ELECTRONIC FRONTIER FOUNDATION

JANUARY 27TH, 2010

A Primer on Information Theory and Privacy

Technical Analysis by Peter Eckersley

If we ask whether a fact about a person *identifies* that person, it turns out that the answer isn't simply yes or no. If all I know about a person is their ZIP code, I don't know who they are. If all I know is their date of birth, I don't know who they are. If all I know is their gender, I don't know who they are. But it turns out that if I know these three things about a person, I could probably deduce their identity! Each of the facts is partially identifying.

There is a mathematical quantity which allows us to measure how close a fact comes to revealing somebody's identity uniquely. That quantity is called *entropy*, and it's often measured in bits. Intuitively you can think of entropy being generalization of the number of different possibilities there are for a random variable: if there are two possibilities, there is 1 bit of entropy; if there are four possibilities, there are 2 bits of entropy, etc. Adding one more bit of entropy doubles the number of possibilities.[1]

Because there are around 7 billion humans on the planet, the identity of a random, unknown person contains just under 33 bits of entropy (two to the power of 33 is 8 billion). When we learn a new fact about a person, that fact reduces the entropy of their identity by a certain amount. There is a formula to say how much:

$\Delta S = - \log_2 \Pr(X=x)$

Where ΔS is the reduction in entropy, measured in bits,[2] and $\Pr(X=x)$ is simply the probability that the fact would be true of a random person. Let's apply the formula to a few facts, just for fun:

Starsign: $\Delta S = - \log_2 \Pr(\text{STARSIGN=capricorn}) = - \log_2 (1/12) = 3.58$ bits of information

Birthday: $\Delta S = - \log_2 \Pr(\text{DOB=2nd of January}) = -\log_2 (1/365) = 8.51$ bits of information

Note that if you combine several facts together, you might not learn anything new; for instance, telling me someone's starsign doesn't tell me anything new if I already knew their birthday.[3]

In the examples above, each starsign and birthday was assumed to be equally likely.[4] The calculation can also be applied to facts which have non-uniform likelihoods. For instance, the likelihood that an unknown person's ZIP code is 90210 (Beverley Hills, California) is different to the likelihood that their ZIP code would be 40203 (part of Louisville, Kentucky). As of 2007, there were 21,733 people living in the 90210 area, only 452 in 40203, and around 6.625 billion on the planet.

Knowing my ZIP code is 90210: $\Delta S = - \log_2 (21{,}733/6{,}625{,}000{,}000) = 18.21$ bits

Knowing my ZIP code is 40203: $\Delta S = - \log_2 (452/6{,}625{,}000{,}000) = 23.81$ bits

Knowing that I live in Moscow: $\Delta S = -\log_2 (10524400/6{,}625{,}000{,}000) = 9.30$ bits

How much entropy is needed to identify someone?

As of 2007, identifying someone from the entire population of the planet required:

$S = \log_2 (1/6625000000) = 32.6$ bits of information.

Conservatively, we can round that up to 33 bits.

So for instance, if we know someone's birthday, and we know their ZIP code is 40203, we have 8.51 + 23.81 = 32.32 bits; that's almost, but perhaps not quite, enough to know who they are: there might be a couple of people who share those characteristics. Add in their gender, that's 33.32 bits, and we can probably say exactly who the person is.[5]

An Application To Web Browsers

Now, how would this paradigm apply to web browsers? It turns out that, in addition to the commonly discussed "identifying" characteristics of web browsers, like IP addresses and tracking cookies, there are more subtle differences between browsers that can be used to tell them apart.

One significant example is the User-Agent string, which contains the name, operating system and precise version number of the browser, and which is sent every web server you visit. A typical User Agent string looks something like this:

```
Mozilla/5.0 (Windows; U; Windows NT 5.1; en-GB; rv:1.8.1.6) Gecko/20070725 Firefox/2.0.0.6
```

As you can see, there's quite a lot of "stuff" in there. It turns out that that "stuff" is quite useful for telling different people apart on the net. In another post, we report that on average, User Agent strings contain about 10.5 bits of identifying information, meaning that if you pick a random person's browser, only one in 1,500 other Internet users will share their User Agent string.

EFF's Panopticlick project is a privacy research effort to measure how much identifying information is being conveyed by other browser characteristics. Visit Panopticlick to see how identifying your browser is, and to help us in our research.

[1] Entropy is actually a generalization of counting the number of possibilities, to account for the fact that some of the possibilities are more likely than

others. You can find a pretty version of the formula here.

2. This quantity is called the "self-information" or "surprisal" of the observation, because it is a measure of how "surprising" or unexpected the new piece of information is. It is really measured with respect to the random variable that is being observed (perhaps, a person's age or where they live), and a new, reduced, entropy for their identity can be calculated in the light of this observation.

3. What happens when facts are combined depends on whether the facts are *independent*. For instance, if you know someone's birthday and gender, you have 8.51 + 1 = 9.51 bits of information about their identity because the probability distributions of birthday and gender are independent. But the same isn't true for birthdays and starsigns. If I know someone's birthday, then I already know their starsign, and being told their starsign doesn't increase my information at all. We want to calculate the change in conditional entropy of the person's identity on all the observed variables, and we can do that by making the probabilities for new facts conditional on all the facts we already know. Hence we see $\Delta S = -\log_2$ Probability(Gender=Female|DOB=2nd of January) = $-\log_2(1/2) = 1$, and $\Delta S = -\log_2$ Probability(Starsign=Capricorn|DOB=2nd of January)=$-\log_2(1) = 0$. In between cases are also possible: if I knew that someone was born in December, and then I learn that they are a Capricorn, I still gain some new bits of information, but not as much as I would have if I hadn't known their month of birth: $\Delta S = -\log_2$ Probability(Starsign=Capricorn|month of birth=December)=$-\log_2$ (10/31) = 1.63 bits.

4. Actually, in the birthday example, we should have accounted for the possibility that someone was born on the 29th of February during a leap year, in which case $\Delta S = -\log_2$ Pr(1/365.25)

5. If you're paying close attention, you might have said, "Hey, that doesn't sound right; sometimes there will be only one person in ZIP code 40203 who has a given birthday, in which case you don't need gender to identify them, and it's possible (but unlikely) that ten people in 40203 were all born on the 2nd of January. The correct way to formalize these issues would be to use the *real* fequency distribution of birthdays in the 40203 ZIP code.

Related Issues: *Online Behavioral Tracking*, *Privacy*

[Permalink: http://www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy]