Diss. ETH No. 22986
TIK-Schriftenreihe Nr. 162

On Data and Privacy Leakage in Web Traffic

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

DAVID GUGELMANN

Master of Science ETH in Electrical Engineering and
Information Technology, ETH Zurich
born on 24.09.1984
citizen of Brittnau (AG)

accepted on the recommendation of
Prof. Dr. Bernhard Plattner, examiner
Prof. Dr. Roberto Perdisci, co-examiner
Dr. Vincent Lenders, co-examiner

2015

Abstract

Web services are an inseparable part of our personal and business life. Nevertheless their widespread use and the large footprint caused by embedded third-party services bring about a massively increased risk of data and privacy leakage. Data leakage, i.e., the disclosure of proprietary data, either through careless or malicious employees, or resulting from digital attacks on information systems, primarily concerns organizations. The universal access to the Web, the complex traffic patterns and the large request volumes caused by employees' browsing make it extremely challenging to prevent and investigate data leakage in outgoing HTTP and HTTPS requests. For individuals, leakage rather takes the form of privacy loss through Web services that track a user's actions to create a detailed profile without the user being aware of it. The trend towards outsourcing Web site functionality to third-party services increases the risk of privacy loss in the Web because every embedded third party can learn which pages a user is visiting. In this thesis, we consider both variations of leakage in Web traffic, i.e., data leakage and privacy loss. Our aim is to analyze problems of existing leakage protection measures and to develop new methods for the investigation and prevention of data and privacy leakage. We develop methods to archive and analyze Web traffic for the investigation of data loss incidents, and we discuss how to identify privacy-intrusive Web services in order to prevent privacy loss during Web browsing.

In the first part of this thesis, we focus on data leakage in organizations. As employees routinely need access to external Web services, an organization cannot simply block all HTTP and HTTPS traffic at the network perimeter using traditional firewalls. Security companies promote data loss prevention (DLP) systems as a silver bullet to solve this problem. DLP solutions monitor outgoing information flow and selectively block the leakage of sensitive data. But as we show in this thesis, DLP systems cannot hold up to their promises.

We systematically analyze data leakage vectors for HTTP requests and show that three DLP solutions of major vendors cannot withstand basic data leakage by disgruntled users or malware, they can only prevent accidental data leakage. Since protection measures are insufficient, organizations need to plan for handling data leakage incidents. As a result, there is a demand for forensic procedures that allow organizations to investigate data leakage incidents in retrospect. But the time span covered by current network forensics solutions is quite limited due to the large traffic volumes transmitted nowadays. To address this issue, we develop an architecture to filter non-relevant HTTP traffic and compress the remaining outgoing HTTP request data. In contrast to DLP systems, our architecture does not aim to identify sensitive data, but data that are irrelevant for incident investigations. Our approach increases the time horizon for forensic investigations significantly. Further, as the complexity of today's Web traffic makes it time-consuming to reconstruct user actions and identify malicious activities, we present a novel visualization methodology. Our methodology allows an investigator to understand the Web activity of a client at a glance and to spot malware and data leakage activities.

In the second part of this thesis, we approach the problem of privacy loss while making use of the Web. The large footprint of today's Web sites not only complicates forensics, but also puts various third parties in the position to create exhaustive user profiles. To get an intuition to which classes of Web services users provide most information, we first conduct a study covering the HTTP traffic of around fifteen thousand IP addresses in a campus network. We introduce a heuristic to estimate the amount of information contained in HTTP requests and find that advertisement and analytics services receive by far most of the information transmitted to third parties during Web browsing. Further, we find that many advertisement and analytics services show distinct traffic properties. Based on this finding, we develop a machine learning-based approach for automatic identification of new privacy-intrusive services. Our approach can complement the blacklists employed by ad blockers and thus reduces privacy loss in the Web.

In conclusion, our work *(i)* extends the time horizon for forensic investigations compared to state-of-the-art solutions, *(ii)* speeds up investigations by our novel HTTP and HTTPS visualization, and *(iii)* better protects individuals and organizations from new privacy-intrusive Web services.

Kurzfassung

Das World Wide Web (WWW) ist nicht mehr aus unserem privaten oder geschäftlichen Leben wegzudenken. Die verbreitete Nutzung und Komplexität der im WWW angebotenen Dienste erhöhen jedoch die Risiken für Datenlecks und gefährden unsere Privatsphäre massiv. Datenlecks, d.h., die Offenlegung von vertraulichen Daten, betreffen in erster Linie Organisationen. Datenlecks können durch digitale Angriffe auf die IT-Infrastruktur einer Organisation und durch nachlässige oder böswillige Mitarbeiter verursacht werden. Durch alltägliches Websurfen werden grosse Anfragevolumen an eine Vielzahl von Webdiensten gesendet, dies macht es sehr schwierig Datenlecks in ausgehenden HTTP- und HTTPS-Anfragen zu verhindern und zu untersuchen. Die Gefährdung der Privatsphäre durch Webdienste, die das Surfverhalten von Nutzern verfolgen, betrifft hauptsächlich Einzelpersonen. Ein wesentlicher Grund für die Gefährdung der Privatsphäre sind Funktionen und Dienste, die durch die Betreiber von Webdiensten in ihre Webseiten eingebunden, aber durch Drittdienste bereitgestellt werden. Jede eingebundene Drittpartei kann grundsätzlich Informationen zum Surfverhalten der Nutzer auf der Webseite sammeln. Der Einbezug von derartigen Diensten bleibt dem Nutzer jedoch verborgen. In dieser Arbeit betrachten wir sowohl Datenlecks, als auch den Verlust von Privatsphäre, beide in Bezug auf Web-Verkehr. Es ist unser Ziel, Probleme bei der Prävention von Datenlecks zu untersuchen und neue Methoden zu entwickeln, um Datenlecks und Risiken für die Privatsphäre zu untersuchen und zu verhindern. Für die Untersuchung von Datenlecks entwickeln wir Methoden zur Archivierung und Analyse von Web-Verkehr. Weiter diskutieren wir, wie Webdienste, die eine Gefahr für die Privatsphäre darstellen, identifiziert werden können, um den Verlust von Privatsphäre im Internet zu verhindern.

