j:=2^j G/g}} \frac{1}{H_j} \sum_{1\leqslant |h|\leqslant H_j} \left| \sum_{\substack{n_1,n_2\asymp N \\ n_1\equiv n_2 \pmod r \\ n_2\equiv (b_2/b_1)n_1 \pmod g}} \beta(n_1)\overline{\beta(n_2)} e_{rg\ell_1\ell_2}(m_0(n_1,n_2)h) \right| .$$

We write $n_1 = n,\ n_2 = n + kr$, replace the n_2 variable with k, and define $m_k(n) = m_0(n_1, n_2)$. To simplify matters shall proceed with r, g, k and j fixed, and then sum over these at the end, so we are reduced to studying

$$\sum_{\substack{\ell_1,\ell_2\asymp L \\ (\ell_1,\ell_2)=1}} \frac{1}{H} \sum_{1\leqslant |h|\leqslant H} \left| \sum_{\substack{n\asymp N \\ (b_2-b_1)n\equiv b_1 kr \pmod g}} \beta(n)\overline{\beta(n+kr)} e_{rg\ell_1\ell_2}(m_k(n)h) \right| \tag{9.5}$$

where $L = Q/g$.

10. Linnik's dispersion method

The proof of Zhang's Theorem, and indeed of all the results in the literature of this type, use Linnik's dispersion method. The idea is to express the fact that n belongs to an arithmetic progression using Fourier analysis; summing up over n gives us a main term plus a sum of exponential sums, and then the challenge is to bound each of these exponential sums.

Often the sums come with weights, and judicious use of Cauchying allows one to work with an unweighted, but more complicated exponential sum. We will discuss bounds on exponential sums later in this section. These exponential sums are often *Kloosterman sums*, which one needs to bound. Individual Kloosterman sums can often by suitably bounded by Weil's or Deligne's Theorem. However, sometimes one needs to get good bounds on averages of Kloosterman sums, a question that was brilliantly attacked by Deshouillers and Iwaniec [7] , using the (difficult) spectral theory of automorphic forms. Indeed all previous work, breaking the $\sqrt{x}$ barrier, such as [12] , [3] uses these types of estimates. One of the remarkable aspects of Zhang's work is that he avoids these penible techniques, and the restrictions that come with them.

Zhang was able to use only existing bounds on Kloosterman sums to prove his Theorem, though he does use the sophisticated estimate of Birch and Bombieri from the appendix of [14] . Polymath8 indicates how even this deeper result can be avoided, so that the proof can be given using only "standard" estimates, which is what we do here.

10.1. **Removing the weights again.** To remove the β weights from (9.5) , we begin by replacing the absolute value in (9.5) by the appropriate complex number c_{h,ℓ_1,ℓ_2} of absolute value 1, and re-organize to obtain

(10.1)
$$\sum_{\substack{n\asymp N\\ (b_2-b_1)n\equiv b_1 kr \pmod g}} \beta(n)\overline{\beta(n+kr)} \sum_{\substack{\ell_1,\ell_2\asymp L\\ (\ell_1,\ell_2)=1}} \frac{1}{H}\sum_{1\leqslant |h|\leqslant H} c_{h,\ell_1,\ell_2} e_{rg\ell_1\ell_2}(m_k(n)h).$$

We now Cauchy on the outer sum, which allows us to peel off the β's in the term

$$\sum_n |\beta(n)\beta(n+kr)|^2 \leqslant \sum_n |\beta(n)|^4 = \|\beta\|_4^4,$$

times the more interesting term

$$\sum_n \left| \sum_{\substack{\ell_1,\ell_2\asymp L\\ (\ell_1,\ell_2)=1}} \frac{1}{H}\sum_{1\leqslant |h|\leqslant H} c_{h,\ell_1,\ell_2} e_{rg\ell_1\ell_2}(m_k(n)h)\right|^2 .$$

We simply expand this sum, and take absolute values for each fixed $h, j, \ell_1, \ell_2, m_1, m_2$, to obtain

$$\leqslant \frac{1}{H^2}\sum_{1\leqslant |h|,|j|\leqslant H_i}\ \sum_{\substack{\ell_1,\ell_2,m_1,m_2\asymp L\\ (\ell_1,\ell_2)=(m_1,m_2)=1}} \left| \sum_{\substack{n\asymp N\\ (b_2-b_1)n\equiv b_1 kr \pmod g}} e_{rg\ell_1\ell_2}(m_k(n)h) e_{rgm_1m_2}(-m_k(n)j)\right| .$$

Finally we have pure exponential sums, albeit horribly complicated.

