Why does the 'hack' for 'squawk' work?

A breezy account of a famous theorem

Doug Hensley

The 'hack' is that if you want to know whether a positive integer u can be written as the sum of two squares, $u = a^2 + b^2$ say, there's a better way to get to the yes or no than to search possible values of a and b until you find a pair that works or run out of possibilities.

You factor u and look at its odd prime factors p. Some of them may be in the arithmetic progression

$$3, 7, 11, (15), 19, 23, (27), 31, (35), \ldots .$$

These are the numbers that are congruent to 3 mod 4, in other words, if you divide any of them by 4 in quotient and remainder arithmetic, the remainder is 3. The hack, or the theorem, is that among positive integers, those that have primes of that form only to even powers (if at all) as factors, and only those, can be written as $a^2 + b^2$. Thus, 205 is a 'squawk' number, but 204 is not, because it is divisible by 3 but not by 9.

A prime p with $p \equiv 3 \pmod 4$ cannot be written as $p = a^2 + b^2$ because a^2 can only be 0 or 1 mod 4, which means that $a^2 + b^2$ can only be 0 or 1 or 2 mod 4, never 3.

It's not so obvious that a number n which has as a factor such a prime cannot be written that way. It's not even true! Consider 45. It's divisible by 3, yet $45 = 9 + 36$. But that's cheap. We're just piggybacking on $5 = 1 + 4$. What we should have said is this:

If $p \mid u$ and $p \equiv 3 \pmod 4$, and the number of powers of p in u is odd, then u cannot be written as $a^2 + b^2$.

This claim is at least true, but that doesn't make it obvious. The numerical evidence for it is good, as some pencil-and-paper work will confirm. In mathematics, though, it can happen that numerical evidence is good, and yet, somewhere out in the wilderness of big numbers, there is a counterexample. So we need a proof. Besides, that's what we do: we prove things. So here goes.

We begin with a lemma. *If $p \equiv 3 \pmod 4$, then there is no integer x such that $p \mid (x^2 + 1)$.*

Equivalently, $x^2 \not\equiv -1 \pmod p$. Suppose there were an x with $x^2 \equiv -1 \pmod p$. We are assuming $p = 4n + 3$ for some nonnegative integer n. Consider $y = x^{4n+2}$. Clearly $y \equiv -1 \pmod p$. On the other hand, from Fermat's little theorem, $x^{p-1} \equiv 1 \pmod p$. But then y is congruent both to 1, and to -1, mod p, a contradiction. This proves the lemma.

Now suppose we had $a^2 + b^2 = u$, with $p \equiv 3 \pmod 4$ and p dividing u to an odd power. We would then have $a^2 + b^2$ exactly divisible by an odd power of p. Perhaps both a and b would be divisible by p. If so, we divide out however many powers of p we can from both a and b, arriving at new integers c and d so that $c^2 + d^2$ is divisible by an odd power of p, and at least one of c and d is not

a multiple of p. But then neither is, for if one is and one is not, then $c^2 + d^2$ would not be a multiple of p.

At this point, we have a nontrivial pair (c, d) so that $c^2 + d^2 \equiv 0 \pmod{p}$.

In arithmetic modulo a prime, integers (such as d) not congruent to 0 have multiplicative inverses. Thus, mod 11, $5 \cdot 9 \equiv 1$. Multiplying by the square of the inverse of d, we get $(d^{-1}c)^2 + 1 \equiv 0 \pmod{p}$. But that makes $d^{-1}c$ into exactly the x we just proved cannot be. This proves that when u has an odd power of a prime of type $4n + 3$, u cannot be written as $a^2 + b^2$.

Now, let's look at 2, which plays a special but happily simple role. The fact is that if u is even, then u can be written as $a^2 + b^2$ if and only if $u/2$ can be written that way. So let $u = 2v$ and suppose first that $v = a^2 + b^2$. Then $2v = (a - b)^2 + (a + b)^2$ (multiply it out!). On the other hand,m if $u = a^2 + b^2$ then, because u is even, a and b are both even, or both odd. Let $c = (a - b)/2$, and $d = (a + b)/2$. Both c and d are integers, and $c^2 + d^2 = u/2 = v$. So powers of 2 are neutral. But it's better than that—they are computably neutral, to coin a phrase. If we have a pair that works for v, we can compute the pair that works for $2v$, and vice-versa.

