Nota de prensa

Madrid, Spain 14 October, 2020

Over 150 million websites among a billion tested include sensitive (and tracked) content

The European General Data Protection Regulation (GDPR) includes specific clauses that put restrictions on the collection and processing of sensitive personal data, defined as any data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, also genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation…

After two years of hard work, and having crunched more than one billion web-sites (most of the English speaking web), an international team, with Nikolaos Laoutaris (Research Professor at IMDEA Networks Institute, Madrid), as well as researchers from TU Berlin and the Cyprus University of Technology, has developed specialised machine learning classifiers that are able to identify sensitive URLs on the web and used them to search for such URLs on a corpus of some 1 billion URLs in total. As a main (and disturbing) result, some 150 million of them were find to include sensitive content related to Health, Political Beliefs, Sexual Orientation, etc … and still be tracked nearly as much as the rest of the web.

A real time detection

Existing legislation about sensitive personal data is targeted mostly for use by humans, e.g., to file complaints, conduct investigations, and even pursue cases in courts of law. With the use of the new automated machine learning classifiers, however, additional proactive measures can also be put in place for the first time. For example, the browser of the user, or an add-on program, can warn him before clicking and following URLs pointing to sensitive content. Upon visiting such sites, trackers can be blocked, and complaints can be automatically filed. Being able to do the above hinges on being able to classify automatically whether a URL is a sensitive one or not, in real time.

The latter is easier said than done. The reason has to do with the ambiguity of terms such as "Health", that are used by legal documents to indicate what types of information are considered as sensitive. Indeed, the word Health can be found in both web-sites about healthy eating, sports, and organic food, but also on web-sites about chronic diseases, sexually transmitted diseases, and cancer. Most of the effort on producing the aforementioned classifier went into collecting sufficient "ground-truth" data for training the classifier and allowing it to distinguish truly sensitive uses of words such as health from less sensitive ones.

The results of the work of the team will be presented, as a scientific paper, in ACM IMC'20 (ACM Internet Measurement Conference 2020, 27-29 October, Pittsburgh, USA). Laoutaris also participates in PIMCity (Building the next generation personal data platforms), the EU-funded

project to increase transparency and provide users with control over their data. "Privacy law is made for use by humans -Laoutaris comments-, typically after a privacy breach has occurred – e.g., an illegal processing of such data- … but how can we teach this law to machines and have them protect us before privacy breaches occur?". The research team is working to bring a technological solution to the user in 2021.

"Tracking people -add the researcher- when they visit websites with content that belongs to the GDPR sensitive categories is the true 'Elephant in the Room' of privacy. Most people don´t mind be tracked about things that they consider innocent, but would be very upset to know that their visit to sensitive websites are being logged and released to unknown third parties. Our study is, by far, the biggest study about tracking of sensitive topics on the web. It shows that a good part of the web includes content of sensitive character. Unfortunately, these sensitive pages appear to be as tracked as the rest of the web".

About Nikolaos Laoutaris

Research professor at IMDEA Networks since December 2018. Laoutaris is a doctor of computational sciences from the University of Athens (Greece) and worked as a researcher at Harvard University and Boston University. His areas of research centre on privacy, transparency and data protection; the network and information economy; smart transport; distributed systems and network protocols and traffic measurements.

Source(s): IMDEA Networks Institute

URL: Over 150 million websites among a billion tested include sensitive (and tracked) content

About us

IMDEA Networks Institute is a research organization on computer and communication networks whose multinational team is engaged in cutting-edge fundamental science and technology. As a growing, English speaking institute located in Madrid, Spain, IMDEA Networks offers a unique opportunity for pioneering scientists to develop their ideas. IMDEA Networks has established itself internationally at the forefront in the **development of future network principles and technologies**. Our team of highly-reputed researchers is designing and creating today the networks of tomorrow.

Some keywords that define us: 5G, Big Data, blockchains and distributed ledgers, cloud computing, content delivery networks, data analytics, energy-efficient networks, fog and edge computing, indoor positioning, Internet of Things (IoT), machine learning, millimeter-wave communication, mobile computing, network economics, network measurements, network security, networked systems, network protocols and algorithms, network virtualization (software defined networks – SDN and network function virtualization – NFV), privacy, social networks, underwater networks, vehicular networks, wireless networks and more…

IMDEA Networks Institute
28918 Leganes (Madrid) Spain
Avda. del Mar Mediterráneo, 22

+34 91 481 6210
mediarelations.networks@imdea.org
www.networks.imdea.org

Twitter: @IMDEA_Networks | Facebook | Instagram | Flickr | YouTube