Worried About Being Watched Online? Cryptanalytica is Here to Address Your Privacy Concerns

Gargi Mitra*
Department of Computer Science & Engineering, IIT Madras
Email: gargim@cse.iitm.ac.in

The advent of easily affordable smartphones and Internet of Things (IoT) devices, coupled with Government initiatives such as the Digital India, has almost brought the whole world at our fingertips. However, while enjoying the luxuries of the digital world, we seldom spare a thought about the privacy of the vast amount of digital information we generate and transmit across the Internet every day. Perhaps we exercise utmost caution only during online payments or while sending some confidential emails. But then, the 'https' tag on the address bar assures us that our information is being *securely* transmitted across the web. But if security is already assured, why is there so much hype around online *privacy*? Doesn't security implicitly guarantee privacy? Can the vast amount of internet traffic traveling around the globe leak some of our personal details? What can be the consequences of such information falling in the wrong hands? If these questions baffle you, then IIT Madras now has a solution that lets you evaluate your online privacy. The proposed solution, called *CryptAnalytica*, lets you discover all the information that leaks when you browse through websites even while using a secure communication protocol. It further lets you understand if the leaked information can pose a threat to your online privacy and also gives an insight into how you can prevent a potential privacy breach on the Internet.

Although 'security' and 'privacy' conceptually sound very similar, there is a fine line that distinguishes the two. For those of you who are not familiar with this subtlety, here is a simple illustration that will help you understand the difference. Imagine that on a fine morning, you get a call from a courier waiting at your office entrance while you are at your desk. Upon meeting, he

* Ms. Gargi Mitra, Ph.D. Scholar from Indian Institute of Technology, Madras, is pursuing her research on "Evaluating and Alleviating Privacy Leakage on the Internet." Her popular science story entitled "Worried about Being Watched Online? Cryptanalytica is Here to Address Your Privacy Concerns" has been selected for AWSAR Award.

delivers you a gift that your friend has sent you for your birthday, parcelled up in a pretty wrapper. However, you do not want to open the parcel in front of your colleagues, so you put the parcel in your drawer and lock it up once you walk back to your desk. Now, your colleagues who have been observing all your activities can easily infer that you have received a parcel by courier. The wrapper of the parcel further indicates that it is a gift. Some of your colleagues might ask you if it is your birthday. Finally, the secrecy you maintain about it makes some people believe that the parcel might contain something expensive. As a result, although you have now successfully hidden the contents of the parcel from your colleagues, by merely observing your actions they come to know of some information about you, which you explicitly did not share with them. Some of this information may be correct and some may be wrong. Hence the inferences made by your colleagues are definitely probabilistic. However, the consequences of such information being leaked can be either pleasant or painful for you, depending on who observes what. A friendly colleague may organize a birthday party for you, while a jealous one may plan to steal the parcel. In the world of Internet communications, the act of locking up your parcel in the drawer ensures 'security', and the inferences made by your colleagues from your behavior and the appearance of the parcel constitute the 'leaked information' that may violate your 'privacy'. All the attributes of the parcel and your activities constitute the 'metadata', and the metadata that helped your colleagues draw inferences about you are called 'side-channels'. Your colleagues are the 'surveillants' here.

The concept of mass-scale digital surveillance garnered worldwide attention when thousands of classified documents belonging to the National Security Agency (NSA) of the USA were leaked by its ex-employee Edward Snowden in 2013. These documents revealed that post the 9/11 attacks, the US Government has been spending heavily on well-organized mass surveillance programs such as the PRISM, XKeyscore, and Tempora. Through these programs, the NSA collects and analyzes Internet communication traffic generated by people across the globe with the help of companies like Google and by intercepting fiber-optic cables around the world. When confronted with charges of unauthorized information access, the NSA attempted to defend themselves by stating that they only collect *metadata* about Internet communications for national interests, and do not break the security of the actual *data* being communicated. However, this sparked a debate regarding the role of metadata in breaching the online privacy of an individual, an institution or a nation.

Internet measurement studies suggest that as of August 2018, 90.4% of all Internet traffic consist of web browsing traffic, and it is also the most sought-after source of information for mass surveillants. CryptAnalytica focuses on identifying side-channels in secure web browsing traffic that may leak information about which web pages in a website are popular among the masses. Such information, when leaked, can help cyber attackers identify their sweetest target points for circulating malwares or other malicious activities. Identification of such side-channels before making a website publicly available will help a website designer devise ways to protect web browsing privacy. Existing mechanisms for evaluating privacy vulnerability of Internet communication assume targeted surveillance on specific people, where the attacker is believed to possess a lot of background information about the victims. Such information includes personal details such as preference of food, possible medical conditions, etc. Possessing such detailed knowledge about a large number of people is not practically feasible. To the best of our knowledge, CryptAnalytica is the first framework that evaluates the vulnerability of web traffic in the face of mass surveillance, assuming no prior knowledge about the targets.

CryptAnalytica operates in two phases - *profiling* and *prediction*. In the profiling phase, it first observes the metadata of Internet traffic generated when a user accesses different web pages of a website. By metadata, we refer to those attributes of Internet traffic which are visible to anyone who can intercept it. Such metadata include the volume of network traffic, the time required to transmit a file, IP address of the web server, IP address of the user, etc. From the metadata, CryptAnalytica identifies the side-channels which might reveal which web resource (image, video, etc.) has been communicated over a secure channel. As we know, web pages are composed of multiple such web resources. So, if a surveillant can infer which web resource has been accessed, he can further infer the webpage accessed. CryptAnalytica then selects those side-channels which have a steady value across different network conditions. This is important since side-channels having different values for different scenarios are not suitable for mass-scale analysis. For instance, the time required to download a video from a website is not a good side-channel for identifying which video has been downloaded, since the download time depends on the network speed, which varies from time to time. Hence, by observing the download time, a surveillant cannot infer which video has been downloaded by a user. However, it has been observed that even when communicated securely, the sizes of the web resources cannot be hidden from a surveillant. Furthermore, these sizes remain constant across various network conditions and user behaviors. So, this forms a stable side-channel. Once such stable side-channels have been identified, CryptAnalytica stores the side-channel values (in our case, the resource sizes) in a database. Thereafter, in the prediction phase, CryptAnalytica uses this information to check if the different resources can be identified uniquely from their side-channel values. Also, from the resource identified, it checks if it is possible to predict the webpage accessed. This analysis is important because in practical cases, a website often hosts different resources having similar sizes. Also, the same resource can be shared by multiple web pages. Owing to such factors, the inferences made from side-channel values are always probabilistic. We evaluated CryptAnalytica on a real website and it was found that the side-channel leakage of the website can allow surveillants to correctly predict web pages browsed by its users in 78% cases.

Apart from the cyber attack point of view, unauthorized inference of web page access statistics can have several other consequences. There has been evidence of dishonest Internet Service Providers, VPN and cloud service providers who sell information about their clients to unauthorized people. If such entities can discover stable side-channels in business websites, they may sell off the web page access statistics to data analysts who can infer further information about the business organizations, such as their consumer bases, popular products, and other confidential business-specific information. On the other hand, if used by authorized personnel, CryptAnalytica can be used to identify malicious contents such as malwares on live web traffic and that may help in controlling the proliferation of such harmful objects.

The research team at IIT Madras that has been working on CryptAnalytica includes - Gargi Mitra, Prasanna Karthik Vairam, Prof. V Kamakoti and Dr. Nitin Chandrachoodan.