ARTIGOS
Matemática Universitária Nº 19, Dezembro de 1995, 16-43

Uma Demonstração Maravilhosa

Fernando Q. Gouvêa[1]

Ninguém sabe ao certo quando foi que Pierre de Fermat fez aquela famosa inscrição na margem da sua cópia da *Aritmética* de Diofanto. Foi provavelmente por volta de 1635, mas foi só após a sua morte que a anotação foi tornada pública. A data do anúncio da demonstração, por outro lado, é conhecida. Foi em 23 de junho de 1993, na Universidade de Cambridge, e o anunciante foi Andrew Wiles, professor na Universidade de Princeton.

O anúncio inicial foi seguido de um período de intensa atividade. O manuscrito foi enviado a vários "referees" para ser avaliado, e o mundo matemático ficou em suspense até dezembro de 1993, quando Wiles anunciou que vários erros haviam sido encontrados. Na sua maioria, os problemas foram resolvidos facilmente, mas um deles havia provado ser realmente difícil, e Wiles concluiu que naquela altura a demonstração permanecia incompleta. Quase um ano mais tarde, em outubro de 1994, pesquisadores em teoria de números receberam uma agradável surpresa: um manuscrito, grande, por Wiles, reduzindo o problema à demonstração de um resultado sobre o anel de operadores de Hecke. Junto desse, um outro manuscrito, por Wiles e Taylor, continha a demonstração do resultado que estava faltando: o teorema estava demonstrado.

A demonstração encontrada por Wiles é uma mistura engenhosa de vários tópicos que tem sido, já há muitos anos, foco de intensa pesquisa em teoria de números: curvas elípticas, formas modulares e representações galoisianas. O objetivo deste artigo é dar acesso às idéias básicas da demonstração aos matemáticos que não são especialistas no assunto. Para isso, é preciso começar por um breve esboço destes tópicos, e em seguida tentar descrever como eles são usados para obter a demonstração. O leitor que estiver interessado principalmente na estrutura da demonstração e que não quer (ou não precisa) rever os conceitos

1 Com o apoio parcial da National Science Foundation, grant DMS-9401313.

básicos poderá apenas passar os olhos na segunda parte deste artigo e depois concentrar sua atenção na terceira, que contém a descrição da demonstração propriamente dita. Nossa discussão inclui algumas poucas observações sobre a história do problema, mas traçar essa história não é nossa intenção principal, e portanto nós apenas mencionados os pontos que parecem ser relevantes ao nosso objetivo de descrever as idéias da demonstração.

Meus sinceros agradecimentos a várias pessoas que leram e comentaram versões preliminares deste artigo, inclusive Barry Mazur, Kenneth Ribet, Serge Lang, Noriko Yui, George Elliot, Keith Devlin e Lynette Millett. Durante a preparação deste artigo o autor teve o apoio financeiro da National Science Foundation, através do "grant" DMS-9401313.

1 - Preliminares

A afirmação que Fermat escreveu na margem é bem conhecida. Ele diz que para qualquer expoente $n \geq 3$ não existem soluções não-triviais em inteiros da equação $x^n + y^n = z^n$. (Aqui, "não-trivial" significa apenas que nenhum dos três inteiros x, y e z deve ser igual a zero.) Fermat afirma, na sua nota, que havia encontrado uma demonstração maravilhosa deste fato, mas que infelizmente a demonstração não caberia na margem.

Este enunciado ficou conhecido como o "último teorema de Fermat," não porque se achasse que este "teorema" tivesse sido o último descoberto por Fermat, mas porque no começo do século XIX esta era a última asserção de Fermat que permanecia em aberto, sem prova ou refutação. No que segue, nós chamaremos a equação $x^n + y^n = z^n$ de "equação de Fermat."

Os primeiros resultados importantes em relação ao "teorema" de Fermat foram teoremas mostrando que a afirmação de Fermat era verdadeira para valores específicos do expoente n. O primeiro destes se deve ao próprio Fermat: poucas de suas demonstrações foram tornadas públicas, mas em uma de suas cartas ele explica como provar que a equação

$$x^4 + y^4 = z^2$$

não tem soluções não triviais em inteiros. Como qualquer solução da equação de Fermat com expoente 4 daria uma solução desta equação também, pode-se concluir que a afirmação de Fermat é verdadeira para o caso $n = 4$.

Uma vez estabelecido este fato, é fácil ver que podemos agora restringir nossa atenção ao caso em que n é um número primo ímpar. Para ver por quê, basta notar, primeiro, que qualquer inteiro maior que 2 é divisível ou por um primo ímpar ou por 4, e depois, que podemos escrever a equação

$$x^{m\,k} + y^{m\,k} = z^{m\,k}$$

como

$$(x^m)^k + (y^m)^k = (z^m)^k.$$

Isto mostra que uma solução para o caso $n = mk$ permite obter de imediato uma solução para o caso $n = k$. Se n não for primo, nós sempre podemos escolher a fatorização de modo que k seja ou 4 ou um primo ímpar. Se já sabemos que não há soluções nesses casos, não pode haver uma solução para n. Assim, o problema fica reduzido aos casos em que o expoente é primo ou é igual a quatro.

No meio do século XVIII, Euler ficou interessado no que Fermat havia descoberto em teoria de números, e começou uma investigação sistemática do assunto. Pouco a pouco, ele foi examinando as muitas asserções de Fermat, encontrando demonstrações para a maioria. Em particular, ele considerou a equação de Fermat para $n = 3$ e $n = 4$ provando outra vez que não existem soluções nesses dois casos. (A demonstração de Euler para $n = 3$ depende do estudo dos "números" que resultam quando se "acrescenta" $\sqrt{-3}$ aos números racionais, uma das primeiras instâncias em que os "número algébricos" aparecem na matemática.) Um bom relato histórico do trabalho de Euler pode ser encontrado em [Wei83].

Nos anos seguintes, vários outros matemáticos generalizaram os resultados de Euler, passo a passo, cobrindo os casos $n = 5,7$ .... Uma descrição das fortunas do "teorema" durante este período pode ser encontrada em [Rib79].

De lá até hoje, métodos mais e mais sofisticados foram sendo descobertos que permitem testar a asserção de Fermat para valores específicos de n. Como resultado, a lista de expoentes para os quais se sabia que o teorema de Fermat era verdadeiro foi ficando mais e mais longa. Em 1992, se sabia que o "teorema" era verdadeiro para expoentes primos até 4 000 000 (cf. [BCS92]).

É claro, entretanto, que para obter resultados gerais é preciso um método geral, que se aplique a todos os valores de n. Uma maneira de se obter isso seria descobrir uma maneira de ligar a equação de Fermat (para um expoente n arbitrário) a alguma outra área da matemática. Isto permitiria usar teoremas sobre esta outra área para estudar a equação de Fermat, e, quem sabe, provar que não há soluções. Nos últimos séculos houve várias tentativas de se estabelecer uma tal ligação; nós mencionamos aqui apenas duas tentativas que foram particularmente bem sucedidas (e omitimos uma grande quantidade de bons resultados, para os quais o leitor pode consultar, por exemplo, [Rib79]).

O primeiro destes sucessos ocorreu no trabalho de E. Kummer, que, no meio do século XIX, estabeleceu uma ligação entre o teorema de Fermat e a teoria de corpos ciclotômicos. Esta ligação permitiu que Kummer provasse que o teorema é verdadeiro no caso em que o expoente é um número primo com uma propriedade particularmente favorável (Kummer chamou tais primos de "regulares"). A demonstração de Kummer é um trabalho impressionante, e foi o primeiro resultado

sobre a equação de Fermat a ter real alcance. Isto fica particularmente claro à luz do fato que, experimentalmente, a maior parte dos primos parece ser regular. Infelizmente, ninguém até hoje conseguiu nem mesmo provar que o conjunto dos primos regulares é infinito. (E, ironicamente, já foi provado que existem infinitos primos que *não* são ragulares.) Uma discussão das idéias de Kummer pode ser achada em [Rib79]; para mais informações sobre a teoria dos corpos ciclotômicos pode-se começar com [Was82].

O segundo sucesso que devemos mencionar é o trabalho de G. Faltings, que, em 1983, provou uma conjetura (devida a Mordell) sobre as soluções racionais de certos tipos de equações polinomiais. Aplicando este resultado às equações de Fermat, se obtém que para cada $n \geq 4$ existe apenas um número *finito* de soluções não-triviais. É um resultado poderoso, mas outra vez o impacto sobre o problema de Fermat propriamente dito acaba sendo menor, porque ainda não se encontrou um método de determinar o número exato de soluções que se espera existir. Para uma introdução ao trabalho de Faltings, o leitor poderá consultar [CS86], que contém também uma tradução (para o inglês) do artigo original (em alemão).

A demonstração de Wiles depende uma outra ligação deste tipo, desenvolvida nos anos oitenta por G. Frey, J.-P.Serre e K. A. Ribet. A idéia é relacionar o teorema de Fermat com a teoria das curvas elípticas, que tem sido objeto de intenso estudo durante todo este século. Através das curvas elípticas, o problema fica também relacionado a toda a maquinaria das formas modulares e das representações galoisianas que, no final, permite obter a demonstração.