10.2. **Exponential sums with complicated moduli.** If $(r,s)=1$ then there are integers a, b for which

$$ar+bs=1.$$

Note that although there are infinitely many possibilities for the pair of integers a, b, the values of $a \pmod s$ and $b \mod r$ are uniquely defined. If we divide the previous equation by rs, and multiply by m, and then take $e(.)$ of both sides, we obtain

$$e_{rs}(m) = e_s(am)\cdot e_r(bm).$$

This allows us to write the exponential, in our last sum, explicitly. After some analysis, we find that our exponential sum take the form

$$\sum_{\substack{n \asymp N \\ n\equiv a \pmod q}} e_{d_1}\left(\frac{C_1}{n}\right) e_{d_2}\left(\frac{C_2}{n+kr}\right), \tag{10.2}$$

for some constants C_1, C_2 (where $d_1 = rg[\ell_1,\ell_2]$, $d_2 = [m_1,m_2]$ and q divides g) which depend on many variables but are independent of n. With a change of variable $n \to qn+a$ we transform this to another sum of the same shape but instead over all n in an interval.

10.3. **Exponential sums: From the incomplete to the complete.** We now have the sum of the exponential of a function of n, over the integers in an interval. There are typically many integers in this sum, so this is unlike what we encountered earlier (when we were summing 1). The terms of the sum are periodic of period dividing $[d_1, d_2]$ and it is not difficult to sum the terms over a complete period. Hence we can restrict our attention to "incomplete sums" where the sum does not include a complete period.

We can now employ (9.3) once more. The coefficients $\hat{I}(h)$ are well understood, but the $\hat{f}(h)$ now take the form

$$\sum_{n \pmod q} e_{d_1}\left(\frac{C_1}{n}+hn\right) e_{d_2}\left(\frac{C_2}{n+\Delta}+hn\right),$$

a "complete" exponential sum.

The trick here is that we can factor the exponential into its prime factor exponentials and then, by the Chinese Remainder Theorem, this sum *equals* the product over the primes p dividing q, of the same sum but now over $n \pmod p$ with the appropriate $e_p(*)$. Hence we have reduced this question to asking for good bounds on exponential sums of the form

$$\sum_{n \pmod p} e_p\left(\frac{a}{n}+\frac{b}{n+\Delta}+cn\right).$$

Here we omit values of n for which a denominator is 0. As long as this does not degenerate (for example, it would degenerate if $p|a,b,c$) then Weil's Theorem implies that this is $\leqslant \kappa p^{1/2}$, for some constant $\kappa > 0$. Therefore the complete sum over $n \pmod q$ is $\leqslant \kappa^{\nu(q)} q^{1/2}$. This in turn allows us to bound our incomplete sum (10.2) , and to bound the term at the end of the previous section.

The calculations to put this into practice are onerous, and we shall omit these details here. At the end one finds that the bounds deduced are acceptably small if

$$x^{1/2} \geqslant N > x^{(2+\epsilon)/5}$$

where $\epsilon > 12\eta + 7\delta$. However this is not quite good enough, since we need to be able to take N as small as $x^{1/3}$.

We can try a modification of this proof, the most successful being where, before we Cauchy equation (10.1) we also fix the ℓ_1 variable. This variant allows us to extend our range to all

$$N > x^{\frac{1}{3}+\epsilon}$$

where $\epsilon > \frac{14}{3}\eta + \frac{7}{2}\delta$. We are very close to the exponent $\frac{1}{3}$, but it seems that we are destined to just fail.

