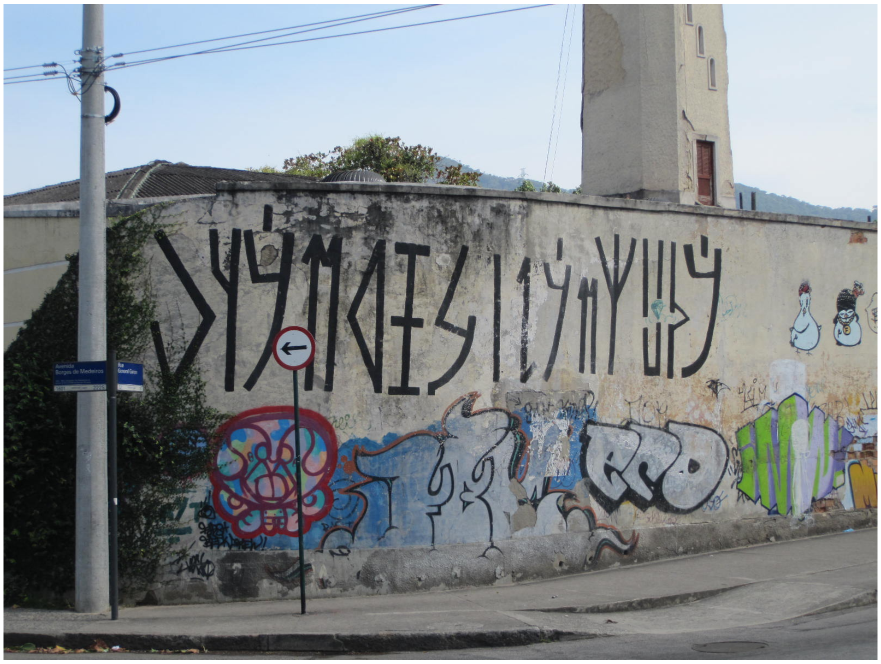
The Probabilities in Pixo

Paulo Orenstein

Published in[1]: *Walls of air: Brazilian Pavilion (Venice Biennale 2018)*,
Sol Camacho, editor, São Paulo: Bienal de São Paulo, 2018.

I can still vividly recall the first time I saw the mysterious symbols, hidden in an underpass in the heart of Rio de Janeiro. It was clearly a language to conceal meaning, with characters that combined the otherworldliness of hieroglyphs and the decisiveness of runes. They could be found across the city's walls in crowded streets, abandoned alleys, by the banks of a lagoon or facing the city's Botanical Gardens. The strokes themselves had a geometric precision to them, buried in a haze of meaningless characters. One could spend a long time contemplating them, but to me, more than just art over a concrete canvas, those symbols were a tantalizing, sprawling puzzle.



[1] Available at `https://issuu.com/murosdear_wallsofair/docs/16a-mia_catalogo_eng_low_completo`.

Satélite
O energético
escolhido pelo
Rock in Rio
CH
CAROLINA HERRERA
L'eau, the new feminine fragrance
L'eau
FRAGRANCE
FRAGRANCE Shopping Leblon • Avenida Afrânio de Mello Franco, 290 - loja 211 H • 2º Piso
FRAGRANCE Shopping da Gávea • Rua Marquês de São Vicente, 52 - loja 178 • 1º Andar