Notação: Nós usaremos os símbolos usuais: Q representa o conjunto dos números racionais, e $\mathbf{Z}$ o dos números inteiros. Os inteiros módulo m serão denotados por[2] $\mathbf{Z}/m\mathbf{Z}$; na maior parte das vezes o inteiro m será uma potência de um número primo p. Se p é primo, o anel $\mathbf{Z}/p\mathbf{Z}$ é um corpo, e nós enfatizaremos este fato usando uma notação diferente:

$$F_p = \mathbf{Z}/p\mathbf{Z} .$$

À medida que formos encontrando os personagens principais da demonstração, nós iremos introduzir outras convenções quanto a notação.

2 Muitos textos elementares usam $\mathbf{Z}_m$ como a notação para os inteiros módulo m; para nós (e para teoria de números em geral) esta notação é inconveniente porque colide com a notação para os inteiros p-ádicos descritos abaixo

2 - Os Personagens

Começamos introduzindo os personagens principais. Primeiro, fazemos uma breve (e informal) introdução aos números p-ádicos. Estes não são tanto atores do drama quanto parte do cenário: instrumentos que permitem que os atores façam o seu trabalho. Em seguida, damos um esboço breve e impressionista das teorias de curvas elípticas, formas modulares e representações galoisianas.

2.1- Números p-ádicos

Os números p-ádicos são uma extensão do corpo dos números racionais que é, de certo modo, análoga aos números reais. Como os números reais, os p-ádicos podem ser obtido a partir de uma noção de distância entre dois números racionais, tornando Q um espaço métrico. Passando ao completamento desse espaço métrico, obtemos um corpo Q_p. Para entender a idéia básica da demonstração nós não precisamos saber muito a respeito deles. Os fatos cruciais são:

1. Para cada número primo p existe um corpo Q_p que é completo com respeito a certa noção de distância e contém os números racionais como um sub-corpo denso.

2. Proximidade na métrica p-ádica se traduz em congruência módulo potências de p. Por exemplo, dois inteiros cuja diferença é divisível por p^n são "próximos" um do outro no mundo p-ádico (quanto maior for n, mais próximos eles são).

3. Como conseqüência, pode-se pensar nos p-ádicos como uma maneira de entender informação sobre congruência: sempre que soubermos algo módulo p^n para todo n, podemos traduzir esta informação numa propriedade p-ádica, e vice-versa.

4. O corpo Q_p contém um subanel Z_p, que se chama o *anel de inteiros p-ádicos*. Z_p é o fecho do conjunto Z dos números inteiros em Q_p.

É claro que há muito mais a dizer, e o leitor encontrará as idéias básicas em várias referência (por exemplo, [Kob84], [Cas86], [Ami75], e mesmo [Gou93]). Os números p-ádicos foram inventados por K. Hensel (um discípulo de Kummer). De fato, muitas das idéias parecem estar implícitas no trabalho de Kummer. Desde então os p-ádicos se tornaram um instrumento fundamental da teoria de números.

2.2. Curvas Elípticas

As curvas elípticas são um tipo especial[3] de curva algébrica. Estas curvas são especialmente interessantes porque têm uma estrutura aritmética muito rica, que vai ser o foco da nossa discussão.

Há várias maneiras sofisticadas de fazer a definição básica, mas para nós é mais razoável simplesmente definir curvas elípticas como o conjunto de pontos que satisfazem um certo tipo de equação polinomial. Especificamente, considere uma equação do tipo

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6 ,$$

onde os a_i são números inteiros (há uma boa razão para a escolha estranha dos índices, mas não vale a pena entrar em detalhes aqui). Nós queremos considerar o conjunto dos pontos (x,y) que satisfazem esta equação. Como estamos fazendo teoria de números, é importante deixar em aberto (por hora, pelo menos) que tipo de números estamos pensando serem as coordenadas x e y: faz sentido imaginá-las como números reais, como números complexos, como números racionais, e mesmo, para qualquer número primo p, como inteiros módulo p (o que equivale a pensar na equação como uma congruência módulo p). Uma maneira de conceber a situação é imaginar que existe um objeto subjacente que nós chamamos de *curva* E e, para cada um dos possíveis corpos nos quais as coordenadas (x,y) podem viver, nós chamamos o conjunto de soluções de conjunto de "pontos de E definidos sobre este corpo." Assim, se considerarmos todas as soluções em números complexos, obtemos o conjunto $E(\boldsymbol{C})$ dos pontos complexos de E. Da mesma forma, podemos investigar os pontos reais $E(\boldsymbol{R})$ os pontos racionais $E(\boldsymbol{Q})$ e o conjunto $E(\boldsymbol{F}_p)$ dos pontos modulo p.

Nós ainda não dissemos quando é que uma equação deste tipo define uma curva elíptica. A condição é simplesmente que se trata de uma *curva lisa.* Se estivermos considerando pontos reais ou complexos, isto significa exatamente o que se espera: a curva não contém pontos "singulares", isto é, existe uma reta tangente bem definida em cada um dos pontos da curva. (No caso dos pontos complexos, a "reta" tangente é um *plano* tangente, i.é, é uma "reta complexa.") Como aprendemos no curso de cálculo, uma equação $f(x,y) = 0$ define uma curva lisa exatamente quando não há pontos na curva em que ambas as derivadas parciais de f se anulem simultaneamente. Em outras palavras, a curva será lisa quando não houver soluções comuns das equações

3 Talvez seja melhor remover a confusão natural logo no princípio: elipses não são curvas elípticas. Na realidade, a ligação entre elipses e curvas elípticas é sutil. A idéia é a seguinte: as curvas elípticas (pensadas em termos dos números complexos) são o "habitat natural" das integrais elípticas que aparecem, por exemplo, quando se tenta calcular o comprimento de arco de uma elipse. Para nós, esta ligação será de muito pouca importância.

$$f(x,y) = 0\ ,\quad \frac{\partial f}{\partial x}(x,y) = 0\ ,\quad \frac{\partial f}{\partial y}(x,y) = 0\ .$$

Notamos, entretanto, que esta condição é realmente uma condição algébrica, porque as derivadas que aparecem são derivadas de polinômios, e podem ser calculadas formalmente. Neste caso, é ainda mais fácil, porque é possível reduzir a condição toda a um polinômio (complicado) nos a_i. Precisamente, existe um polinômio $\Delta = \Delta\,(a_1,a_2,a_3,a_4,a_6)$ nos coeficientes a_i tal que E é uma curva lisa se e somente se $\Delta(E) \neq 0$. Isto permite dar uma definição completamente formal, que faz sentido mesmo sobre $\boldsymbol{F}_p$. O número Δ é chamado o *discriminante* da curva E.

Definição 1. *Seja K um corpo. Uma curva elíptica sobre K é uma curva algébrica definida por uma equação do tipo*

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6\ ,$$

onde os a_i pertencem a K e satisfazem a condição $\Delta\,(a_1,a_2,a_3,a_4,a_6) \neq 0$.

Os especialistas provavelmente gostariam de re-escrever a definição de modo a permitir outros tipos de equação, desde que se possa transformá-las em equações do nosso tipo através de uma mudança de variável.

Já é tempo de dar alguns exemplos. Para tornar as coisas mais fáceis, vamos focalizar a nossa atenção no caso especial em que a equação é da forma $y^2 = g(x)$, com $g(x)$ um polinômio de grau 3 (na notação acima, estamos assumindo que $a_1 = a_3 = 0$) . Neste caso, é fácil decidir quando é que podem haver pontos singulares, e mesmo que tipo de pontos singulares eles serão. Se pusermos $f(x,y) = y^2 - g(x)$, teremos

$$\frac{\partial f}{\partial x}(x,y) = -g'(x) \quad \text{e} \quad \frac{\partial f}{\partial y}(x,y) = 2y,$$

e a condição para um ponto ser "mal comportado" fica sendo

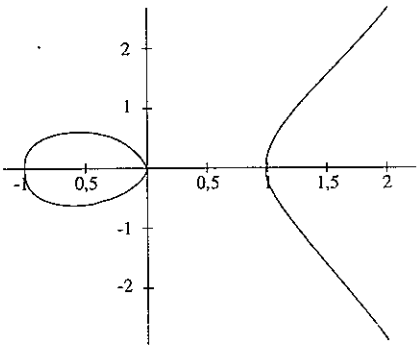
$$y^2 = g(x) \qquad g'(x) = 0 \qquad 2y = 0,$$

que se reduz a $y = g(x) = g'(x) = 0$. Em outras palavras, um ponto será ruim exatamente quando sua ordenada y é zero e sua abcissa x é uma *raiz dupla* do polinômio $g(x)$. Como $g(x)$ é de grau 3, isto dá apenas três possibilidades.

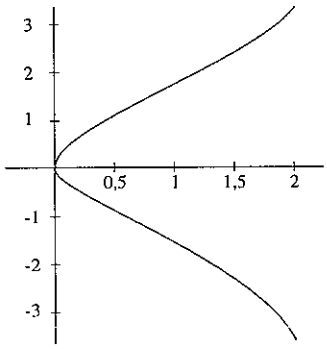
- $g(x)$ não tem raízes múltiplas, e a equação define uma curva elíptica;
- $g(x)$ tem uma raiz dupla;
- $g(x)$ tem uma raiz tripla.

Vamos considerar um exemplo de cada caso, e fazer o gráfico dos pontos reais da curva correspondente.

