Leakage of private information from popular websites is common, new study finds

June 2 2011

A study of more than 100 popular websites used by tens of millions of people has found that three quarters directly leak either private information or users' unique identifiers to third-party tracking sites. The study, co-authored by Craig Wills, professor of computer science at Worcester Polytechnic Institute (WPI), also demonstrated how the leakage of private information by many sites, including email addresses, physical addresses, and even the configuration of a user's web browser -- so-called browser fingerprints -- could permit tracking sites to link many disparate pieces of information, including browsing histories contained in tracking cookies and the contents of searches on health and travel sites, to create detailed profiles of individuals.

The study, presented last week at the Web 2.0 Security and Privacy conference in Oakland, Calif., concluded that efforts made to date to curb the leakage of personal information from websites and online social networking sites, including proposals made in a 2010 Federal Trade Commission (FTC) report on protecting consumer privacy, would be largely ineffective in preventing the identified leakage and linkage. They asserted that websites need to take greater responsibility for privacy protection.

"Despite a number of proposals and reports put forward by researchers, government agencies, and privacy advocates, the problem of privacy has worsened significantly," Wills said. "With the growing disconnect between the existing and proposed privacy protection measures and the increasing and increasingly worrisome linkage of personal information

from all sorts of websites, we believe it is time to move beyond what is clearly a losing battle with third-party aggregators and examine what roles first-party sites can play in protecting the privacy of their users."

The researchers, who had previously brought attention to the leakage of personal information from many popular social networking sites (preview.tinyurl.com/4y583ru), decided to explore the handling of private information by conventional websites, an area that has gone largely unexamined, Wills said. They focused on sites that encourage users to register, since users often share personal and personally identifiable information, including their names, physical address, and email address, during the registration process. They also examined popular health and travel sites, since users conduct searches on these sites that can point to their health issues or reveal their travel plans.

They found that information is leaked through a number of routes to third-party sites that track users' browsing behavior for advertisers. In some cases, information was passed deliberately to the third-party sites. In others it was included, either deliberately or inadvertently, as part of routine information exchanges with these sites. Depending on the site, the leakage occurred as users were creating, viewing, editing, or logging into their accounts, or while navigating the websites. They also observed sensitive search terms (such as pancreatic cancer) being leaked by health sites and travel itineraries being leaked by travel sites.

The researchers examined the types of information being leaked by the websites and rated them according to their sensitivity and their ability to identify users. A user's name, phone number, or email address rated highest on the identifiability scale, for example, while health information and travel itineraries rated highest on the sensitivity scale. While the majority of leaked information rated low on both scales, the authors said this does not necessarily suggest that users need not be concerned about privacy leaks from websites.

They noted that third-party tracking sites receive a wide range of information from popular websites that could be used to connect diverse bits of leaked information and connect them to an individual user's identity. These include the user ID that a website assigns to a user (leaked by nearly half of the sites studied), unique identifiers like email addresses or home addresses, and browser fingerprints—information on how an individual browser is configured, including the list of installed plugins, which the authors found is leaked by a number of sites.

The study also evaluated a range of actions that web users could take to prevent their information from being leaked, including blocking the setting of cookies and using an advertising blocking utility or the blocking features built into the newest versions of some popular browsers. They found that all of these techniques miss some types of leakage; ad blockers, for example, do not reliably block leakage to so-called hidden third-party sites and also impair the usability of some websites.

They also reviewed proposals included in a December 2010 report on online privacy release by the FTC. "The report advocates the Privacy by Design initiative, which seeks proactive embedding of privacy at the design stage, defaults to be set to private, transparence about users' information, and access to all user-related sensitive data stored in aggregators," the study notes. But even these proposals fail to provide safeguards against the linkage of user information by third-party sites or leakage to hidden third parties, and they do not include methods for either verifying that third-party sites abide by the guidelines or penalizing those that do not.

"A key failure of the FTC report is that it largely ignores the responsibility of websites in safeguarding the privacy of their users," Wills said. "These sites should play a custodial role in protecting their users and preventing the leakage of their sensitive or identifiable

information. Third-party sites have a powerful economic incentive to continue to collect and aggregate user information, so relying on them to protect user privacy will continue to be a losing battle. It is time to put the focus on what first-party sites can and should do."

Provided by Worcester Polytechnic Institute