3 6 0°
PLAY
the
GAME
H
A
R

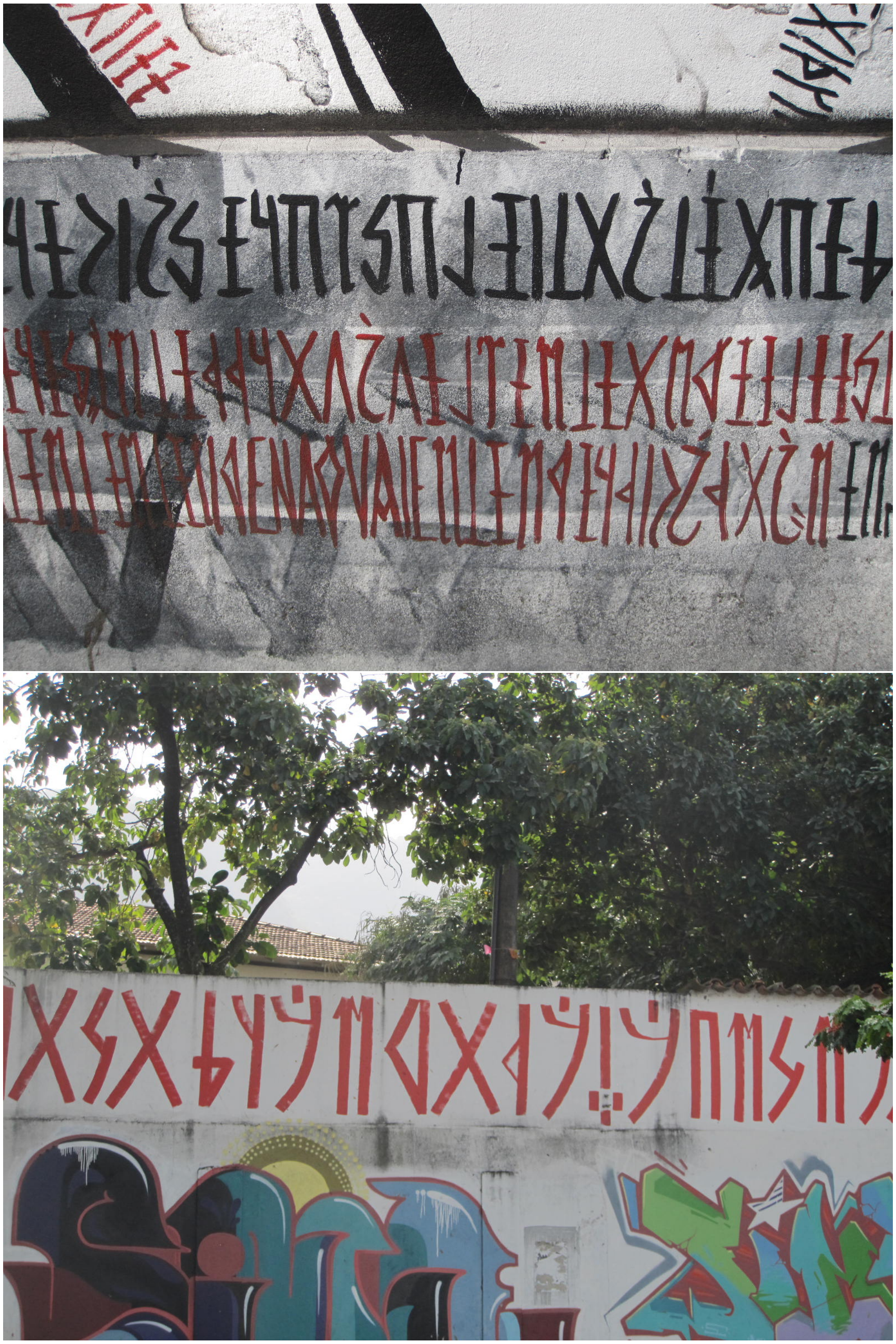


Figure 1: examples of the ciphered symbols across Rio de Janeiro

It was 2011 and I was halfway through my undergraduate studies. A professor close to me had read an article about an enigmatic artist, called Joana César, who wrote the ciphered inscriptions I had seen throughout the city's walls. Pixo is not usually written to be broadly understood, but this was different: these weren't disfigured letters upon the walls, but an entirely new alphabet. In the article, Joana said she was laying bare all of her innermost feelings for the city to see, but hidden so no one could read it. Her diary was coloring Rio's urban landscape: she wrote in giant letters about her childhood goals, failed dreams, grievances, recollections and wishes, even erotic fantasies. Yet, no one could know. She was playing a game of hide and seek with the entire city.

The professor, Carlos Tomei, challenged me and one of his postdocs, Juliana Freire, to make sense of this intimate and intentional mess. After all, seen through the right prism, this was a fascinating mathematical problem. His advice: to read through the muddle of characters using the algorithmic precision of a computer.

So how does one turn street symbols into mathematics, and then mathematics into language? Years prior, a Stanford professor wrote a paper suggesting a way to do it[2]. The author, Persi Diaconis, along with one of this students, was trying to read ciphered messages exchanged by inmates in a Californian prison; our task didn't seem that much different. And, as mathematical ideas run at an abstract level, they can be reshaped to fit many purposes. In our case, we wanted to make this distinct kind of pixo legible.