11. Complete exponential sums: Combining information the Graham-Ringrose way

The "square-root cancellation" for incomplete exponential sums of the form $|\sum_n e_q(f(n))|$ for various moduli q, with the sum over n in an interval of length $N < q$ is not quite good enough to obtain our results.

Graham and Ringrose [**17**] proved that we can improve the (analogous) incomplete character sum bounds when q is smooth. Here we follow Polymath8 [**34**], who showed how to modify the Graham-Ringrose argument to incomplete exponential sums. This will allow us to reduce the size of N in the above argument and prove our result.

11.1. **Formulating the improved incomplete exponential sum result.** For convenience we will write the entry of the exponential sum as $f(n)$, which should be thought of as taking the form $a/n + b/(n + \Delta) + cn$, though the argument is rather more general. We assume that $N < q$, so that the Weil bound gives

$$\left|\sum_n e_q(f(n))\right| \ll \tau(q)^A q^{1/2}. \tag{11.1}$$

for some constant A which depends only on the degree of f.

In what follows we will assume that q is factored as $q = q_1 q_2$, and we will deduce that

$$\left|\sum_n e_q(f(n))\right| \ll \left(q_1^{1/2} + q_2^{1/4}\right) \tau(q)^A (\log q) N^{1/2}. \tag{11.2}$$

If q is y-smooth then we let q_1 be the largest divisor of q that is $\leqslant (qy)^{1/3}$ so that it must be $> (q/y^2)^{1/3}$, and so $q_2 \leqslant (qy)^{2/3}$. Hence the last bound implies

$$\left|\sum_n e_q(f(n))\right| \ll \tau(q)^A (qy)^{1/6} (\log q) N^{1/2}.$$

It is this bound that we insert into the machinery of the previous section, and it allows use to extend our range to all

$$N > x^{\frac{3}{10}+\epsilon}$$

where ϵ is bounded below by a (positive) linear combination of η and δ. In order that we can stretch the range down to *all* $N > x^{\frac{1}{3}}$, this method requires that

$$162\eta + 90\delta < 1.$$

11.2. **Proof of (11.2).** We may assume

$$q_1 \leqslant N \leqslant q_2$$

else if $N < q_1$ we have the trivial bound $\leqslant N < (q_1 N)^{1/2}$, and if $N > q_2$ then (11.1) implies the result since $q^{1/2} = (q_1 q_2)^{1/2} < (q_1 N)^{1/2}$.

The main idea will be to reduce our incomplete exponential sum mod q, to a sum of incomplete exponential sums mod q_2. Now

$$e_q(f(n + kq_1)) = e_{q_1}(f(n)/q_2)\ e_{q_2}(f(n + kq_1)/q_1)$$

so that, by a simple change of variable, we have

$$\sum_n e_q(f(n)) = \sum_n e_q(f(n + kq_1))) = \sum_n\ e_{q_1}(f(n)/q_2)\ e_{q_2}(f(n + kq_1)/q_1).$$

Now, if we sum this over all $k, 1 \leqslant k \leqslant K := \lfloor N/q_1 \rfloor$, then we have

$$K \sum_n e_q(f(n)) = \sum_n e_{q_1}(f(n)/q_2)\ \sum_{k=1}^{K} e_{q_2}(f(n + kq_1)/q_1),$$

and so

$$\begin{aligned}\left| K \sum_n e_q(f(n)) \right|^2 &\leqslant \left(\sum_n \left| \sum_{k=1}^{K} e_{q_2}(f(n + kq_1)/q_1) \right| \right)^2 \\ &\ll N \sum_n \left| \sum_{k=1}^{K} e_{q_2}(f(n + kq_1)/q_1) \right|^2 \\ &= N \sum_{1 \leqslant k,k' \leqslant K}\ \sum_n e_{q_2}(g_{k,k'}(n)),\end{aligned}$$

where $g_{k,k'}(n) := (f(n+kq_1) - f(n+k'q_1))/q_1 \pmod{q_2}$ if $n+kq_1,\ n+k'q_1 \in I$, and $g_{k,k'}(n) := 0$ otherwise. If $k = k'$ then $g_{k,k}(n) = 0$, and so these terms contribute $\leqslant KN^2$.

We now apply the bound of (11.1) taking $f = g_{k,k}$ for $k \neq k'$. Calculating the sum yields (11.2).

11.3. **Better results.** In **[34]** the authors obtain better results using somewhat deeper techniques.

By replacing the set of y-smooth integers by the much larger class of integers with divisors in a pre-specified interval (and such that those divisors have divisors in a different pre-specified interval, etc., since one can iterate the proof in the previous section) they improve the restriction to

$$84\eta + 48\delta < 1.$$

Following Zhang they also gained bounds on certain higher order convolutions (of the shape $\alpha * 1 * 1 * 1$), though here needing deeper exponential sum estimates, and were then able to improve the restriction to (slightly better than)

$$43\eta + 27\delta < 1.$$

11.4. **Final remark.** It is worth noting that one can obtain the same quality of results only assuming a bound $\ll p^{2/3-\epsilon}$ for the relevant exponential sums in finite fields.

References

[1] E. Bombieri, *On the large sieve*, Mathematika **12** (1965), 201–225.

[2] E. Bombieri and H. Davenport *Small difference between prime numbers*, Proc. Roy. Soc. Ser. A **293** (1966), 1-18.

[3] E. Bombieri, J. Friedlander and H. Iwaniec*Primes in arithmetic progressions to large moduli*, Acta Math. **156** (1986), no. 3-4, 203–251.

[4] E. Bombieri, J. Friedlander and H. Iwaniec, *Primes in arithmetic progressions to large moduli. II*, Math. Ann. **277** (1987), no. 3, 361–393.

[5] E. Bombieri, J. Friedlander and H. Iwaniec, *Primes in arithmetic progressions to large moduli. III*, J. Amer. Math. Soc. **2** (1989), no. 2, 215–224.

[6] P. Deligne, *La conjecture de Weil. II*, Publications Mathématiques de l'IHÉS **52** (1980), 137–252.

[7] J.-M. Deshouillers and H. Iwaniec, *Kloosterman Sums and Fourier Coefficients of Cusp Forms*, Inventiones mathematicae **70** (1982/83), 219-219.

[8] P. D. T. A. Elliott and H. Halberstam, *A conjecture in prime number theory*, Symp. Math. **4** (1968), 59–72.

[9] E. Fouvry, *A new form of the error term in the linear sieve*, Acta Arith., **37** (1980), 307–320.

[10] E. Fouvry, *Autour du théorème de Bombieri-Vinogradov*, Acta Math. **152** (1984), no. 3-4, 219–244.

[11] E. Fouvry and H. Iwaniec, *On a theorem of Bombieri-Vinogradov type*, Mathematika **27** (1980), no. 2, 135–152 (1981).

[12] E. Fouvry and H. Iwaniec, *Primes in arithmetic progressions*, Acta Arith. **42** (1983), no. 2, 197–218.

[13] J. Friedlander and A. Granville, *Limitations to the equi-distribution of primes. I*, Ann. of Math. **129** (1989), 363-382.

[14] J. Friedlander and H. Iwaniec, *Incomplete Kloosterman sums and a divisor problem*, With an appendix by Bryan J. Birch and Enrico Bombieri. Ann. of Math. (2) **121** (1985), no. 2, 319–350.

[15] D. Goldston, J. Pintz and C. Yıldırım, *Primes in tuples. I*, Ann. of Math. **170** (2009), no. 2, 819–862.

[16] D. Goldston, S. Graham, J. Pintz and C. Yıldırım, *Small gaps between primes or almost primes*, Trans. Amer. Math. Soc. **361** (2009), no. 10, 5285–5330.