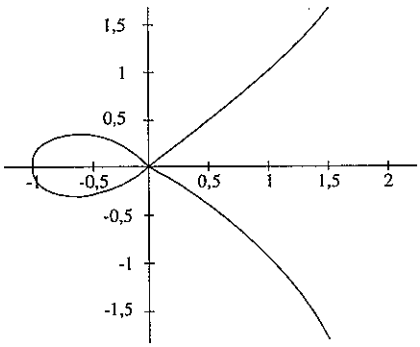
Para o primeiro caso, considere a curva definida por $y^2 = x^3 - x$. Seu gráfico é a figura 1(a) (mais precisamente, este é o gráfico dos pontos reais da curva). Um outro exemplo do mesmo caso é a curva definida pela equação $y^2 = x^3 + x$; veja a figura 1(b). (A razão da diferença entre estes gráficos é que nós estamos olhando apenas os pontos reais da curva; um destes polinômios tem três raízes reais, e o outro tem uma só. Sobre os números complexos, estas duas curvas são isomorfas.)



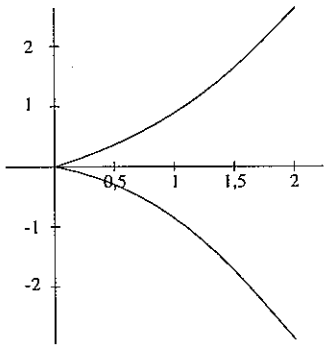
(a) $y^2 = x^3 - x$



(b) $y^2 = x^3 + x$



(c) $y^2 = x^3 + x^2$: um nó



(d) $y^2 = x^3$: uma cúspide

Figura 1 - Gráficos de várias equações cúbicas

Quando há pontos "ruins", o que acontece é que ou duas das raízes de $g(x)$ coincidem, ou as três coincidem. No primeiro caso, temos um laço; no ponto de cruzamento, que se costuma chamar um "nó", a curva tem duas retas tangentes diferentes. Veja a figura 1(c), que é o gráfico da equação $y^2 = x^3 + x^2$ (raiz dupla em $x = 0$).

No terceiro caso, não são apenas as três raízes que coincidem, mas também as duas retas tangentes do nó coincidem para formar uma espécie de "tangente dupla" (isto pode ser tornado preciso com um pouco de álgebra de polinômios, mas é mais divertido pensar a respeito geometricamente, imaginando as raízes chegando cada vez mais perto umas das outras). O gráfico fica parecido com a figura 1(d), e dizemos que a curva tem uma singularidade cuspidal.

Como é que tudo isso se liga ao discriminante Δ mencionado acima? Bom, se r_1, r_2 e r_3 são as raízes do polinômio $g(x)$, o discriminante da equação $y^2 = g(x)$ é igual a

$$\Delta = \alpha(r_1 - r_2)^2 (r_1 - r_3)^2 (r_2 - r_3)^2,$$

onde α é uma constante. Isto faz exatamente o que nós queremos: se duas das raízes são iguais, $\Delta = 0$; se não, não. Além disso, não é difícil ver que Δ é um polinômio nos coeficientes de $g(x)$, como nós afirmamos acima. Em outras palavras, o discriminante nos dá um procedimento algébrico direto para determinar se há pontos singulares.

Esta análise foi feita especificamente para curvas de forma $y^2 = g(x)$, mas na realidade os resultados são igualmente válidos para todas as equações do tipo que estamos estudando: existe no máximo um ponto singular, e a singularidade ou é um nó ou é cuspidal.

Uma última idéia geométrica: como se vê nos gráficos, estas curvas não são fechadas. É conveniente torná-las fechadas. Fazemos isto acrescentando um outro ponto à curva. Normalmente, falamos de acrescentar "um ponto no infinito". Isto pode ser feito de modo preciso usando a idéia do plano projetivo, e simplesmente tomando o fecho da curva quando pensada como um subconjunto do plano projetivo. Para nós, entretanto, a única coisa importante é lembrar que nós temos mais um ponto em nossas curvas. (Devemos imaginá-lo "infinitamente longe no eixo y", mas lembrando que há um ponto só, de modo que o *mesmo* ponto é "infinitamente longe para cima" e "infinitamente longe para baixo.")

Com estes exemplos em mãos, podemos entrar em águas mais profundas. Para entender a ligação que vamos estabelecer entre curvas elípticas e o teorema de Fermat, precisamos rever uma boa parte do que se sabe sobre a rica estrutura aritmética destas curvas.

A primeira coisa a notar é que se pode definir uma operação no conjunto dos pontos de uma curva elíptica, isto é, uma maneira de "somar" dois pontos. Esta

operação, que é bastante natural, torna a curva elíptica um grupo abeliano. (Em outras palavras, "somar" pontos é uma operação comutativa, associativa, com elemento neutro e com inversos.) O elemento neutro deste grupo é o ponto no infinito (seria mais honesto, talvez, dizer que nós *escolhemos* o ponto do infinito para cumprir esse papel).

Não é preciso, neste artigo, entrar em detalhes sobre como se somam pontos numa curva elíptica. Há várias definições equivalentes, e cada uma tem suas vantagens. Nas referências, o leitor poderá encontrar detalhes sobre a definição, e a demonstração de que a operação de fato tem as propriedades que definem um gurpo abeliano. Por agora, a coisa que nós precisamos saber sobre a definição é que ela é feita de modo a preservar o corpo de definição dos pontos: a soma de dois pontos racionais é um ponto racional, e assim por diante.

O resultado é que para cada escolha de corpo de base, nós acabamos conseguindo um grupo, o grupo dos pontos da curva com coordenadas naquele corpo. Dessa forma, a curva elíptica nos dá um monte de grupos diferentes, todos, é claro, parentes uns dos outros. (O parentesco vem do fato de que todos provêm da mesma curva; a relação exata que isto estabelece entre os grupos é às vezes misteriosa.) Assim, dada uma curva E, podemos considerar os seus pontos complexos $E(\boldsymbol{C})$ que formam um grupo de Lie complexo que topologicamente é um toro, ou podemos olhar para o gurpo de Lie real $E(\boldsymbol{R})$ que acaba sendo isomorfo ou ao círculo S^1 ou ao produto direto $\mathbf{Z}/2\mathbf{Z} \times S^1$ (Veja os exemplos de curvas elípticas acima; você consegue decidir quais curvas correspondem a quais grupos de pontos reais?)

Do ponto de vista da aritmética, o mais interessante desses grupos é o grupo dos pontos racionais, $E(\boldsymbol{Q})$. Um ponto $P \in E(\boldsymbol{Q})$ corresponde a uma solução da nossa equação cúbica em números racionais. Achar tais soluções é resolver uma *equação diofantina*, um tipo de problema que tem uma longa história. No caso do grupo $E(\boldsymbol{Q})$ um fato adicional torna a situação ainda mais interessante: na década de vinte, L. Mordell e A. Weil provaram que este é um grupo abeliano *finitamente gerado*. O que isto significa, na prática, é o seguinte: existe uma lista finita de pontos racionais da curva (ou, equivalemente, de soluções da equação em números racionais) tais que todos os outros pontos racionais (ou outras soluções) são obtidos a partir desses usando a operação de somar pontos. Como a operação é relativamente simples, uma lista deste tipo seria uma solução completa do problema de achar pontos racionais. Os pontos nessa lista são chamados os *geradores* do grupo $E(\boldsymbol{Q})$ que costuma ser chamado o *grupo de Mordell-Weil* de E.

As curvas elípticas que nós consideramos acima têm grupos de Mordell-Weil bem simples. No caso da curva $y^2 = x^3 - x$ (figura 1a), o grupo tem quatro elementos, isto é, há quatro pontos racionais: (0,0), (±1,0) e o ponto no infinito.

No caso de $y^2 = x^3 + x$ (figura1b), há dois pontos racionais. É fácil achar exemplos mais interessantes. Um, escolhido, ao acaso nas tabelas de [Cre92], é o seguinte: se E é a curva definida por $y^2 + y = x^3 - x^2 - 2x + 2$, o grupo de Mordell-Weil $E(\boldsymbol{Q})$ é um grupo cíclico infinito, gerado pelo ponto (2,1). (Em outras palavras, todos os pontos racionais desta curva são obtidos a partir de (2,1) através da operação de somar pontos.)

Saber que o grupo de Mordel-Weil é finitamente gerado é interessante, mas para poder realmente usar esta informação nós precisamos achar algum modo de determinar (ou estimar) o número de geradores, ou, ainda melhor, um método para achar os geradores. Ambos estes probelmas continuam em aberto, embora haja conjeturas precisas sobre qual deve ser a resposta. Para muitas curvas específicas, tanto o número quanto os geradores foram completamente determinados (veja, por exemplo, as tabelas em [Cre92]), mas o problema em geral ainda parece bem difícil.

Uma parte importante da estratégia conjetural para se determinar os geradores é considerar, para cada número primo p a redução da curva módulo p. A idéia básica é a seguinte: a equação de uma curva elíptica tem coeficientes inteiros, de modo que é possível "reduzir a equação módulo p", isto é, considerar a equação como uma equação no corpo finito $\boldsymbol{F}_p$(os inteiros módulo p). Isto define um grupo finito[4] $E(\boldsymbol{F}_p)$ cuja estrutura deve ser um pouco mais fácil de analisar que a estrutura do grupo $E(\boldsymbol{Q})$. É uma idéia bem simples, e que dá certo em grande parte, mas várias coisas complicam[5] a nossa vida.