This brings us to the question of what happens when u is odd, and all the (odd) primes p that divide u to an odd power are congruent to 1 mod 4. These are numbers that our numerical evidence suggests can all be written as $u = a^2 + b^2$.

Again, we begin by thinking about $x^2 \equiv -1 \pmod{p}$, but this time, with $p \equiv 1 \pmod{4}$. The proof we gave for the other case, that this couldn't happen, doesn't work here, but that alone is not sufficient reason why there must be solutions x.

Here's a handy little fact: if $a^2 \equiv b^2 \pmod{p}$, then $(a - b)(a + b) \equiv 0 \pmod{p}$. Thus, because p is prime, one of $(a - b)$ or $(a + b)$ is 0 mod p. In other words, the only way that $a^2 \equiv b^2 \pmod{p}$ is when $a \equiv \pm b \pmod{p}$.

We begin this part of the proof with another theorem about primes: for any p, $(p - 1)! \equiv -1 \pmod{p}$. This is because the numbers from 2 to $p - 2$ can be paired off into (number, inverse) with none of them being its own inverse. The product of all those, then, resolves into products of pairs each of which multiply to 1 mod p, and that leaves 1 and $p - 1$. The product of the whole batch resolves into 1, times a bunch of 1's, times -1, mod p.

Now when $p = 4n + 1$, there's another perspective. We write

$$(p - 1)! \equiv 1(-1)(2)(-2) \cdots (2n)(-2n) \equiv -1 \pmod{p}.$$

Thus

$$\prod_{k=2}^{2n} (-k^2) \equiv 1$$

, so

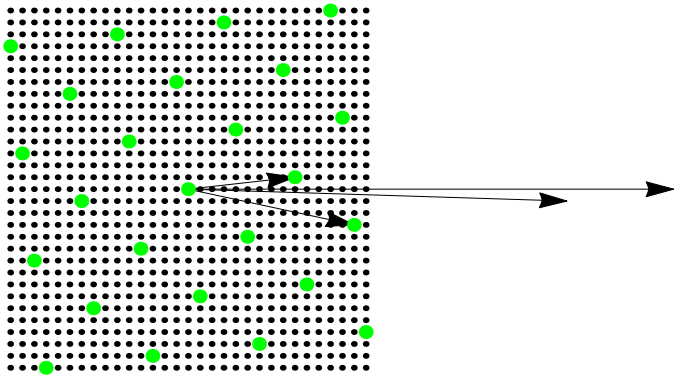
$$\prod_{k=2}^{2n} k^2 \equiv -1.$$

(Here, $\prod$ means multiply, just like $\sum$ means add.) Let $x = \prod_{k=2}^{2n} k = (2n)!$. Then, as we have just shown, $x^2 \equiv -1 \pmod{p}$.

Computationally, it would be somewhat difficult to arrive at x this way if n is large. Happily, there's another approach. Recall Fermat's little theorem that $y^{p-1} \equiv 1 \pmod{p}$? There's a more refined version of this: $y^{(p-1)/2} \equiv \pm 1 \pmod{p}$. As it happens, though we will not go into the details of why, half the values of y give -1, and half give 1. It is easy to compute high powers of y mod p, because we can write $(p-1)/2$ as a sum of powers of 2, and it is easy to square a number repeatedly mod p. So if we just thrash around like a fish out of water, we shall eventually luck upon a y so that $y^{(p-1)/2} \equiv -1$. With $p = 4n+1$, that means that $y^2 n \equiv -1$, so that $(y^n)^2 \equiv -1$. In other words, not only does there exist x so that $x^2 \equiv -1 \bmod p$ when $p = 4n+1$, this x is not so hard to find, if only we have a bit of luck.