[17] S. W. Graham and C. J. Ringrose, *Lower bounds for least quadratic nonresidues*, Analytic number theory (Allerton Park, IL, 1989), 269–309, Progr. Math., 85, Birkhäuser Boston, Boston, MA, 1990.

[18] A. Granville and K. Soundararajan, *Multiplicative number theory; the pretentious approach*, to appear.

[19] B.J. Green and T. Tao, *The primes contain arbitrarily long arithmetic progressions*, Annals of Mathematics **167** (2008), 481-547.

[20] B.J. Green, T. Tao and T. Ziegler*An inverse theorem for the Gowers $U^{s+1}[N]$-norm*, Annals of Mathematics **176** (2012), 1231-1372.

[21] G. H. Hardy and J. E. Littlewood, *Some problems of "Partitio Numerorum", III: On the expression of a number as a sum of primes*, Acta Math. **44** (1923), 1–70.

[22] D. R. Heath-Brown, *Prime numbers in short intervals and a generalized Vaughan identity*, Canad. J. Math. 34 (1982), no. 6, 1365–1377.

[23] H. A. Helfgott, *Major arcs for Goldbach's theorem*, to appear.

[24] D. Hensley and I. Richards, *On the incompatibility of two conjectures concerning primes*, Analytic number theory (Proc. Sympos. Pure Math., Vol. XXIV, St. Louis Univ., St. Louis, Mo., 1972), pp. 123–127. Amer. Math. Soc., Providence, R.I., 1973.

[25] D. Hensley and I. Richards, *Primes in intervals*, Acta Arith. **25** (1973/74), 375–391.

[26] H. Iwaniec and E. Kowalski, *Analytic number theory*, AMS Colloquium Publications, **53** (2004).

[27] H. D. Kloosterman, *On the representation of numbers in the form* $ax^2 + by^2 + cz^2 + dt^2$, Acta Mathematica **49** (1926), pp. 407–464.

[28] H. Maier, *Small differences between prime numbers*, Michigan Math. J. **35** (1988), 323–344.

[29] J. Maynard, *Small gaps between primes*, preprint.

[30] L. J. Mordell, *On a sum analogous to a Gauss's sum*, Quart. J. Math. Oxford Ser. **3** (1932), 161–167.

[31] Y. Motohashi, *An induction principle for the generalization of Bombieri's Prime Number Theorem*, Proc. Japan.. Acad. **52** (1976) 273–275.

[32] Y. Motohashi and J. Pintz, *A smoothed GPY sieve*, Bull. Lond. Math. Soc. **40** (2008), no. 2, 298–310.

[33] J. Pintz, *Polignac Numbers, Conjectures of Erdős on Gaps between Primes, Arithmetic Progressions in Primes, and the Bounded Gap Conjecture*, preprint.

[34] D.H.J. Polymath, *A new bound for small gaps between primes*, preprint.

[35] A. Schinzel, *Remarks on the paper "Sur certaines hypothéses concernant les nombres premiers"*, Acta Arith. **7** (1961/1962) 1–8.

[36] A. Selberg, *On elementary methods in prime number-theory and their limitations*, in Proc. 11th Scand. Math. Cong. Trondheim (1949), Collected Works, Vol. I, 388–397, Springer-Verlag, Berlin-Göttingen-Heidelberg, 1989.

[37] P. Shiu, *A Brun-Titchmarsh theorem for multiplicative functions*, J. Reine Angew. Math. 313(1980), 161–170.

[38] K. Soundararajan, *Small gaps between prime numbers: the work of Goldston-Pintz-Yildirim*, Bull. Amer. Math. Soc. (N.S.) **44** (2007), no. 1, 1–18.

[39] T. Tao, *private communication.*

[40] R. C. Vaughan, *Sommes trigonométriques sur les nombres premiers*, C. R. Acad. Sci. Paris Sér. A **285** (1977), 981–983.

[41] A. I. Vinogradov, *The density hypothesis for Dirichlet L-series*, Izv. Akad. Nauk SSSR Ser. Mat. **29** (1965), 903–934.