O problema principal é que é perfeitamente possível que a redução módulo p de uma curva elíptica *não* seja uma curva elíptica. Para ver por quê, lembre que para decidir se uma curva é de fato uma curva elíptica (isto é, não tem pontos singulares) nós precisamos verificar que o discriminante Δ não é zero. Mas é claro que é perfeitamente possível que Δ seja diferente de zero (de modo que a curva sobre $\boldsymbol{Q}$ é elíptica) mas divisível por p (de modo que a curva sobre $\boldsymbol{F}_p$ não é). Este fenômeno se chama *má redução*, e é fácil achar exemplos. Considere $p = 5$ e a curva $y^2 = x^3 - 5$. Esta é uma curva elíptica sobre $\boldsymbol{Q}$, mas sua redução módulo 5 vai ser $y^2 = x^3$ que tem uma singularidade. Dizemos, então, que esta curva tem má redução em 5. De fato, o discriminante é $\Delta = -10800$, que é claramente divisível por 2, 3 e 5, de modo que a curva tem má redução em cada um destes. (Nos três casos, a singularidade é uma cúspide).

4 O grupo é finito porque, além do ponto no infinito, há no máximo p^2 outros pontos. Na realidade, o número máximo de pontos é bem menor que isso, mas provar isso é um pouco mais complicado.

5 Pode parecer um pouco pessimista dar tanta atenção às complicações, mas vai ser preciso, mais adiante, que nós tenhamos pelo menos um pouco de informação sobre o que pode acontecer.

Nós precisamos classificar os tipos possíveis de redução, mas há um problema minha que precisamos resolve antes de podermos fazer isso. Considere a curva cuja equação é $y^2 = x^3 - 625x$. A primeira vista, ela parece ainda pior do que a curva que nós consideramos acima, e o descriminante, que é $\Delta = -15625000000$, é *muito* divisível por 5. Mas vejam só: se mudarmos as variáveis pondo $x = 25u = 5^2u$ e $y = 125v = 5^3v$, a equação se torna

$$(5^3v)^2 = (5^2u)^3 - 625(5^2u),$$

que simplifica:

$$5^6v^2 = 5^6u^3 - 5^6u,$$

ou, cancelando o 5^6,

$$v^2 = u^3 - u,$$

que é uma curva elíptica bem comportada que tem *boa* redução em 5. Este exemplo mostra, então, que *curvas que são isomorfas sobre Q podem ter reduções módulo p diferentes.*

Como lidar com isso? O fato é que entre todas as equações possíveis para a nossa curva nós podemos escolher uma equação que é *mínima,* no sentido em que o seu discriminante é divisível por menos primos que o discriminante de qualquer outra equação. Como os primos que dividem o discriminante são os primos onde há má redução, uma equação mínima terá as melhores propriedades de redução possíveis para aquela curva. Quando estudamos as propriedades de redução de uma curva, então, estaremos sempre trabalhando com a equação mínima. (Há algoritmos que permitem partir de uma equação para a curva e obter uma equação mínima, de modo que isto não é muito difícil de fazer.)

Consideremos, então, uma curva elíptica E definida por uma equação mínima. Considerando propriedades de redução, podemos dividir os números primos em três grupos:

- *Primos de boa redução:* aqueles que não são divisores do discriminante da equação mínima. A curva módulo p é uma curva elíptica, e define um grupo $E(F_p)$.
- *Primos de redução multiplicativa:* aqueles para os quais a curva módulo p tem um nó. Se o ponto singular é (x_0,y_0) o conjunto $E(F_p) - \{(x_0,y_0)\}$ tem uma estrutura de grupo, que é isomorfo ou ao grupo multiplicativo $F_p - \{0\}$ ou a um variante do grupo multiplicativo.
- *Primos de redução aditiva:* aqueles para os quais a curva módulo p tem uma cúspide. Se o ponto singular é (x_0,y_0) o conjunto $E(F_p) - \{(x_0,y_0)\}$ tem uma estrutura de grupo, que é isomorfo ao grupo aditivo F_p.

Nenhuma curva tem boa redução em todos os primos, de modo que sempre haverá alguns primos de má redução. Por outro lado, o fato é que redução multiplicativa é um pouco melhor que redução aditiva. (Há várias razões técnicas para se avaliar as coisas desta forma, mas nós não vamos precisar dos detalhes neste artigo.) Se não é possível obter boa redução em todos os primos, podemos pelo menos pedir que a redução nunca seja aditiva. Curvas cuja redução é sempre ou boa ou multiplicativa se chamam *semi-estáveis*.

É conveniente pôr a informação sobre o tipo de redução nos vários primos em um número, chamado o *condutor* da curva. O condutor é definido como um produto $N = \prod p^{n(p)}$ onde p percorre os números primos e

$$n(p) = \begin{cases} 0 & \text{se } E \text{ tem boa redução em } p \\ 1 & \text{se } E \text{ tem redução multiplicativa em } p \\ \geq 2 & \text{se } E \text{ tem redução aditiva em } p \end{cases}$$

(O valor exato do expoente $n(p)$ no caso de redução aditiva depende de propriedades um tanto sutis da redução; na maior parte dos casos, o expoente é simplesmente igual a 2.) O resultado desta definição é que nós podemos detetar, simplesmente olhando o condutor, qual é o tipo de redução em cada primo. Nas tabelas, curvas normalmente são organizadas em ordem crescente do condutor. Vale notar que o condutor é divisível exatamente pelos primos de má redução, que são também os primos que dividem o discriminante.

As curvas elípticas que vão ser mais importantes para nós vão ser as curvas semi-estáveis, cuja redução é sempre ou boa ou multiplicativa. Do ponto de vista do condutor, a exigência é que, para todo número primo p, N não seja divisível por p^2. Tais números costumam ser chamados *livres de quadrados*.

Definição 2. *Uma curva elíptica se diz semi-estável se todas as suas reduções são ou boas ou multiplicativas. Equivalentemente, a curva é semi-estável se o seu condutor for livre de quadrados.*

Será crucial, quando formos aplicar o teorema de Wiles ao teorema de Fermat, verificarmos que certa curva é semi-estável. Como ponto de referência, aqui vão alguns exemplos.

1. Seja E_1 a curva $y^2 = x^3 - 5$ que nós consideramos acima. Verifica-se que a equação é mínima, e que a curva tem redução aditiva em 2, 3, e 5, e portanto não é semi-estável. O condutor é 10800. Neste caso, o condutor é essencialmente o mesmo que o discriminante.

2. Seja E_2 a curva $y^2 + y = x^3 + x$. Esta curva tem redução multiplicativa em 7 e 13 (verificar isto é um bom exercício), e boa redução em todos os outros primos. Assim, E_2 é semi-estável, e o seu condutor é 91.

3. Seja E_3 a curva $y^2 = x^3 + x^2 + 2x + 2$ (a equação é mínima). Ela tem discriminante $\Delta = -1152 = -2^7 \cdot 3^2$ de modo que os primos de má redução são 2 e 3. A redução é aditiva em 2, multiplicativa em 3, e a curva não é semi-estável. O condutor é 384.

4. *Para nós, o exemplo principal:* Sejam a, b, e c inteiros relativamente primos tais que $a + b + c = 0$. Considere a curva E_{abc} cuja equação[6] é $y^2 = x(x - a)(x + b)$. Dependendo dos valores dos parâmetros, esta equação pode não ser mínima; vamos fazer as hipóteses adicionais que $a \equiv -1 \pmod 4$ e $b \equiv 0 \pmod{32}$. Neste caso, a equação *não* é mínima. A equação mínima acaba sendo

$$y^2 + xy = x^3 + \frac{b - a - 1}{4} x^2 - \frac{ab}{16},$$

que se obtém com a mudança de variáveis $x \to 4x$, $y \to 8y + 4x$. A partir daí, não é difícil calcular que o discriminante é $\Delta = a^2b^2c^2/256$ (não deve surpreender: a menos de um fator constante, é o quadrado das diferenças das raízes da cúbica), e verificar que a curva é semi-estável. Os primos de má redução são aqueles que dividem abc (isto se pode ver da equação original, porque se um primo p divide a diferença entre duas raízes, a equação vai ter raiz dupla modulo p), e portanto o condutor é o produto destes primos:

$$N = \prod_{p \mid abc} p$$

(este número é chamado o *radical de abc*). Quando formos ligar o teorema de Wiles ao problema de Fermat, vamos usar uma curva da forma E_{abc} com uma escolha muito especial para os parâmetros a, b, e c.

Há um último ponto da teoria de curvas elípticas que nós precisamos considerar. É interessante pensar sobre o que acontece com o número de pontos do grupo $E(F_p)$ quando nós variamos o primo p. (Para simplificar, olhamos só os primos de boa redução, o que só elimina um número finito de primos.) Uma motivação para esta pergunta é a seguinte: se o grupo de Mordell-Weil $E(Q)$ for grande (i.é, há muitas soluções racionais da equação), esperar-se-ia que para muitos primos p muitos dos pontos de $E(Q)$ "sobreviveriam" a redução módulo p,

6 Pode parecer estranho que c esteja ausente da equação. Tendo em mente, entretanto, que c é completamente determinado por a e b de modo que não é tão crucial incluir c. O ponto crucial é que as raízes do polinômio de terceiro grau que aparece na equação são 0, a, e $-b$ de modo que, a menos de sinal, as diferenças entre as raízes são exatamente a, b, e c. Esta é a propriedade essencial.

e portanto o grupo $E(F_p)$ seria grande. Assim, seria natural conjeturar que se os $E(F_p)$ forem grandes para muitos primos p, o grupo $E(Q)$ também será grande.