Here is one way to start: collect all the characters used by the artist. For now, let's assume we have 26 symbols, just like our alphabet, and fix them in any order we want (see Figure 2). We can write all possible ways of organizing the 26 letters in our alphabet to match the cipher: [abcdefghijklmnopqrstuvwxyz] is one, where the first symbol is 'a', the second 'b' , etc; [bacdefghijklmnopqrstuvwxyz] is another, where the first symbol is 'b', the second 'a', etc; and [mlpnkobjivhucgyxftzdrseawq] is yet another, where the first symbol is now 'm', the second 'l', etc. Of course, there are many ways to organize these 26 letters, but we know one of them must be the actual cipher being used by the artist[3].

We have now turned the symbols into mathematics: the task of finding the right cipher is as simple as finding the right configuration of the 26 letters in the alphabet. We can think of each possible combination as a point in space (see Figure 3). We could then just get a sample of the artist's writings on the wall, and use a computer to try each possible configuration to turn her symbols into our alphabet.

[2] Diaconis, Persi. "The Markov chain Monte Carlo revolution." *Bulletin of the American Mathematical Society 46.2* (2009): 179-205.

[3] There are many other encryption methods she could have used, besides the plain substitution of a symbol for a letter; she had told a reporter that she invented this alphabet when she was very young, which lead us to believe this was such a simple cipher, known as a *substitution cipher*.

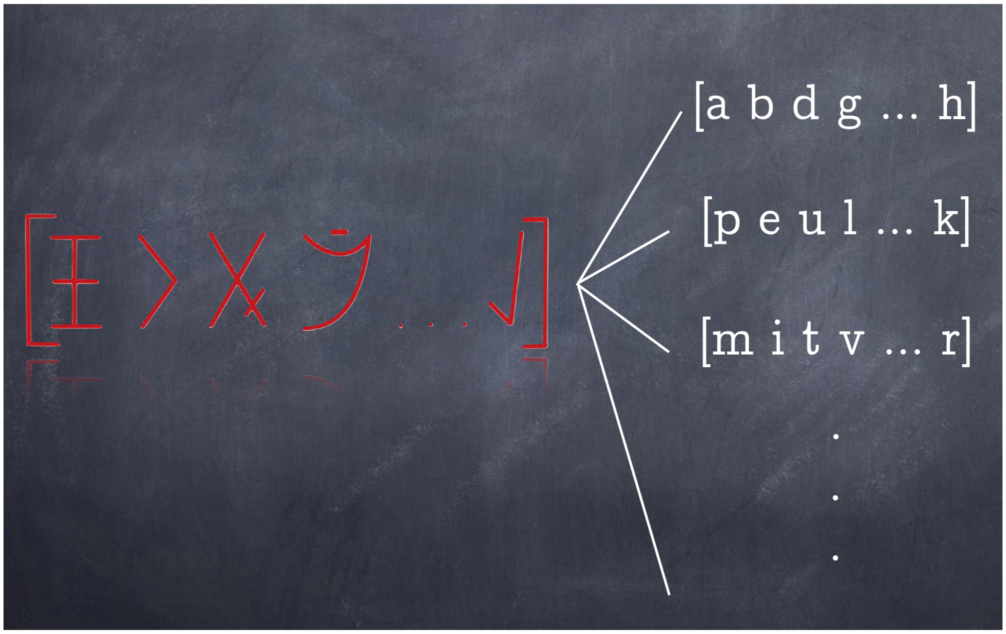


Figure 2: transforming the symbols back to our alphabet, with many possible ciphers

Most of the attempts will be incorrect, and return a completely meaningless array of letters, but, once we stumble upon the correct one, we should just get back faultless Portuguese.

The problem with this approach is that there are many, many possible combinations of the 26 letters in our alphabet to try. If we went about decoding at random, even with a computer trying a million ciphers per second (and a person checking whether the result resembles Portuguese at the same speed!), it would take far longer than a trillion years to get it done. What could be a better way?

First, while computers are really fast, humans are generally not. Hence, it would be helpful if we could teach the computer to automatically recognize Portuguese. That way, once it tries decoding the artist's symbols with a given cipher, it can automatically detect whether what it reads looks like Portuguese or not (much like Figure 4, with runic characters being translated to English). Put another way, we wanted the computer to look at a collection of letters and decide how much it resembles Portuguese as opposed to just a random string of characters. We had to give the computer a way to assign a number, or a grade, that should be high if the characters could be coming from a Portuguese text, and low if it couldn't.