OK, so what? How does this let us find $a^2 + b^2 = p$? Well, let's think about vectors. Let's think about an example. [This example is so typical of the general situation that once we have it figured out, the general case becomes evident.] Consider $p = 41$. If we try $y = 2$, we have $2^5 \equiv 32 \equiv -9 \pmod{41}$, so $2^{10} \equiv 81 \equiv 40 \equiv -1 \pmod{41}$. Thus $2^{20} \equiv 1$. Out of luck. It works out that $3^{10} \equiv 9 \pmod{41}$, so 9 is our square root of minus 1 mod 41. That means that the vector $(9, 1)$, and while we're at it, the vector $(41, 0)$, both serve as pairs of numbers (c, d) such that $c^2 + d^2 \equiv 0 \pmod{41}$.

Suppose we add these vectors? What happens? Yep, the new vector shares this property.



The set L of combinations $s(9,1)+t(41,0)$ forms a sublattice of the set of integer points in the plane. In every row, 1 of every 41 integer points is in L, because at row k, $k(9,1)$ is in the row, so also $k(9,1) + j(41,0)$ for any j.

Thus overall, one of every 41 integer points belongs to L. Why should one of these points lie on the circle $x^2 + y^2 = 41$? Well, if not, then because 41 divides $x^2 + y^2$ for (x, y) in L, there wouldn't be any inside the circle $x^2 + y^2 = 82$. But as we've seen, $2p$ and p go together: if one can be represented, so can the other. So now we're out to $x^2 + y^2 = 164$. But this circle is too big to avoid capturing, inside it, one of the lattice points of L. Any (two dimensional)

lattice can be described as the integer combinations of some two of its vectors. We started using the vector pair $(41, 0)$, $(9, 1)$ as a *basis*. But we could better use $(41, 0) - (9, 1)$, $(9, 1)$ as a basis. The area of the parallelogram built using the origin, those two, and their sum, as corners, is the same as the area of the original, skinnier parallelogram. The combinations of these two form the same set L as the combinations of the original pair.

In this way, we work back to $(32, -1)$ and $(9, 1)$, $(23, -2)$ and $(9, 1)$, $(14, -3)$ and $(9, 1)$, $(5, -4)$ and $(9, 1)$, and now, though we cannot get any further with this one-pony show, we can switch gears and subtract $(5, -4)$ from $(9, 1)$ to get a new basis $\{(5, -4), (4, 5)\}$ for L. At this point we have a couple of short-as-possible vectors in L. The angle between the two vectors is a perfect 90 degrees, but in any reduced basis, the angle would be somewhere between 60 and 90 degrees. With the two basis vectors being more or less, at any rate, (and in our example and in general in this particular setting, exactly) perpendicular, the area of a single tile in the lattice is not far from (equal to) the product of the lengths of the vectors. If both vectors of the reduced basis lay outside $x^2 + y^2 = 4p$, the area would be too large. The product of the lengths would be at least $4p$, while actually the area is exactly p.

There's another perspective on this. The *Gaussian integers* are numbers of the form $a + ib$ where i is the square root of -1 and a and b are ordinary integers. Gaussian integers obey the ususal rules of integer arithmetic, including, crucially, unique factorization into primes. But primes in the Gaussian integers still can spring a few surprises. First, in place of ± 1 which don't count as primes or as parts of a factorization, we have ± 1 and $\pm i$. Second, in the context of Gaussian integers, 3 is prime, but 5 is not, because $5 = (2 + i)(2 - i)$.

Associated with a Gaussian integer $a+ib$ we have its *norm* a^2+b^2. The norm of a product of Gaussian integers is the product of the norms. Any ordinary integer that can be written $a^2 + b^2$ is the norm of a Gaussian integer. So all we have to do is to work out which norms can occur. The green dots we were seeing in the graphic can be reinterpreted as being all the multiples, with the Gaussian integers, of $5 - 4i$. Since we can swap signs around and not change $a^2 + b^2$, we get another batch of solutions, the Gaussian integer multiples of $5 + 4i$. The two are not multiples of each other.

This perspective makes obvious what is otherwise a bit tricky: if u and v can be written as $a^2 + b^2$, then so can uv. The classical proof of this is to observe that if $u = a^2 + b^2$ and $v = c^2 + d^2$ then

$$uv = (ac - bd)^2 + (ad + bc)^2.$$

How did they think of that? Well, they had time enough to fool around with it and try things. But the easy way to come up with this is just to multiply out $(a + ib)(c + id)$.