Após certa elaboração e refinamento, esta idéia nos levaria à conjetura de Birch e Swinnerton-Dyer, que é muito importante mas fora do nosso assunto. Mas mesmo esta versão grosseira serve para sugerir que a variação do número de pontos de $E(F_p)$ contém informação sobre a aritmética da curva. A melhor maneira de "registrar" essa variação é a seguinte: note, primeiro, que a reta (projetiva) sobre F_p tem exatamente $p + 1$ pontos (os p elementos de F_p mais o ponto do infinito). Tomando isso como o número "normal" de pontos de uma curva sobre F_p nós olhamos $E(F_p)$ para cada p e anotamos quão longe estamos desse padrão. Em termos precisos, dada uma curva elíptica E e um primo de boa redução p, nós definimos um número a_p pela fórmula

$$\#E(F_p) = p + 1 - a_p.$$

Para os outros primos (de má redução), não é muito difícil achar uma maneira de estender a definição: no caso de redução multiplicativa, temos $a_p = \pm 1$ (com uma regra precisa para decidir qual dos dois); para redução aditiva $a_p = 0$.

Uma maneira de fazer uma "tabela" da seqüência dos a_p é usá-los para construir uma função analítica de uma variável complexa chamada a *função L* da curva elíptica E. A notação é $L(E,s)$, onde s é um número complexo, e a idéia é que a seqüência dos a_p determina a função e vice-versa, de modo que se pode esperar que as propriedades dos a_p se traduzam em propriedades da função L. Agora, existem muitas "funções L" em teoria de números, e a experiência com essas funções sugere fazer duas conjeturas. Primeiro, que a função L tem propriedades analíticas parecidas com as de outras funções L. Segundo, que é possível descobrir coisas sobre a curva E estudando sua função L. Esta é uma longa história que não podemos contar aqui, mas que na realidade tem ligações com o que vamos discutir mais adiante. Por agora, basta dizer que este procedimento nos dá uma função

$$L(E,s) = \sum_{n=1}^{\infty} \frac{a_n}{n^s},$$

onde os a_p são exatamente os mesmos que os números que nós determinamos acima, os a_n são calculados a partir dos a_p através de uma expressão da função L como um "produto Euler", e é possível provar que a série converge quando $\mathrm{Re}(s) > 3/2$. Conjetura-se que a função L pode ser extendida a uma função analítica em todo o plano complexo, e que a função, assim extendida, satisfaz uma equação funcional que relaciona seus valores em s e em $2 - s$.

Já é tempo de introduzir os outros personagens da trama e explicar como eles

se relacionam às curvas elípticas. O leitor que desejar se aprofundar nessa teoria tem muitas opções. Como introdução informal, seria possível começar com o artigo [Sil 93], que discute relações entre curvas elípticas e vários problemas envolvendo "somas de dois cubos", inclusive o famoso número do táxi de Ramanujan. Há vários textos introdutórios, inclusive [Cas91], [Hus87], [Kna92], [Sil86], [ST92]. Cada um destes tem qualidades diferentes; o último é dirigido a alunos de graduação.

Além dos textos, o leitor poderá achar interessante explorar vários programas que permitem estudar curvas elípticas com o computador. O programa PARI-GP inclui várias funções especificamente para curvas elípticas, e tais funções podem ser adicionadas a *Mathematica* usando o pacote *Elliptic-CurveCalc* de Silverman, e a *Maple* usando *Apecs*, criado por I. Connell. Cf. [BBCO], [SvM], [Con].

2.3. Formas Modulares

Formas modulares são, inicialmente, objetos analíticos (ou, talvez de forma um pouco mais precisa, objetos ligados à teoria de representações de grupos), mas elas acabam sendo importantes em teoria de números também. Nesta seção, vamos fazer um esboço (muito) breve das definições e explicar a ligação entre formas modulares e curvas elípticas.

Seja $\mathcal{H} = \{x + iy \mid y > 0\}$ o semi-plano superior complexo, isto é, o conjunto dos números complexos cuja parte imaginária é positiva. Não é muito difícil verificar que o grupo $SL_2(\mathbf{Z})$ das matrizes 2×2 com entradas inteiras e de determinante 1 age nos pontos do semiplano da seguinte forma. Se $\gamma \in SL_2(\mathbf{Z})$ é a matriz

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

(de modo que a,b,c,d são inteiros satisfazendo $ad - bc = 1$), e $z \in \mathcal{H}$, definimos

$$\gamma \cdot z = \frac{az + b}{cz + d}.$$

Não é difícil verificar que se $z \in \mathcal{H}$ então $\gamma \cdot z \in \mathcal{H}$, e também que $\gamma_1 \cdot (\gamma_2 \cdot z) = (\gamma_1 \gamma_2) \cdot z$.

Nós queremos considerar funções definidas no semi-plano superior que são "tão invariantes quanto for possível" sob a ação de $SL_2(\mathbf{Z})$, ou talvez sob a ação de algum subgrupo. Os subgrupos que nós vamos considerar são os "subgrupos de congruência" que são definidos exigindo que os termos das matrizes satisfaçam certas congruências. No nosso caso, para cada inteiro positivo N queremos considerar o grupo

$$\Gamma_0(N) = \left\{ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbf{Z}) \mid c \equiv 0 \pmod{N} \right\}.$$

Agora podemos começar a definir formas modulares. Em primeiro lugar, uma forma modular é uma função homolorfa f: $\mathcal{H} \to \boldsymbol{C}$, que se transforma de modo razoável sob a ação de um dos grupos $\Gamma_0(N)$. Especificamente, exigimos que exista um inteiro positivo k tal que

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^k f(z)$$

para toda matriz $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N)$.

Aplicando esta fórmula no caso em que a matriz é

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

nós vemos que uma função deste tipo é periódica: $f(z+1) = f(z)$. Portanto, f tem uma expansão de Fourier

$$f(z) = \sum_{n=-\infty}^{\infty} a_n q^n \qquad \text{onde } q = e^{2\pi i z}.$$

Nós exigimos, para uma forma modular, que esta expressão na realidade inclua apenas potências não-negativas de q (e fazemos a mesma exigência para um número finito de outras expansões semelhantes, que os especialistas chamam de "expansões de Fourier nas outras pontas"). Uma função satisfazendo todas estas condições se chama uma *forma modular de peso k sobre* $\Gamma_0(N)$. O número N é normalmente chamado o *nível* da forma modular f.

É importante, em certos momentos, considerar um sub-espaço especial do espaço das formas modulares de um dado peso e nível. Em vez de exigir que apareçam potências não-negativas apenas, podemos exigir que apareçam apenas potências *positivas* (tanto na expansão principal quanto nas "outras pontas"). Tais formas modulares são chamadas formas modulares *parabólicas*; elas são a parte mais interessante do espaço das formas modulares.

Finalmente, é preciso fazer uma observação sobre a relação entre os espaços de formas modulares de vários níveis: se N divide M, toda forma de nível N (e peso k) produz várias formas de nível M (e peso k). O espaço gerado por todas as formas de nível M e peso k que aparecem desta forma (a partir de formas cujos níveis são divisores próprios de M) é chamado o espaço das *formas velhas* de nível M. O espaço de todas as formas de nível M e peso k tem um produto interno natural, e podemos então considerar o complemento ortogonal do espaço das formas velhas. Este complemento se chama o espaço das *formas novas*, e contém as formas que vão ser mais interessantes para nós.

Até aqui estamos basicamente fazendo análise complexa. O que introduz o

elemento "aritmético" é a existência de uma família de operadores lineares, os *operadores de Hecke*, em cada um dos espaços de formas modulares. Estes operadores comutam entre si, e portanto geram uma álgebra comutativa de operadores. Em vez de dar a definição formal destes operadores (que é natural mas complicada), notamos apenas os seguintes fatos, que são o crucial para nós:

- Para cada inteiro positivo n relativamente primo com o nível N, existe um operador T_n agindo no espaço de formas modulares de peso k e nível N.
- Os operadores T_n comutam uns com os outros.
- Se m e n são relativamente primos, temos $T_{nm} = T_n T_m$.

Nós estaremos especialmente interessados em formas modulares que são autovetores para todos os operadores de Hecke simultaneamente, i.é, formas para as quais existem números λ_n tais que $T_n(f) = \lambda_n f$ para cada n (Sabe-se que os λ_n serão necessariamente números reais pertencendo a um corpo de números algébricos). Nós chamamos formas que têm esta propriedade de *autoformas*.

Isto tudo é bem estranho e complicado, de modo que talvez valha a pena traçar de imediato a conexão entre formas modulares e curva elípticas. Suponha que tenhamos uma forma modular que seja

- de peso 2 e nível N,
- parabólica,
- nova,
- uma autoforma.

Neste caso, podemos, após multiplicar por um fator constante, supor que a expansão de Fourier é

$$f(z) = \sum_{n=1}^{\infty} a_n q^n \qquad \text{com } a_1 = 1.$$

Suponha que, uma vez feita esta normalização,

- todos os coeficientes de Fourier são a_n inteiros.

Então pode-se provar que existe uma curva elíptica cuja equação tem coeficientes inteiros, cujo condutor é N, e cujos a_n (obtidos contando pontos módulo p) são exatamente os mesmos que os a_n que aparecem na expansão de Fourier de f. Em particular, a função L da curva E pode ser determinada a partir de f (através de uma transformação de Mellin), e o fato que f tem boas propriedades analíticas permite provar que a função L tem as propriedades analíticas que nós mencionamos acima: uma continuação analítica para o plano complexo todo e uma equação funcional.

