A Walk on the Web's Wild Side

STUDIENARBEIT

für die Prüfung zum

Bachelor of Science

des Studiengangs Informatik
Studienrichtung Angewandte Informatik

an der

Dualen Hochschule Baden-Württemberg Karlsruhe

von

Samuel Philipp
Daniel Brown
Jan-Eric Gaidusch

28. April 2017

Bearbeitungszeitraum	6 Monate
Matrikelnummern	9207236, 3788021, 8296876
Kurs	TINF14B2
Ausbildungsfirma	Fiducia & GAD IT AG
Gutachter der Studienakademie	Dr. Martin Johns

Abstract *TODO* Daniel

Erklärung

(gemäß §5(3) der „Studien- und Prüfungsordnung DHBW Technik“ vom 29.9.2015)

Wir versichern hiermit, dass wir unsere Studienarbeit mit dem Thema:

„A walk on the web’s wild side“

selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt haben. Wir versichern zudem, dass die eingereichte elektronische Fassung mit der gedruckten Fassung übereinstimmt.

Karlsruhe, den 28. April 2017	
Ort, Datum	Samuel Philipp
Karlsruhe, den 28. April 2017	
Ort, Datum	Daniel Brown
Karlsruhe, den 28. April 2017	
Ort, Datum	Jan-Eric Gaidusch

Inhaltsverzeichnis

Abkürzungsverzeichnis

WWW World Wide Web

HTML Hypertext Markup Language

CSS Cascading Style Sheets

UI User Interface

JVM Java Virtual Machine

API Application Programming Interface

DRY Don't Repeat Yourself

REST Representational State Transfer

URI Uniform Ressource Identifier

NIDS Network Intrusion Detection System

Abbildungsverzeichnis

Tabellenverzeichnis

Listings

1 Einleitung

1.1 Einführung

TODO Samuel

1.2 Hintergrund

TODO Jani

1.3 Team

TODO Needs review Das Entwicklerteam besteht aus drei Studenten der Richtung Angewandte Informatik: Samuel Philipp, Daniel Brown und Jan-Eric Gaidusch. Der Name der Arbeitsgruppe ist *SecuritySquad* [1].

Die Studienarbeit wird von Dr. Martin Johns betreut, der an der DHBW Karlsruhe die Vorlesung Datensicherheit hält. Hauptberuflich ist er Forscher ebendieses Gebietes am CEC Karlsruhe der SAP AG[2].

[1] Der Name *SecuritySquad* ist angelehnt an den Titel des US-amerikanischen Actionfilms *Suicide Squad*.

[2] **johnsProfile**

Abbildung 1: Secutitysquad - Logo

1.4 Aufgabenstellung

Anbieter von zwielichtigen Web-Angeboten greifen ihre User mit diversen Client-seitigen Methoden an. Beispiele für solche Angriffe sind Malware Downloads, Phishing, JavaScript Intranet Angriffe, oder Browser Exploits.

Ziel der Arbeit ist eine systematische Untersuchung der Aktivitäten von semi-legalen Webseiten im World Wide Web (WWW). Das erwartete Ergebnis ist ein Prüfportal, auf dem jene Webseiten automatisiert analysiert werden und Ergebnisse präsentiert werden sollen.

Nach dem ersten Schaffen einer Übersicht von interessanten Zielen, wie z.B. One-Click-Hoster oder File-sharing Sites sollen ausgewählte Webseiten manuell untersucht werden. Außerdem sollen verschiedene Angriffsszenarien zur weiteren Prüfung ausgewählt werden. Der Untersuchungsprozes der Webseiten soll im Verlauf dieser Arbeit stückweise automatisiert und in den Rahmen einer Prüfanwendung gebracht werden.

Abschließend sollen eine Vielzahl von Webseiten mit der Anwendung getestet und die Ergebnisse ausgewertet und dokumentiert werden.

1.5 webifier

Abbildung 2: webifier - Logo

webifier ist eine Anwendung, mit der Webseiten auf deren Seriosität und mögliche clientseitige Angriffe auf den Nutzer geprüft werden können. Sie besteht aus mehreren eigenständigen Teilanwendungen. Im Zentrum steht der Tester, welcher die einzelnen Tests verwaltet, ausführt und anschließend die Ergebnisse auswertet. Jeder einzelne Test ist eine weitere isolierte Teilanwendung des Testers. So kann jeder Test unabhänig von allen anderen betrieben werden.

