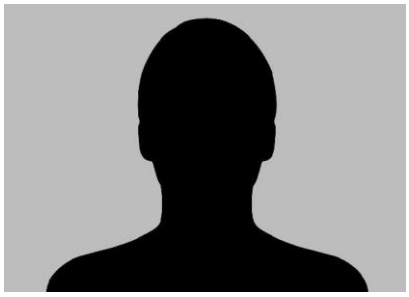
Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

InIT Institut für angewandte Informationstechnologie

Sniffing out the Bad Guys - Classifying URLs using Active Probing and 3rd-Party Data Sources



Diplomand/in
Olivier Favre

Dozent/in
Bernhard Tellenbach

Lange Zeit konzentrierten sich Anti-Malware-Bemühungen auf die Identifizierung von Malware durch statische und dynamische Analyse von Dateien, die sich im Transit, im Ruhezustand oder im Gebrauch befinden. Eine neuere Ergänzung im Arsenal der Verteidiger sind dateiinhaltsunabhängige Ansätze, die darauf abzielen, die Infrastruktur der Malware-Verbreitung zu identifizieren, um damit ausgetauschte Daten zu erkennen, zu blockieren oder, wenn möglich, zu löschen. In dieser Arbeit schlagen wir eine Reihe von inhaltsunabhängigen Indikatoren vor, welche im Idealfall beurteilen können, ob der Zugriff auf eine bestimmte Datei-URL eine Bedrohung darstellt. Ausserdem zeigen wir wie man einen ausgewogenen Datensatz mit bösartigen und gutartigen Datei-URLs erstellen kann, der ausschliesslich auf öffentlich verfügbaren Datenquellen basiert. Unsere Indikatoren konzentrieren sich auf drei öffentlich zugängliche Informationsquellen: Webcrawler, Whitelists und Web-Archivierungsdienste.

Bei Quellen, die sich auf das Webcrawler beziehen, liegt der Schwerpunkt auf Suchmaschinen, wo wir zeigen, dass die Verwendung sorgfältig konstruierter Suchanfragen Informationen über Datei-URLs und den zugrunde liegenden Fussabdruck aufdecken kann.

Bei Whitelists analysieren wir verschiedene Exemplare auf ihre Stärken und Schwächen. Anschliessend zeigen wir, dass der traditionelle Ansatz, bei dem Dateien auf der Grundlage ihrer kryptographischen Hashes verglichen werden, nicht ausreicht. Folglich schlagen wir die Verwendung von Fuzzy-Hashes als Lösung vor. In unseren Experimenten zeigen wir, dass es möglich ist, die Reichweite einer Whitelist durch die Verwendung von Fuzzy-Hashes zu erhöhen, selbst wenn die Fuzzy-Hashes vor Jahren erzeugt wurden.

Im dritten Bereich, der sich auf Archivierungsdienste bezieht, zeigen wir, wie man Momentaufnahmen einer Website nutzen kann, um einen Überblick über ihren Fussabdruck zu gewinnen.

Abschliessend schlagen wir eine Reihe von Merkmalen vor, die auf den Resultaten aufbauen und für die Verwendung mit einem binären Klassifikator geeignet sind. Unsere Evaluation mit einem Satz von bösartigen Datei-URLs aus Blacklists und gutartigen aus dem CommonCrawl-Datensatz konnte die korrekte Klasse einer Datei-URL mit einer Genauigkeit von 93% vorhersagen.

Collage von Organisationen und Services, welche als Drittpartei Quellen gebraucht wurden