Esta ligação entre formas e curvas elípticas é extremamente poderosa, porque liga análise de um lado e álgebra do outro. Foi natural investigar a questão mais a fundo. O primeiro a sugerir que talvez *todas* as curvas elípticas (com coeficientes inteiros) sejam obtidas a partir de formas modulares foi Y. Taniyama, na década

de 50. A sugestão só penetrou a cultura matemática uma ou duas décadas mais tarde, devido em grande parte ao trabalho de G. Shimura, e ela foi tornada mais precisa por A. Weil, que determinou o papel exato do condutor. Vamos chamá-la de "Conjetura da Modularidade". Aqui está:

Conjetura 1 (Modularidade) *Seja E uma curva elíptica cuja equação tem coeficientes inteiros. Seja N o condutor de E, e para cada n seja a_n o número que aparece na expressão da função L de E. Então existe uma autoforma parabólica de peso 2, nova de nível N, cuja expansão de Fourier é* $\sum a_n q^n$.

Para uma curva específica, não é muito difícil decidir se a conjetura é verdadeira. Pega-se a curva E, e se determina o condutor N e vários dos coeficientes a_n. Como o espaço das formas modulares de peso 2 e nível N é de dimensão finita, saber suficientes a_n determina a forma, e nós podemos examinar o espaço, diagonalizar a ação dos operadores de Hecke, e ver se a forma está lá. (Em geral, dada uma lista de a_n não é possível decidir a priori se $\sum a_n q^n$ é a expansão de uma forma modular, de modo que é preciso proceder ao contrário: acha-se uma base do espaço de todas as formas, e depois procura-se obter a nossa possível forma como uma combinação linear.) Se acharmos uma forma cuja expansão de Fourier tem os a_n corretos (para os n que nós escolhemos), isto é um primeiro passo para achar que a curva satisfaz a conjetura. Para decidir a questão, pode-se usar uma versão do teorema de Cobotarev para achar um inteiro M tal que se os a_n estão certos para $n \leq M$ então todos os a_n estão certos. Na prática, basta verificar a_p onde p é primo. Métodos desse tipo foram usados para verificar a conjetura para um grande número de curvas cujos condutores são pequenos (veja, por exemplo, [Cre92]).

A conjetura da modularidade tem um papel crucial na teoria das curvas elípticas. Curvas para as quais a conjetura é correta são chamadas "curvas elípticas modulares", e um grande número de teoremas sobre curvas elípticas foram provados apenas para curvas que têm esta propriedade. Se a conjetura da modularidade for falsa (o que, depois de Wiles, parece altamente improvável), haverá curvas elípticas que serão completamente misteriosas.

Para mais informações sobre formas modulares, boas referências iniciais são o último capítulo de [Ser73] e as notas [Gou89]. Para mais detalhes, as referências padrão são [Lan76], [Miy80] ou [Shi71]. Quanto à conjetura da modularidade, uma introdução, em termos completamente diferentes, aparece no artigo [Maz91] de Mazur; veja-se também [Lan91].

2.4. Representações Galoisianas

Resta introduzir as representações galoisianas. Começamos com o grupo de Galois de uma extensão do corpo dos números racionais. Para entender este grupo de Galois um pouco melhor, é natural tentar "representar" os elementos do grupo

como matrizes. Em outras palavras, podemos tentar achar um espaço vetorial em que o nosso grupo de Galois age; se tivermos sucesso em construir algo assim, isso nos dará um meio de associar uma matriz a cada elemento do grupo. Pela definição de uma ação de um grupo, esta correspondência será um homomorfismo do grupo de Galois com imagem no grupo de matrizes. Este homomorfismo não precisa ser injetor; quando ele é nós dizemos que a representação é "fiel".

Em vez de trabalhar com extensões específicas de Q é mais fácil trabalhar com o grupo de Galois $G = \mathrm{Gal}\ (\overline{Q}/Q)$ do fecho algébrico de Q . Este é um grupo enorme (tornado um pouco mais manejável porque tem uma topologia natural) que esconde dentro de si uma quantidade enorme de informação aritmética. As representações que nós queremos estudar tomam valores em grupos de matrizes 2×2 sobre vários corpos e anéis, e elas serão (na sua maioria) obtidas a partir de curvas elípticas e de formas modulares.

Para construir uma representação do grupo de Galois a partir de uma curva elíptica, começamos com uma curva E cuja equação tem coeficientes em Z. Escolha um número primo p. Como os pontos de E (com coordenadas complexas) formam um grupo, podemos procurar os pontos de ordem p deste grupo. Pode-se provar que (sobre C) existem exatamente p^2 pontos deste tipo, e eles formam um subgrupo que nós denotaremos $E[p]$. Mais ainda, sabe-se que esse subgrupo é isomorfo ao grupo aditivo do produto de duas cópias dos inteiros módulo p:

$$E\,[p] \cong \boldsymbol{F}_p \times \boldsymbol{F}_p\,.$$

Os pontos que pertencem a $E[p]$ têm, a priori, coordenadas complexas. Um exame mais cuidadoso revela, entretanto, que todas as coordenadas destes pontos pertencem a um corpo que é uma extensão finita dos racionais. Mais ainda, vê-se que transformando as coordenadas de um ponto de ordem p por um elemento do grupo de Galois G obtemos *outro* ponto de ordem p, e esta transformação preserva a estrutura de grupo de $E[p]$. Como $E[p]$ pode ser visto como um espaço vetorial de dimensão dois sobre F_p , isto significa que cada elemento de G age como uma transformação linear neste espaço vetorial, o que resulta numa representação

$$\overline{\rho}_{E,p} : G \to GL_2\,(\boldsymbol{F}_p).$$

(Nós usamos a barra para lembrar que esta é uma representação "módulo p".)

Agora, $GL_2\,(\boldsymbol{F}_p)$ é um grupo finito e G é um grupo muito infinito, de modo que esta representação, embora interessante, não pode ser a história toda. Usando números p-ádicos, conseguimos obter muito mais. Em vez de considerar apenas os pontos de ordem p podemos considerar os pontos de ordem p^n para cada inteiro n. Isto produz uma cadeia de subgrupos

$$E[p] \subset E[p^2] \subset E[p^3] \subset \ldots$$

e portanto uma seqüência de representações, com valores primeiro em $GL_2(F_p)$, depois em $GL_2(\mathbf{Z}/p^2\mathbf{Z})$, depois em $GL_2(\mathbf{Z}/p^3\mathbf{Z})$... Como os p-ádicos são uma forma de encapsular informação módulo p^n para todo n, juntando todas estas representações nós obteremos uma representação p-ádica

$$\rho_{E,p} : G \to GL_2(\mathbf{Q}_p)$$

que contém todas as outras dentro de si. As representações $\rho_{E,p}$ capturam uma grande quantidade de informação aritmética sobre a curva E.

E do lado das formas modulares? O caso fácil é quando se trata de autoformas novas de peso 2 com coeficientes inteiros: a forma corresponde a uma curva elíptica, e portanto, pela construção acima, a uma representação. O caso geral é bem mais difícil, mas o fato se generaliza. É uma conseqüência do trabalho de vários matemáticos (M. Eichler, G. Shimura, P. Deligne, e J.-P. Serre) que a toda autoforma (de qualquer peso) é possível associar uma representação

$$\rho_{f,p} : G \to GL_2(\mathbf{Q}_p)$$

que se liga a f num sentido preciso que é muito técnico para se explicar aqui. (A construção da representação é bastante difícil).

A coisa crucial para nós sabermos é que *quando uma curva elíptica E provém de uma forma modular f, as representações* $\rho_{E,p}$ *e* $\rho_{f,p}$ *são equivalentes.* Melhor ainda, a recíproca também é verdade: dada uma curva E, se conseguirmos achar uma forma modular f tal que as representações $\rho_{E,p}$ e $\rho_{f,p}$ são equivalentes, então E é uma curva elíptica modular.

3 - A Trama

Finalmente estamos no momento de tentar usar toda esta teoria para atacar o Teorema de Fermat. A idéia é assumir que o teorema é falso (portanto, que uma solução existe) e usar esta hipótese para construir uma curva elíptica muito estranha.

3.1. Ligando o Teorema de Fermat às Curvas Elípticas

Começamos, então, supondo que o teorema de Fermat é falso, isto é, que existem três inteiros u, v e w diferentes de zero, tais que $u^p + v^p + w^p = 0$ (como nós já sabemos, basta considerar o caso de um expoente primo, e portanto ímpar, e isto permite reescrever a equação desta forma). Como nós já sabemos que o teorema é verdadeiro para $p = 3$ podemos supor que $p \geq 5$. Portanto supor, também, que u, v e w são relativamente primos, e isto significa que exatamente um deles é par. Reordenando, seja v par. Olhando a equação módulo 4, nós vemos que v^p é divisível por 4, e portanto que um dos dois outros tem que ser congruente

$a - 1$ módulo 4, e o outro tem que ser congruente a 1. Digamos que $u \equiv -1 \pmod 4$.

