

Computer scientists study data security in Internet website tracking

June 17 2020



When browsing on the Internet, companies collect data not only about accessed websites, but also about the time of access or location information. (Photo: Amadeus Bramsiepe, Markus Breig, KIT) Credit: Amadeus Bramsiepe, Markus Breig, KIT

Tracking our browsing behavior is part of routine Internet use.



Companies use it to adapt ads to the personal needs of potential clients or to measure their range. Many providers of tracking services advertise secure data protection by generalizing datasets and anonymizing data in this way. Computer scientists of Karlsruhe Institute of Technology (KIT) and Technische Universität Dresden (TUD) have now studied how secure this method is and reported their findings in a scientific paper for the IEEE Security and Privacy Conference.

Tracking services collect large amounts of Internet user <u>data</u>. These data include the websites accessed, but also <u>information</u> on the end devices used, the time of access (timestamp) or location information. "As these data are highly sensitive and have a high personal reference, many companies use generalization to apparently anonymize them and to bypass data security regulations," says Professor Thorsten Strufe, Head of the 'Practical IT Security' Research Group of KIT.

By using generalization, the level of information detailing is reduced, such that identifying individuals is supposed to be impossible. For example, <u>location information</u> is restricted to the region, the time of access is limited to the day, or the IP address is shortened by some figures. Strufe, together with his team and colleagues of TUD, have now studied whether this method really allows no conclusions to be drawn with respect to the individual.

With the help of a large volume of metadata of German websites with 66 million users and over 2 billion page views, the computer scientists succeeded in not only drawing conclusions with for websites accessed, but also to the chains of page views, the so-called 'click traces.' The data were made available by INFOnline GmbH, an institution measuring the data range in Germany.

The Course of Page Views Is of High Importance



"To test the effectiveness of generalization, we analyzed two application scenarios," Strufe says. "First, we checked all click traces for uniqueness. If a click trace, that is the course of several successive page views, can be distinguished clearly from others, it is no longer anonymous." They found that information on the website accessed and the browser used has to be removed completely from the data to prevent conclusions to be drawn with respect to persons. "The data will only become anonymous when the sequences of single clicks are shortened, which means that they are stored without any context, or when all information, except for the timestamp, is removed," Strufe says. "Even if the domain, the allocation to a subject, such as politics or sports, and the time are stored on a daily basis only, 35 to 40 percent of the data can be assigned to individuals." For this scenario, the researchers found that generalization does not correspond to the definition of anonymity.

A Few Observations Are Sufficient to Identify User Profiles

In addition, the researchers checked whether even subsets of a click trace allow conclusions to be drawn about individuals. "We linked the generalized information from the database to other observations, such as links shared on social media or in chats. If, for example, the time is generalized precisely to the minute, one observation is sufficient to clearly assign 20 percent of the click traces to a person," says Clemens Deusser, doctoral researcher of Strufe's team, who was largely involved in the study. "Another two observations increase the success to more than 50 percent. Then, it is easily obvious from the database which other websites were accessed by the person and which contents were viewed." Even if the timestamp is stored with the precision of a day, only five additional observations are needed to identify the person.

"Our results suggest that simple generalization is not suited for effectively anonymizing web tracking data. The data remain sharp to the person and anonymization is ineffective. To reach effective data



protection, methods extending far beyond have to be applied, such as noise by the random insertion of minor misobservations into the data," Strufe recommends.

Provided by Karlsruhe Institute of Technology

Citation: Computer scientists study data security in Internet website tracking (2020, June 17) retrieved 26 April 2024 from https://techxplore.com/news/2020-06-scientists-internet-website-tracking.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.