February 22, 2016

Carnegie Mellon, Stanford Researchers Devise Method To Share Password Data Safely

Yahoo! Releases Password Statistics of 70 Million Users For Cybersecurity Studies

By Byron Spice / 412-268-9068 / bspice@cs.cmu.edu

An unfortunate reality for cybersecurity researchers is that real-world data for their research too often comes via a security breach. Now computer scientists have devised a way to let organizations share statistics about their users’ passwords without putting those same customers at risk of being hacked.

The work at Carnegie Mellon University and Stanford University, part of an emerging field on rigorous human authentication, persuaded Yahoo! to publicly share password frequency statistics for about 70 million of its users.

“This is the first time a major company has released frequency information on user passwords,” said Anupam Datta, associate professor of computer science and electrical and computer engineering at CMU. “It’s the kind of information that legitimate researchers can use to assess the impact of a security breach and to make informed decisions about password defenses. This is extremely valuable, so we hope other organizations will follow Yahoo’s lead.”

The researchers are presenting their method on Wednesday at the Network and Distributed System Security Symposium in San Diego. Their method distorts numbers in the dataset so the list is “differentially private,” a precise mathematical definition that guarantees the released statistics don’t reveal whether any specific individual’s password is included in the dataset.

The information at issue isn’t actual passwords or user IDs, but password frequency lists — the number of times passwords are selected by a group of users. In a simplified case involving 10 users, if eight users select “123456” as a users, if eight users select “123456” as a password, and two users select “abc123,” the frequency list would be (8,2).

Password frequency lists for large user groups can be analyzed to help organizations set authentication policies that balance security with usability, or to predict which user accounts are most vulnerable, said Jeremiah Blocki, a post-doctoral researcher at Microsoft Research who began this study while a post-doc at Carnegie Mellon.

But getting access to frequency lists is difficult because of the potential for misuse. Alone, frequency lists don’t help hackers identify individual passwords, Blocki said, but they could potentially provide important clues if cross-referenced to other databases. For instance, in the earlier example, if an adversary knew the passwords for nine of 10 users, it would be child’s play to figure out the 10th password knowing that the frequency was (8,2).

Most companies are reluctant to provide access to their frequency lists, so researchers make do with data that has been inadvertently released, such as the 32 million user accounts of the defunct RockYou social app site, which suffered a data breach in 2009.

Several years ago, Joseph Bonneau, a Stanford post-doctoral researcher and a technology fellow with the Electronic Frontier Foundation, obtained samples of password frequency from Yahoo. He was able to publish some aggregate statistics, but Yahoo wouldn’t let him publicly share the raw data because of potential privacy concerns.

“Here was this data that was incredibly useful to people like me, but we couldn’t get access to it.”

people like me, but we couldn't get access to it," Blocki said.

So Blocki, Datta and Bonneau created a new algorithm to add just enough distortion to the frequency lists to make them useless to hackers, but still enable researchers to see the high-level patterns they seek in the data.

Their algorithm is based on a powerful differentially private tool called the exponential mechanism, which introduces minimal distortion but is not computationally efficient in general. By exploiting the inherent mathematical structure of a password frequency list, the researchers were able to develop a computationally efficient version of the exponential mechanism tailored to the lists.

“With our new approach, we can provide precise guarantees about privacy,” Bonneau said. “I hope this convinces more organizations to share data publicly about passwords and potentially other data that might be useful for security.”

Blocki said getting additional organizations to release password frequency lists would enable researchers to explore the impact of differing password policies. The method also might be extended to social networks — enabling the study of degree distribution lists that track the number of friends users have — and to more complicated data structures.

This research was supported by the National Science Foundation, the Air Force Office of Scientific Research, the Simons Institute for the Theory of Computing and the Open Technology Fund.

The Piper: Campus & Community News Official Events Calendar

Carnegie Mellon
University
5000 Forbes Avenue
Pittsburgh, PA 15213
412-268-2900

Legal Info |
www.cmu.edu