What's in a Name?

Maria Gonzalez, Juan Khan, and IT Security

Many security systems still rely on personal questions as a backup means of human authentication. A study of human naming patterns finds this is highly insecure and reveals fascinating trends in how people choose names.

I spent August, in Cambridge, poring over census data and lesser-known papers in information theory. My background is in computer science and math, but I have gotten used to searching out obscure bits of knowledge to see how real-world systems fail. While many Gates scholars are hyper-specialised, I've been cramming in economics, math, electronics, and sociology—and my supervisor has just handed me a thick psychology textbook. That's life as a PhD student in security engineering.

When Passwords Fail

Our field still struggles to authenticate humans. How does Amazon.com really know it's me authorising that £100 charge? Passwords remain the standard, but suffer many problems. Many of us write them down on post-it notes, enter them into the wrong places, and re-use them between high-security accounts and throw-away ones. The biggest challenge is that we forget them with alarming regularity. This is why most systems deploy automated backup authentication, usually by asking a personal question such as the archetypical "What's your mother's maiden name?" Researchers have already proved this easy to find using public records, and the growth of social networks has made it even easier.

As a result, many sites have turned to questions that yield answers that are easy to remember, but more difficult to find (e.g. "What was your kindergarten teacher's first name?"). Almost all of these questions solicit the proper name of a person, place, or animal. Empirically, this appears to be the only type of knowledge that is both memorable and easy to query; yet, we have no good measure of its security. How hard is it to guess a name?

Modeling the Problem

Answering this basic question required developing a formal security model for personal knowledge questions. Security people are good at this. We define whom we're defending against, what their capabilities are, and what we can do to stop them. In our case, we're worried about a miscreant guessing likely names at the password reset interface of many users' accounts until he breaks one. In July, this was exactly the method used by a French hacker who gained access to a corporate email account at Twitter, leaking all of the company's internal documents to the web.

Such an attacker does not care who your mother is, only that her maiden name has a few very likely values. The attacker will probably guess "Smith," "Jones," and "Johnson," and then move on. If we know the distribution of names in the population, using results from "guessing theory" developed in the 1990s by information theorists, we can predict how secure these questions will be. After six dense pages of equations, I produced the formal security model necessary to reason about this attacker. Then it was time for the fun part: plugging in some real-world data values.

Names around the World

My colleagues and I sought out census data of all sorts: from lists of Norwegian surnames to the names of all dogs registered in Los Angeles. We searched through Facebook to get a corpus of 66 million first name/last name pairs. Data like that is hard to come by for privacy reasons, but was a goldmine for analysis.

Our data confirmed some common-sense patterns with hard numbers. Surnames are generally about twice as hard to guess as forenames. Naming patterns vary greatly around the world. The United States, with its diverse mix of cultures, has the greatest variety of names. South Korea, insular and homogeneous, has the lowest (nearly half of the population are named "Kim," "Lee," or "Park"). American surnames are 10 to 1,000 times more difficult to guess in different attack scenarios.

We also found some curious results. In most societies we studied, female names are twice as hard to guess as male names. Perhaps parents devote more time searching for a beautiful name for a baby girl than for a baby boy— although dog names are still a bit harder to guess! Looking at baby registration data from the US Social Security administration over the last 60 years, we found that human forenames are steadily becoming more diverse. Previously uncommon names like "Madison" and "Aiden" are becoming more popular than "Joseph" and "Mary."

Maria and Juan

The most interesting data analysis was correlating first and last names. For security analysis, it's important to realize that they are not picked independently from each other. "Maria Gonzalez" is the most common name on Facebook, despite "Maria" and "Gonzalez" not being close to the most popular names independently. Those would be "David" and "Smith," although there are 50% more people named "Maria Gonzalez" than there are named "David Smith." Why is that? Well, it's partly a matter of culture. For similar reasons, "Juan Khan" is a highly uncommon name, and is conspicuously absent from the data set. Despite the two names both being popular, they are a cultural mismatch.

When designing security systems, we must fully expect an attacker to exploit these statistical anomalies. However, one may not need to. Across the board, human-generated proper names provide terrible security, far lower than commonly thought. There are just too many women named "Maria Gonzalez" and too many dogs named "Lucky." As a result of this research, the demise of personal knowledge questions may be accelerated. The security community is already hard at work on practical replacements.

JOSEPH BONNEAU
Class of '08, PhD candidate in Computer Science