Vamos usar os três números u^p, v^p e w^p para construir uma curva elíptica, seguindo uma idéia devida a G. Frey (cf. [Fre86], [Fre87a], [Fre87b]). Nós consideramos a curva

$$y^2 = x(x - u^p)\,(x + v^p),$$

que é normalmente chamada de "curva de Frey". Como nós já discutimos acima as curvas E_{abc}, já sabemos um bocado sobre a curva de Frey. Aqui está um resumo:

1. Como v é par e $p \geq 5$, nós temos $v^p \equiv 0 \pmod{32}$. Nós sabemos também que $u^p \equiv -1 \pmod p$. Isto nos posiciona corretamente para usarmos os resultados acima sobre E_{abc}.
2. O discriminante mínimo da curva de Frey é
$$\Delta = \frac{(uvw)^{2p}}{256}.$$
3. O condutor da curva de Frey é o produto de todos os primos que dividem u^p, v^p e w^p ou, o que é o mesmo, o produto de todos os primos que dividem uvw.
4. A curva de Frey é semi-estável.

Na década de 80, Frey observou que esta curva é muito estranha; tão estranha que parece provável que ela não exista. Por exemplo, o condutor desta curva é muito pequeno em relação ao seu discriminante (por causa daquele expoente $2p$). Pior ainda, as representações galoisianas associadas a esta curva são muito esquisitas. Logo havia um número grande conjeturas, cada uma das quais (se fosse provada) implicaria que a curva de Frey não pode existir. Se a curva não pode existir, então os números u, v e w não podem existir, e o teorema de Fermat estaria demonstrado: a equação não tem soluções.

3.2. O "Teorema" é Conseqüência da Conjetura da Modularidade

Frey já tinha sugerido que era improvável que a sua curva pudesse ser modular, e que portanto, se ela existisse, a conjetura da modularidade teria que ser falsa. Faltava uma demonstração sólida de que isto era de fato assim. Serre, numa carta a J.-F. Mestre, formulou um resultado preciso que precisava ser provado para que se pudesse concluir que a curva de Frey não pode ser modular. Na carta (publicada mais tarde como [Ser87a]), Serre descreve a situação com a frase "modularidade + ε implica Fermat". Por causa disso, o teorema que precisava ser provado ficou conhecido, por um tempo como "conjetura epsilon". Esta conjetura foi provada por K. A. Ribet em [Rib90], e com isso a ligação entre

modularidade e o teorema de Fermat ficou estabelecida. (Para mais detalhes, cf. [Lan91].)

O que Serre notou foi que a representação módulo p

$$\overline{\rho}_{E,p} : G \to GL_2(F_p)$$

obtida da curva de Frey era muito estranha. Ela tinha a cara de uma representação que seria obtida a partir de uma forma modular de peso 2, mas quando se aplicava a "receita" usual para adivinhar o nível da forma modular, a resposta obtida era $N = 2$. A forma claramente tinha que ser parabólica. Mas isso levaria a uma contradição, porque *não existem formas modulares parabólicas de peso 2 e nível 2!*

Suponha, então, que existe uma solução da equação de Fermat para algum primo p. Usando essa solução nós construímos uma curva de Frey E. Seja N o condutor de E (que nós calculamos acima). Se a conjetura da modularidade for verdadeira para E, tem que existir uma forma modular de peso 2 e nível N cuja representação galoisiana é a mesma que a da curva E. Temos então a seguinte situação curiosa: uma representação $\overline{\rho}$ que nós sabemos provir de uma forma modular de peso 2 e nível N, mas que *parece*, módulo p vir de uma forma de nível mais baixo.

É aqui que entra o teorema de Ribet: ele provou que (sob certas hipóteses que são verdadeiras no nosso caso) quando isto acontece a forma modular de nível mais baixo tem que existir! Em outras palavras, nestas circunstâncias tem que existir uma forma modular de nível mais baixo (e peso 2) cuja representação, reduzida módulo p é a mesma que aquela com que nós começamos o processo.

Isto permite concluir o seguinte:

Teorema 1 (Ribet) *Suponha que a conjetura da modularidade seja verdadeira para toda curva elíptica semi-estável. Então o teorema de Fermat é verdadeiro.*

Isto é verdade porque, se o teorema de Fermat fosse falso, nós poderíamos escolher uma solução da equação de Fermat e usá-la para construir uma curva de Frey, que seria semi-estável, e portanto modular (já que estamos assumindo a conjetura da modularidade). Assim, existe uma forma modular cuja representação é a mesma da curva. Pelo teorema de Ribet e a observação de Serre, tem que existir uma forma modular parabólica de peso 2 e nível 2 que produz a mesma representação módulo p. Mas isto é uma contradição: tais formas não existem. Logo, a solução da equação de Fermat não existe, e o teorema fica provado.

3.3. Deformando Representações Galoisianas

É agora que chegamos ao trabalho de Wiles. A idéia foi atacar o problema de demonstrar a conjetura da modularidade usando as representações galoisianas, e

em particular usando "deformações" de representações galoisianas. Como nós observamos acima, uma representação p-ádica pode ser entendida como um "compêndio" de representações módulo p^n para todo $n \geq 1$. A representação módulo p contida no compêndio se chama a "redução" da representação p-ádica. A idéia fundamental é inverter este processo: dada uma representação módulo p consideramos *todas* as representações p-ádicas cuja redução é a nossa representação inicial. Estas podem ser descritas como "deformações" porque, do ponto de vista p-ádico elas estão "perto" da representação inicial.

O projeto de classificar deformações foi introduzido por B. Mazur em [Maz89]. Mazur mostrou que em muitos casos existe uma "deformação universal," isto é, uma representação em $GL_2(R)$ onde R é um anel enorme, tal que todas as deformações possíveis são determinadas por esta. Se nós sabemos que a representação módulo p vem de uma forma modular, então nós podemos construir outro anel e outra representação, esta contendo todas as deformações que vêm de formas modulares. A teoria geral de deformações nos dá um homomorfismo entre estes dois anéis, e podemos então tentar mostrar que o homomorfismo é mesmo um isomorfismo. Se for assim, segue que todas as deformações são modulares.

É esta a idéia básica do artigo de Wiles. Ele considera, não todas as deformações, mas apenas aquelas cujas propriedades sugerem que é possível que elas sejam associadas a formas modulares de peso 2. Isto dá um anel e uma deformação universal. (Ela é universal entre as deformações que satisfazem as condições de Wiles.) Do lado modular, Wiles constrói, a partir do anel dos operadores de Hecke, um anel e uma deformação que pode (de forma um tanto imprecisa) ser considerada universal entre as deformações que provém de formas modulares de peso 2. Wiles então se propõe a demonstrar, usando um arsenal de idéias importantes descobertas nos últimos anos, que estes anéis são isomorfos.

Não é muito difícil ver que o homomorfismo ligando os dois anéis é sobrejetor. A dificuldade é provar que ele também é injetor. Para fazer isso, Wiles desenvolveu duas estratégias. A primeira estratégia é baseada em cálculos de grupos de cohomologia. Basicamente, ele mostra que para provar o isomorfismo entre os anéis basta provar uma desigualdade limitando o tamanho de um certo grupo de cohomologia. Embora não tenha sido possível provar esta desigualdade diretamente,[7] este ponto de vista permite fazer uma simplificação: Wiles mostra que se o teorema for demonstrado num caso especial (o caso da "deformação mínima"), é possível deduzir o teorema geral.

Resta, então, provar o teorema para a deformação mínima. Neste caso, usa-se o outro critério de Wiles. Lembre que um dos anéis em consideração é determinado

7 A demonstração original procurava provar a desigualdade diretamente usando um "sistema de Euler geométrico." Infelizmente, o argumento parece não funcionar, e foi substituído pelo argumento descrito adiante.

a partir do anel dos operadores de Hecke. Vamos chamá-lo de "anel de Hecke local." O outro critério diz que se o anel de Hecke local for um anel de interseção completa,[8] então os dois anéis são isomorfos.

É este fato que é provado no artigo conjunto de Taylor e Wiles: o anel de Hecke local em questão é um anel de interseção completa. A demonstração depende da construção de uma família infinita de anéis de deformação e de anéis de Hecke satisfazendo várias condições técnicas, e deduz o fato crucial através de uma espécie de "passagem ao limite."

No final, temos uma demonstração de que os dois anéis são de fato isomorfos. Traduzindo de volta à linguagem de representações, isto significa que se começarmos com uma representação módulo p que satisfaz certas condições técnicas *e que é modular*, então todas as suas deformações (do tipo considerado por Wiles) também são modulares.

3.4. Juntando tudo...

Agora suponha que tenhamos uma curva elíptica E cuja representação módulo p nós saibamos (de algum modo) ser associada a uma forma modular. (Note que não é E que estamos supondo ser modular, mas só esta representação módulo p.) Suponha também que esta representação satisfaça as condições técnicas do teorema de Wiles. Então podemos concluir que toda deformação desta representação é modular. Mas a representação p-ádica $\rho_{E,p}$ é uma deformação! Segue que esta representação é modular, e portanto que E é modular.

Para concluir que uma curva elíptica é modular, então, basta achar uma semente: precisamos achar um jeito de provar que a representação módulo p associada a E é modular, e aí podemos argumentar como acima. O primo p nós podemos escolher à vontade, e é isto que faz o argumento de Wiles funcionar.

Começamos com uma curva elíptica semi-estável, e olhamos a representação módulo 3 obtida da curva. Há duas possibilidades. A representação, como observamos acima, vem de uma ação do grupo de Galois no espaço vetorial $F_3 \times F_3$. Pode acontecer que haja um sub-espaço invariante sob todos os elementos do grupo de Galois. Neste caso, se diz que a representação é *redutível*; caso contrário, que ela é *irredutível.*

Muito bem, consideremos $\overline{\rho}_{E,3}$. Esta representação pode ou não ser irredutível. Se ela for, Wiles lança mão de um teorema famoso de R. P. Langlands e J. Tunnell (cf. [Lan80], [Tun81]) para mostrar que a representação é modular. Segue da teoria de deformação, então, que a curva E é modular.