Im ersten Teil dieser Dissertation fokussieren wir auf Datenlecks, die Or-

ganisationen betreffen. Da den Mitarbeitenden kaum der Zugang zu externen Webdiensten systematisch verwehrt werden kann, ist es keine Lösung, den ganzen HTTP- und HTTPS-Verkehr an dem Perimeter des eigenen Netzwerks mit herkömmlichen Firewalls zu blockieren. Sicherheitsunternehmen bewerben daher Data Loss Prevention (DLP) Systeme als ideale Alternative um dieses Problem zu lösen. DLP-Systeme überwachen ausgehende Informationsflüsse und blockieren den Abfluss von sensitiven Daten selektiv. Allerdings zeigen unsere Untersuchungen, dass DLP-Systeme die gemachten Versprechen nicht einhalten können. Unsere systematische Analyse von Datenabflussmöglichkeiten in HTTP-Anfragen zeigt, dass drei DLP-Lösungen von grossen Anbietern grundlegende Datenlecks, die durch verärgerte Mitarbeiter oder bösartige Software verursacht werden, nicht verhindern können. Unsere Experimente zeigen, dass diese Lösungen lediglich versehentliche Datenlecks verhindern können. Da die Schutzmassnahmen unzureichend sind, können Datenlecks nicht grundsätzlich verhindert werden. Somit müssen Organisationen wenigstens in der Lage sein, nach der Entdeckung eines Datenlecks den Schaden zu analysieren, um die Verantwortlichen zur Rechenschaft zu ziehen. Viele Zwischenfälle werden aber über Wochen bis Jahre nicht bemerkt, daher gibt es einen Bedarf für forensische Verfahren, die es erlauben Datenlecks im Nachhinein zu untersuchen. Die Zeitspanne, die von aktuellen Netzwerkforensiklösungen abgedeckt wird, ist jedoch aufgrund der grossen Datenmengen, die heutzutage übertragen werden, recht begrenzt. Um dieses Problem anzugehen, entwickeln wir eine Architektur, die im Gegensatz zu DLP-Systemen nicht versucht sensitive Daten zu identifizieren, sondern Daten die nicht sensitiv und somit irrelevant sind. Durch das Filtern von irrelevanten HTTP-Daten und einer effizienten Speicherung der verbleibenden ausgehenden Daten, kann der Zeithorizont für forensische Untersuchungen deutlich verlängert werden. Da es durch die Komplexität des heutigen Web-Verkehrs zeitaufwendig ist, Benutzeraktivitäten zu rekonstruieren und bösartige Aktivitäten zu identifizieren, präsentieren wir zudem eine neuartige Visualisierungsmethodik. Unsere Methodik erlaubt es einem Ermittler, die Web-Aktivitäten eines Gerätes auf einen Blick zu verstehen und die Aktivitäten von bösartiger Software und Datenlecks visuell zu identifizieren.

Im zweiten Teil dieser Arbeit betrachten wir Mechanismen, die zum Verlust von Privatsphäre beim Websurfen führen können. Die grosse Anzahl an Diensten, welche beim Websurfen involvierten sind, erschwert nicht nur die forensische Analyse, sondern ermöglicht es auch verschiedensten Drittparteien umfassende Benutzerprofile zu erstellen. Wir führen zunächst eine Studie über den HTTP-Datenverkehr von rund fünfzehntausend IP-Adressen in

einem Universitätsnetzwerk durch, um zuverlässige Statistiken über die Arten von Webdiensten zu erhalten, an welche die Benutzer am meisten Daten übermitteln. Wir entwickeln eine Heuristik um abzuschätzen, wie viel Information in HTTP-Anfragen enthalten ist. Dabei stellen wir fest, dass Tracking- und Werbedienste während dem Surfen mit Abstand die meisten der an Dritte übertragenen Informationen erhalten. Zudem finden wir, dass der Verkehr spezielle statistische Eigenschaften aufweist, welche die Tracking- und Werbedienste erkennbar machen. Basierend auf dieser Erkenntnis entwickeln wir einen Ansatz für die automatische Identifizierung von Webdiensten, welche die Privatsphäre von Nutzern gefährden können. Unser auf maschinellem Lernen basierender Ansatz erlaubt es, die von Anti-Werbesoftware eingesetzten schwarzen Listen zu ergänzen und somit den Verlust der Privatsphäre im Web zu reduzieren.

Die in unserer Forschung entwickelten Verfahren *(i)* verlängern den Zeithorizont für die forensische Untersuchung von Zwischenfällen im Vergleich zum aktuellen Stand der Technik deutlich, *(ii)* beschleunigen die Untersuchungen von Zwischenfällen durch unsere neuartige HTTP- und HTTPS-Visualisierung und *(iii)* schützen Einzelpersonen und Organisationen besser vor neuen Webdiensten, die ein Risiko für die Privatsphäre darstellen.