Intuitively, if pairs of letters that appear often in the decoded text also appear often in Portuguese, then it is more likely the decrypted text is in Portuguese. In English, the most popular pairs are 'th', 'er', 'on', 'an', and a configuration used for decryption should be deemed more plausible if the text it outputs

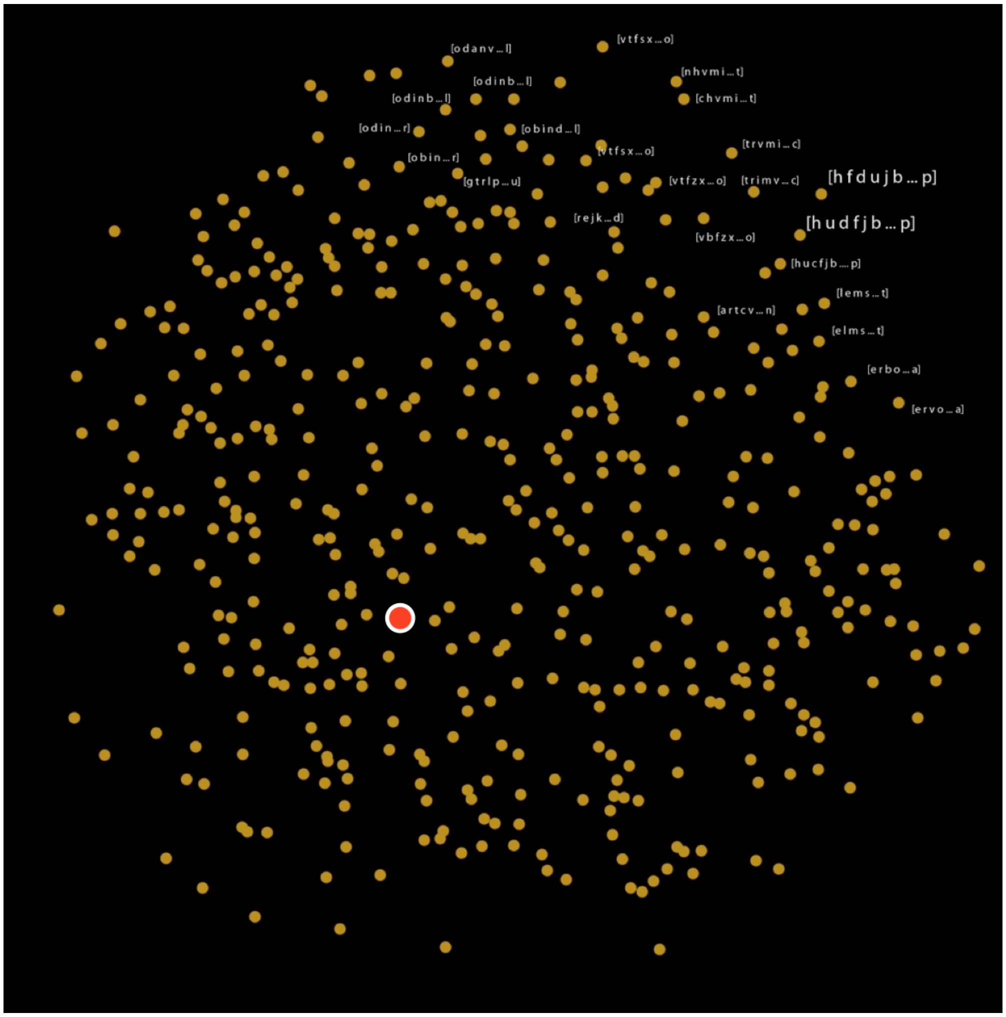


Figure 3: visualizing possible ciphers as points in space; there is a single correct one we are looking for, shown in red