Die Platform ist eine Webanwendung welche den Endnutzern eine grafische Oberfläche zur Verfügung stellt, um Webseiten zu überprüfen. Im Hintergrund setzt die Plattform auf den Tester auf. webifier Mail ist ein Dienst mit dem Links aus E-Mails überprüft werden können. Anschließend erhält der Sender eine E-Mail mit den Resultaten zurück.

Eine weitere Teilanwendung von webifier ist das Data-Modul. Es stellt eine Schnittstelle für den Tester bereit, um alle Testergebisse sammeln zu können. Das Statisitik-Modul ist die letzte Teilanwendung von webifier. Es setzt auf das Data-Modul auf und stellt Funktionen zur Auswertung aller Testergebnisse bereit.

Um die Techniken und Algorithmen von webifier verstehen zu können sind einige Grundlagen erforderlich, welche nun im nächsten Kapitel genauer vorgsetellt werden.

2 Grundlagen

In diesem Kapitel werden die Grundlagen, welche für das weitere Verständnis der Arbeit und der gesamten Anwendung notwendig sind, näher beschrieben. Zunächst werden die verschiedenen Technologien und Frameworks, sowohl des Frontends, als auch des Backends dargestellt. Anschließend werden einige gängige Angriffstypen im WWW erläutert.

2.1 Frontend Technologien und Frameworks

TODO Daniel

- Hypertext Markup Language (HTML)
- Cascading Style Sheets (CSS)
- JavaScript
- jQuery
- Bootstrap

2.2 Backend Technologien und Frameworks

In diesem Abschnitt werden nun alle Technologien und Frameworks vorgestellt welche in den Backends der einzelnen Teilanwendungen zum Einsatz kamen.

Wohl am häufigsten kam die Programmiersprache Java zum Einsatz. Java ist eine universal einsetzbare, nebenläufige, klassenbarierte und objektorientierte Programmiersprache. Sie wurde möglichst einfach gestaltet um von vielen Entwicklern genutzt zu werden. In ihrer Syntax ähnelt sie den Programmiersprachen C und C++. Außerdem ist sie stark und statisch typisiert. Vorallem aber zeichnet sich Java durch seine plattformunabhängigkeit aus. Diese wird dadurch umgesetzt, dass Java-Quellcode in plattformunabhängigen Byte-Code kompiliert wird, welcher von einer Java Virtual Machine (JVM) ausgeführt wird. Java ist eine Hochsprache, die mit Hilfe des so genannten „Garbage Collectors“ eine automatische Speicherverwaltung bereitstellt.[3]

In einigen Teilprojekten wurde das auf Java basierende *Spring*-Framework verwendet. *Spring* stellt eine vereinfachte Möglichkeit auf den Zugriff auf viele Application Programming Interface (API) der Standard-Version zur Verfügung. Ein weiterer wesentlicher Bestandteil des *Spring*-Frameworks ist die *Dependency Injection*. Hierbei suchen sich Objekte ihre Referenzen nicht selbst, sondern bekommen diese Anhand einer Konfiguration injiziert. Dadurch sind sie eigenständig und können in verschiedenen Umgebungen eingesetzt werden. Des weiteren bringt *Spring* eine Unterstützung für aspektorientierte Programmierung mit, wodurch mit verschiedenen Abstraktionsschichten einzelne Module abgekapselt werden können.[4]

Aufbauend auf dem *Spring* Basis-Modul werden noch weitere Module, wie beispielsweise Spring Security, Sprint Boot, Spring Integration, Spring Data, Spring Session oder Spring MVC.[5] Im folgenden werden die *Spring*-Mudule naäher erläutert, die für das weitere Verständnis der Arbeit notwendig sind.