8 Anéis de interseção completa são um tema tradicional da álgebra comutativa. Basicamente, eles são a versão algébrica da idéia de se determinar um conjunto geométrico construindo a interseção do "número certo" de hipersuperfícies. A definição é simples mas técnica, de modo que remetemos o leitor interessado às referências canônicas sobre álgebra comutativa.

Se $\overline{\rho}_{E,3}$ não for irredutível, Wiles usa um truque. Ele mostra que existe uma outra curva E' que tem a mesma representação módulo 5 que a nossa curva original, mas cuja representação módulo 3 *é* irredutível. Pelo argumento do primeiro caso, E' é modular. Logo, sua representação módulo 5 vem de uma forma modular. Mas esta é a mesma representação que a representação módulo 5 de E. Logo, podemos usar a teoria de deformações para concluir que E é modular também.

O resultado final é:

Teorema 2 (Wiles, Taylor) *Toda curva elíptica semi-estável é modular.*

Logo, como a curva de Frey seria semi-estável mas não modular,

Corolário 1 (Fermat, Taylor, Wiles) *Dado $n \geq 3$, não existem soluções inteiras não-triviais da equação $x^n + y^n = z^n$.*

É claro que este é apenas *um* dos corolários do teorema de Wiles, e não há dúvida que outros ainda estão por vir. Por exemplo, como Serre observou em [Ser87b], pode-se usar as idéias a Frey a outras equações diofantinas que são tão difíceis quanto a equação de Fermat. Considere, por exemplo, uma equação da forma

$$x^p + y^p = Mz^p,$$

onde p é primo e M é um inteiro. O argumento de Serre, mais o teorema de Wiles, mostram que:

Corolário 2 *Seja p um número primo e seja M uma potência de um dos seguintes primos:*

3, 5, 7,11, 13, 17, 19, 23, 29,53, 59.

Suponha que $p \geq 11$ e que p não é divisor de M. Então não existem soluções inteiras não-triviais da equação

$$x^p + y^p = Mz^p.$$

A demonstração é completamente análoga do que já fizemos: dada uma solução, construimos uma curva de Frey, e consideramos a forma modular correspondente. Usando o teorema de Ribet para reduzir o nível, obtemos uma previsão de que existe uma forma modular de peso 2 e nível pequeno. Aí estudamos as formas modulares daquele nível para ver se a forma predita por Ribet de fato existe. Se não há tal forma, não há solução da equação.

Além de aplicações, também já existem extensões do teorema. A mais espetacular se deve a F. Diamond, que demonstrou que a conjetura de modulari-

dade vale para curvas elípticas que têm redução boa ou multiplicativa em 3 e em 5 (podendo ter redução aditiva nos outros primos). Isto chega perto de demonstrar a conjetura completamente!

A comunidade matemática começou logo o processo de reflexão sobre os métodos e os resultados de Wiles, e vários artigos (e até livros) já foram escritos. Talvez o mais útil nesta altura seja o artigo de Ribet [Rib95], que contém uma discussão mais técnica das idéias centrais e inúmeras referências bibliográficas.

Reunidos, os trabalhos de Ribet, Wiles e Taylor são uma "demonstração maravilhosa" que não só resolve o problema que Fermat nos legou mas abre novos horizontes para a teoria de números. Trata-se de fato de um grande feito.

Referências

[Ami75] Y. Amice, *Les nombres p-adiques,* Press Universitaires de France, Paris, 1975.

[BBCO] C. Batut, D. Bernardi, H. Cohen, and M. Olivier, *GP-PARI,* disponível por "ftp" em megrez. math. u-bordeaux. fr.

[BCS92] J. P. Buhler, R. E. Crandall, and R. W. Sompolski, *Irregular primes to one million,* Math. Comp. 59 (1992), 717-722.

[Cas86] J. W. S. Cassels, *Local fields*, Cambridge University Press, Cambridge, 1986.

[Cas91] J. W. S. Cassels, *Lectures on elliptic curves,* Cambridge University Press, Cambridge, 1991.

[Con] Ian Connell, *APECS: Arithmetic of Plane Elliptic Curves,* Disponível por "ftp" em math. mcgill. ca. Depende de *Maple.*

[Cre92] J.E. Cremona, *Algorithms for modular elliptic curves,* Cambridge University Press, Cambridge, 1992.

[CS86] Gary Cornell and Joseph H. Silverman (eds.), *Arithmetic geometry,* Springer-Verlag, Berlin, Heidelberg, New York, 1986.

[Fre86] G. Frey, *Links between stable elliptic curves and certain diophantine equations,* Annales Univesitatis Saraviensis, Series Math. 1 (1986), 1-40.

[Fre87a] G. Frey, *Links between elliptic curves and solutions of A - B = C* J. Indian Math. Soc. 51 (1987), 117-145.

[Fre87b] G. Frey, *Links between solutions of A - B = C and elliptic curves,* Number Theory, Ulm 1987 (Berlin, Heidelberg, New York) (H. P. Schlickewei and E. Wirsing eds.), Lecture Notes in Mathematics, vol. 1380, Springer-Verlag, 1987.

[Gou89] Fernando Q. Gouvêa, *Formas modulares: uma introdução,* Monografias de Matemática 47, Instituto de Matemática Pura e Aplicada, Rio de Janeiro, 1989.

[Gou93] Fernando Q. Gouvêa, *p-adic numbers: an introduction,* Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[Hus87] Dale Husemöller, *Elliptic curves,* Springer-Verlag, Berlin, Heidelberg, New York, 1987.

[Kna92] Anthony W. Knapp, *Elliptic curves,* Princeton University Press, Princeton, 1992.

[Kob84] N. Koblitz, *p-adic numbers, p-adic analysis, and zeta-functions,* second ed., Springer-Verlag, Berlin, Heidelberg, New York, 1984.

[Lan76] Serge Lang, *Introduction to modular forms,* Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[Lan80] R. P. Langlands, *Base change for* GL(2), Ann. of Math. Stud., vol. 96, Princeton University Press, Princeton, NJ,, 1980.

[Lan91] Serge Lang, *Number theory III,* Encyclopedia of Mathematical Sciences, vol. 60, Springer-Verlag, Berlin, Heidelberg, New York, 1991.

[Maz89] Barry Mazur, *Deforming Galois representations,* Galois Groups Over $\mathbb{Q}$ (Berlin, Heidelberg, New York) (Y. Ihara, K. A. Ribet, and J.-P. Serre, eds.), Springer-Verlag, 1989.

[Maz91] Barry Mazur, *Number theory as gadfly,* Amer. Math. Monthly 98 (1991), 593-610.

[Miy89] Toshitsune Miyake, *Modular forms,* Springer-Verlag, 1989.

[Rib79] Paulo Ribenboim, *13 lectures on Fermat's Last Theorem,* Springer-Verlag, Berlin, Heidelberg, New York, 1979.

[Rib90] Kenneth A. Ribet, *On modular representations of* Gal $\overline{\mathbb{Q}}/\mathbb{Q}$ *arising from modular forms,* Invent. Math. 100 (1990), 431-476.

[Rib95] K. A. Ribet, *Galois representations and modular forms,* Bull. Amer. Math. Soc. (N.S.) 32 (1995), 375-402.

[Ser73] Jean-Pierre Serre, *A course in arithmetic,* Springer-Verlag, Berlin, Heidelberg, New York, 1973.

[Ser87a] Jean-Pierre Serre, *Lettre à J-F Mestre,* Current Trends in Arithmetical Algebraic Geometry (Providence) (Kenneth A. Ribet, ed.), Contemporary Mathematics, vol. 67, American Mathematical Society, 1987.

[Ser87b] Jean-Pierre Serre, *Sur les représentations modulaires de degré 2 de* Gal $\overline{\mathbb{Q}}/\mathbb{Q}$ Duke Math. J. 54 (1987), 179-230.

[Shi71] G. Shimura, *Introduction to the arithmetic theory of automorphic forms,* Princeton University Press, 1971.

[Sil86] Joseph H. Silverman, *The arithmetic of elliptic curves,* Springer-Verlag, Berlin, Heidelberg, New York, 1986.

[Sil93] Joseph H. Silverman, *Taxicabs and sums of two cubes,* Amer. Math. Monthly 100 (1993), nº 4, 331-340.

[ST92] Joseph H. Silverman and John Tate, *Rational points on elliptic curves,* Springer-Verlag, Berlin, Heidelberg, New York, 1992.

[SvM] J. H. Silverman and P. van Mulbregt, *Elliptic curve calculator,* Disponível por "ftp" em gauss. math. brown. edu. Depende de *Mathematica.*

[Tun81] J. Tunnell, *Artin's conjecture for representations of octahedral type,* Bull, Amer, Math. Soc. (N.S.) 5 (1981), 173-175.

[Was82] Larry C. Washington, *Introduction to cyclotomic fields,* Springer-Verlag, Berlin, Heidelberg, New York, 1982.

[Wei83] Andre Weil, *Number theory: an approach through history, from Hammurapi to Legendre,* Birkhäuser, 1983.

Colby College
Department of Mathematics and Computer Science
Waterville, ME 04901
e-mail: fqgouvea@colby.edu