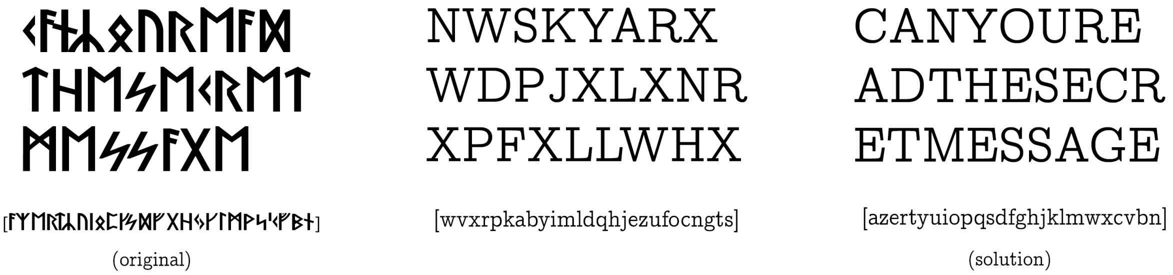


Figure 4: trying to decode a text with runic characters (left) using two different ciphers; the first guess (middle) doesn't look like English, but the second (right) does

contains many such pairs. More mathematically, here is a way to assign such a "plausibility grade":

$$\text{Plausibility}(C) = \prod_{\text{letter pairs}} (\text{port}(\text{pair})^{\text{code}_C(\text{pair})}).$$

In words: call C a given configuration of letters, say, [mlpnkobjivhucgyxftzdrseawq]. Now use that configuration to transform the artist's codes into a text with our usual characters, so, after fixing an arbitrary order for her symbols, the first one becomes an 'm', the second an 'l', etc. To assign a grade to C, go through every pair of letters in the alphabet, aa, ab, ac, . . ., zz, and count how many times we see that pair in a usual Portuguese text[4], and exponentiate that by the number of times we see that pair in the decrypted text; finally, multiply together the resulting number for each pair. This way, configurations with high grades are the ones that make pairs of letters that show up frequently in Portuguese appear often in the decrypted text. To summarize, the above formula lets the computer assign a grade to each configuration, with a higher grade if that configuration returns a text that resembles Portuguese.

text	*cipher*	*grade*
		-
YPOJSXBSSTXSJEUMGIXPEXMRYPKCKXPKVCTMXMR	[dzhgksmtxryilfnauwqjpbvceo]	246
OPYJSXBSSTXSJEUMGIXPEXMROPKCKXPKVCTMXMR	[dzhgksmtxroilfnauwqjpbvcey]	278
...	...	...
WHYNSTGSSUTSNELMIBTHETMOWHARATHAFRUMTMO	[xkpiasmutowbqcjdlzvnhgfrey]	943387
WHYNSTGSSUTSNALMIBTHATMOWHERETHEFRUMTMO	[xkpiesmutowbqcjdlzvnhgfray]	951103
WHYNSTGSSUTSNALIMBTHATIOWHERETHEFRUITIO	[xkpmesiutowbqcjdlzvnhgfray]	988529
WHYNOTGOOUTONALIMBTHATISWHERETHEFRUITIS	[xkpmeoiutswbqcjdlzvnhgfray]	1050285

Figure 5: for each cipher, we can assign a grade; the higher the more correct the decrypted text looks

Now that the computer can judge whether a particular configuration of letters is likely to be correct, we just need to navigate through them until it finds the right one. As we saw before, trying ciphers at random would take forever. How can we make the computer navigate these possibilities in a smarter way, using our grades? Here is an idea: to define a notion of 'neighbor' cipher, start with any given configuration, let's say [abcdefghijklmnopqrstuvwxyz]. Consider the combinations that could result from switching any two letters, so, switching 'a' and 'b' in the configuration gives [bacdefghijklmnopqrstuvwxyz], switching 'a' and 'c' gives [cbadefghijklmnopqrstuvwxyz], all the way to switching 'y' and 'z' to

[4] For what a "usual Portuguese text" looks like, we counted the frequencies of pairs of letters found in Machado de Assis' *Dom Casmurro*, which is a representative text in Portuguese as much as Melville's *Moby Dick* in English.