[42] A. Weil, *Numbers of solutions of equations in finite fields*, Bulletin of the American Mathematical Society **55** (1949), 497–508.

[43] Y. Zhang, *Bounded gaps between primes*, to appear, Annals of Mathematics.

Département de mathématiques et de statistiques, Université de Montréal, Montréal QC H3C 3J7, Canada.

E-mail address: andrew@dms.umontreal.ca

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 11, 2013 (San Diego, CA)

Wei Ho, Columbia University
How many rational points does a random curve have?

Sam Payne, Yale University
Topology of nonarchimedean analytic spaces

Mladen Bestvina, University of Utah
Geometric group theory and 3-manifolds hand in hand: the fulfillment of Thurston's vision for three-manifolds

Lauren Williams, University of California, Berkeley
Cluster algebras

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

2014 CURRENT EVENTS BULLETIN

Committee

Hélène Barcelo, *Mathematical Sciences Research Institute*

Mladen Bestvina, *University of Utah*

Robert Bryant, *Duke University*

Laura DeMarco, *University of Illinois, Chicago*

David Eisenbud, *University of California, Berkeley, Chair*

Susan Friedlander, *University of Southern California*

Ben Green, *University of Oxford*

Christopher Hacon, *University of Utah*

Wei Ho, *Columbia University*

Richard Karp, *University of California, Berkeley*

Izabella Laba, *University of British Columbia*

John Morgan, *Stony Brook University*

David Nadler, *University of California, Berkeley*

Asaf Naor, *Courant Institute of Mathematical Sciences*

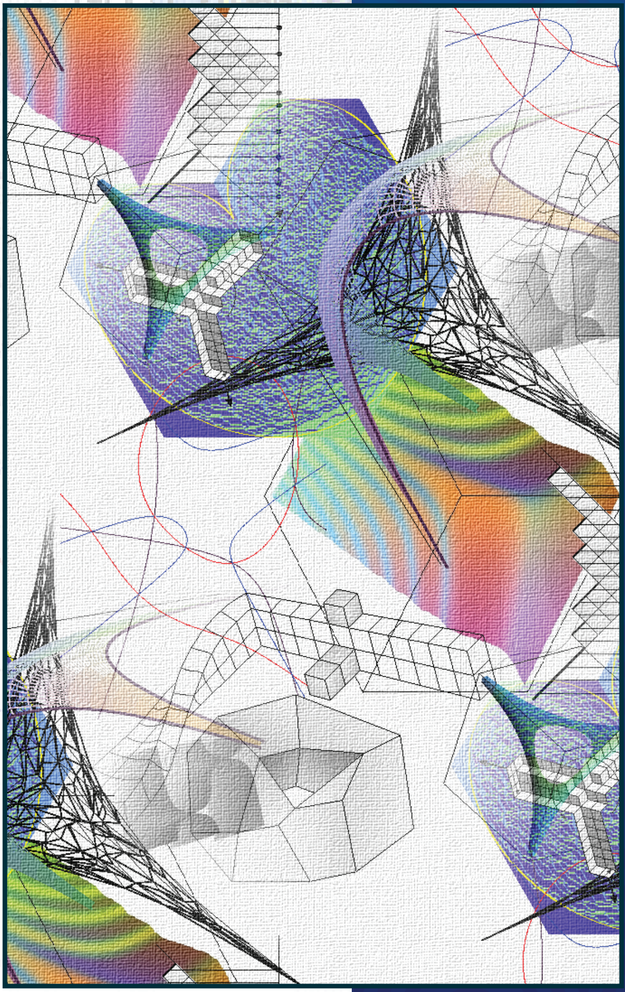
Sam Payne, *Yale University*

Richard Taylor, *Institute for Advanced Study*

Ulrike Tillmann, *Oxford University*

Akshay Venkatesh, *Stanford University*

Lauren Williams, *University of California, Berkeley*



The back cover graphic is reprinted courtesy of Andrei Okounkov.

Cover graphic associated with Karen Vogtmann's talk courtesy of Jos Leys.