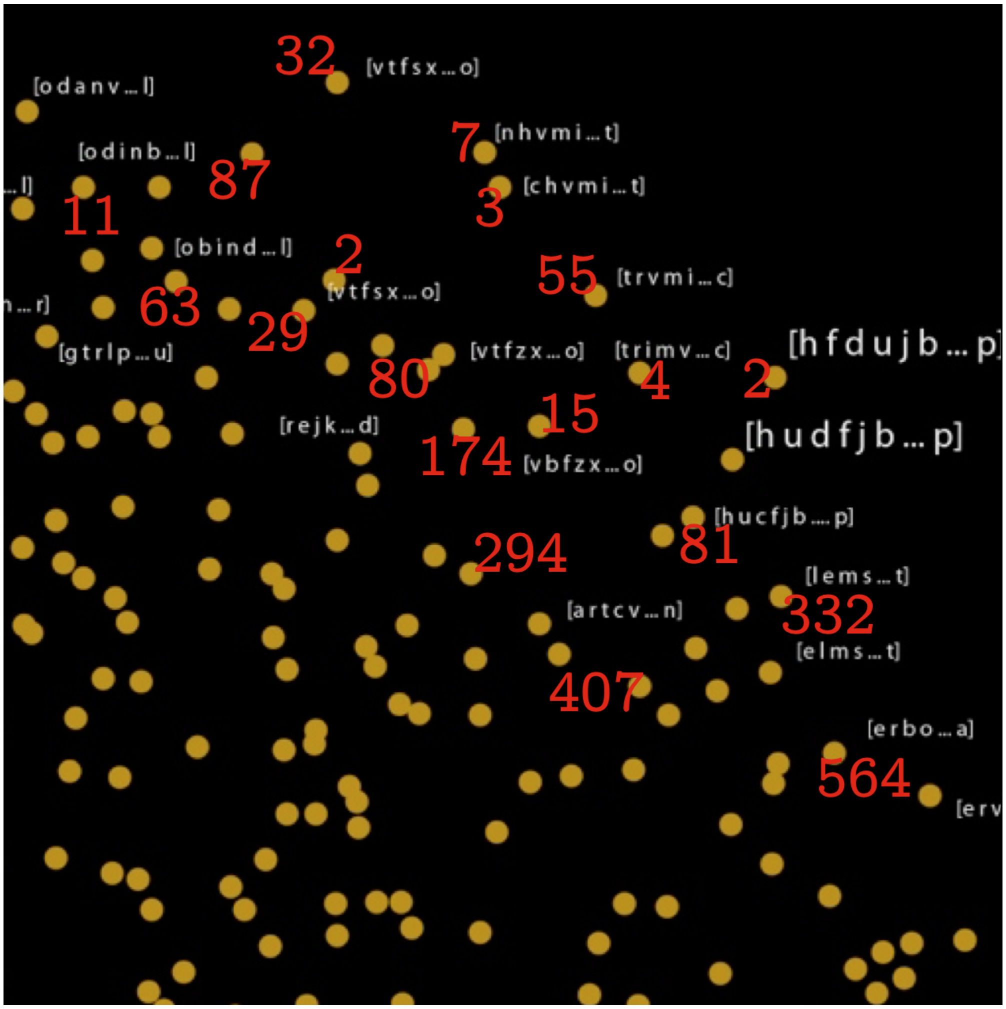


Figure 6: pictorially, each dot (a cipher) now has a grade attached to it

get [abcdefghijklmnopqrstuvwxzy]. There are 325 such combinations. Suppose the computer adds a line from one point to the other if they are connected this way (see Figure 7). This means the computer can navigate through the points by getting to a configuration, looking at the 325 neighbors, and then picking one at random and trying it. The advantage is that now, at every point, we only need to consider moving to its neighbors, not all possible ciphers. By itself, however, this idea doesn't add much: we're still just navigating randomly.

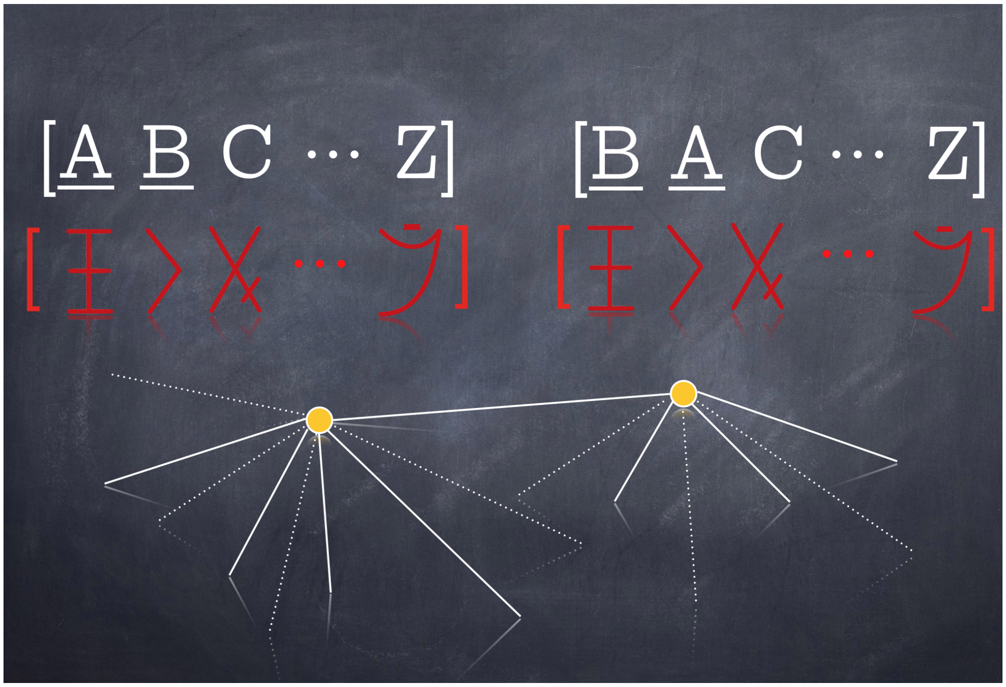


Figure 7: joining points (that is, ciphers) that can be reached by swapping a pair of letters

But here's a better way to pick which neighbor to follow using our plausibility grades: select one of the 325 at random, and if the grade is higher than the point where we're currently at, go to that configuration (that is, switch the letters in the cipher). If it doesn't, pick another neighbor. Keep doing that for a long time, until there are no neighbors with higher plausibility grade; then stop and use that final configuration as your solution. Note that, by design, whenever we switch to a different configuration, the plausibility grade increases, so hopefully after a long time we find a configuration with a very good plausibility grade. This seems like a straightforward idea, but note we now only need to search through a much smaller number of configurations, since we never follow the ones with low grade. That is enough to reduce the computation time from trillions of years to mere seconds.

Finally, we add one extra ingredient to our algorithm. If we always go to a configuration that improves

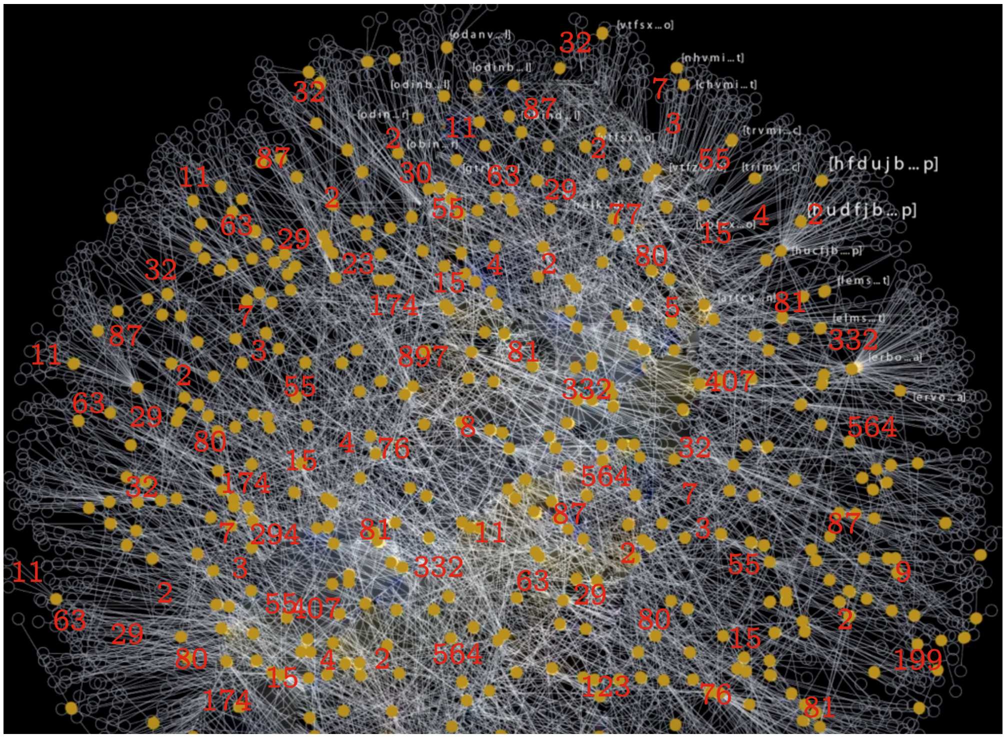


Figure 8: we assign a grade and a set of 'neighbors' to each cipher; the computer can walk around looking for the cipher with highest grade

the grade among the neighbors, we might end up trapped in a configuration that is strong relative to its neighbors, but still not quite Portuguese — just because there are no better neighbors doesn't mean it is right. Hence, instead of never following configurations with lower grades, every once in a while we allow the algorithm to pick a configuration whose grade is worse. Remarkably, adding just the right amount randomness to this otherwise deterministic process usually helps.

Now, we are done with the algorithm! We can pick a configuration at random, and let the computer keep switching letters in that configuration to generate new ones. If we randomly pick a neighbor to the current configuration that has a higher grade, change to that configuration; if we randomly pick a configuration with lower grade, then with high probability pick another neighbor, but with low probability change to that configuration. Run this for a minute or two, and then print the pixo from the artist using the cipher found.

Although the algorithm is finished, in reality there is much more to the problem than that. From a mathematical standpoint, deciding just how often we go to the neighbors with lower grades is far from a trivial matter, and requires intricate calculations[5]. There were other, non-mathematical challenges, as well: for example, Joana, the artist, wasn't using 26 characters, there were 32. Some could be punctuation marks, or accents, or they could just be bogus symbols intended to confuse anyone that tries to read it. We also didn't know if she was actually writing in Portuguese, or English, or any other language. Perhaps she was writing from right to left? We didn't have much information to go with, but we did make some adjustments to the algorithm to account for these possibilities. Furthermore, we needed data! For everything to work, the algorithm must first translate her ciphered texts, so we had to go through the streets of Rio de Janeiro photographing and copying down what she wrote, until we had almost 2000 characters. Then, finally, we were ready.

After running the algorithm for a minute, we got back some meaningful sentences. Translated to English, they would look like:

NIMNAPERSONVROMTHISCRAZYCITY

or

VAMILYOVVILTHYPIGGS.

It was clear there was some work to do. For example, we were getting Vs in place of Fs. Also, some letters seemed duplicated, while others lacked duplication. But these were fixes we could solve with some manual modifications to our solution cipher, and soon we could read almost all of her secrets.

[5] The algorithm we used is called the *Metropolis-Hastings algorithm*, which tells, in particular, how small these probabilities should be.

From then on, every time we walked about town, we would discover something new about her: some first love memories forgotten on a bridge, some medical issues left on a bus stop. Each corner of the city gave us a new perspective on this unknown, but now intimate, person.

While the puzzle was solved, the story was far from over. Over the following months, and after some hesitation on both sides, Joana and I communicated, and finally we agreed to meet. At first, we were both apprehensive: when I showed her the broken cipher, she reacted with a snarl — “bastard!”. However, the fact that we intended to keep her cipher a secret created an immediate bond between us. Very soon, Joana and I became good friends. In fact, she once told me she had considered erasing everything she had written when she learned someone could read it, but after our meeting she realized there was no need. It was still an intimate and intentional mess, but now shared with three more people.

Indeed, a recurring theme in Joana’s work is the idea of concealing. There was more to it than just using a made-up alphabet: for example, she would often paint over her works, sometimes dozens of times, so her thoughts became literally buried. Fortunately, after having her secrets read by us, she became more willing to expose her paintings in art galleries. Today, she is a well known artist in Rio de Janeiro. While her work can still be seen in Rio’s urban landscape, she has lately been hugely successful painting on canvases.

In the following months after we met, Joana would often take me in many of her forays to paint over the city’s walls. She had her own alphabet, and I had one too: mathematics. I didn’t invent mine, however: many amazing people had helped craft that language, and it indeed stands out as one of humanity’s greatest accomplishments. We painted together in many places, from architectural exhibitions to favela rooftops.

I’m now pursuing a PhD at Stanford, advised by Persi Diaconis — the same professor who wrote the original paper with the idea we used to make meaning out of Joana’s codes. He’s fond of saying one can find mathematics everywhere: from the stars above, to the rivers below, from arcane magic tricks to the mundane bubbles in a coffee mug. Now, if you pay some attention, you can also find it spattered across the walls in Rio de Janeiro.