Spring Boot

Mit Spring Boot können Anwendungen, welche das *Spring*-Framework nutzen, einfacher eintwickelt und ausgeführt werden, da dadurch eigenständig lauffähige Programme erzeugt werden können, welche nicht von externen Services abhängig sind. Hierfür bringt Spring Boot einen Integrierten Server mit, auf welchem die Anwendung bereitgestellt wird.[6]

[3] **javaspecification**
[4] **spring3**
[5] **springPivotal**
[6] **springBoot**

Spring MVC

Spring MVC ist sehr gut geeignet um Webanwendungen zu implementieren.[7] Hierfür können die diese in mehrere Abstraktionsschichten gegliedert werden. Beispielsweise in das User Interface (UI), die Geschäftslogik und die Persistenzschicht.[8]

Spring Data

Spring Data vereinfacht Datenbankzugriffe ungemein. Das Modul stelt APIs für fast alle gängigen Datenbankzugriffsschichten, wie JDBC (Java Database Connectivity), Hibernate, JDO (Java Data Objects) zur Verfügung. Aber nicht nur relationale Datenbanken werden unterstützt, sondern beispielsweise auch NoSQL-Datenbanken und Key/Value-Stores können Problemlos eingesetzt werden.[9]

In Verbindung mit Spring Data wurde eine *MongoDB* zur Speicherung der Ergebnisse eingesetzt. *MongoDB* ist eine Dokument orientierte anpassungsfähige und skalierbare Datenbank. Sie vereint viele nützliche Eigenschaften von Relationalen Datenbanken, wie Sekundärindizes, Auswahlabfragen und Sortierung mit Skalierbarkeit, MapReduce-Aggregationen und raumbezogenen Indizes. Außerdem gibt es bei MongoDB keine festen Schemata, weshalb großen Datenmigrationen normal nicht notwendig sind.[10]

Gewonnene und gespeicherte Daten müssen danach auch noch aufbereitet und visualisiert werden. Webifier setzt dafür auf die Programmiersprache R. R ist eine freie Programmiersprache, entwickelt für statistische Auswertungen und Visualisierungen. Sie zählt zu den prozeduralen Programmiersprachen. Die quelltextoffene Programmiersprache wird ständig weiterentwickelt. Zusätzlich gibt es eine Vielzahl an Packages, welche weitere Funktionalität bereitstellen. Diese sind über ein zentrales Repository abrufbar und so leicht einbindbar in den Quelltext.[11]

Ein wichtiger Bestandteil jedes großen Software-Projektes ist ein gutes Build-Management-Tool. Für webifier wurde *Gradle* als solches gewählt. Ein Build-Prozess besteht grundsätzlich aus zwei Teilschritten. Zum Einen aus dem kompilieren des Codes und zum anderen aus dem verlinkten der benutzen Bibliotheken.[12] Da das manu-

[7] **spring3**
[8] **springMvc**
[9] **springData**
[10] **mongodb**
[11] **R**
[12] **buildprozess**

elle Einbinden von Bibliotheken und compilieren des Codes bei großen Projekten sehr aufwändig und mühsam sein kann wird hier auf Build-Management-Tools wie *Gradle* zurückgegriffen. Um den Build für den Nutzer möglichst einfach zu gestalten verfolgt Gradle zwei Prinzipien. Das erste Prinzip ist *Convention over Configuration*, was bedeutet, dass soweit es geht ein Standardbuildprozess definiert ist und der Anwender nur die Parameter ändern muss die Projektspezifisch abweichen. Das zweite Prinzip nennt sich Don't Repeat Yourself (DRY). Hierbei geht es darum Redundanzen in der Konfiguration des Buildes zu vermeiden. Diese beiden Prinzipien helfen Gradle, dass meist kurze Build-Skripte ausreichen um komplexe Prozesse abzubilden.[13]

Die Kommunikation zwischen Server und Client erfolgt über Representational State Transfer (REST). Hierbei wird jedes Objekt in REST als Ressource definiert, welche über einen eindeutigen Uniform Ressource Identifier (URI) adressiert werden können. Über die HTTP-Methoden GET, PUT, POST und DELETE können diese Ressourcen geladen, erstellt, geändert oder auch gelöscht werden.[14]

Das Testen von potenziell gefährlichen Webseiten soll natürlich nicht direkt auf dem Server geschehen, da es sonst diesen potenziell gefährden könnte. Deshalb wird hierfür eine Virtualisierung benötigt um die Tests abgekapselt vom Gesamtsystem auszuführen. Dafür wurde Docker als Tool eingesetzt. Docker ist eine Open-Source-Software zur Virtualisierung von Anwendungen. Hierbei wird auf die Container-Technologie gesetzt. Container sind vom Betriebssystem bereitgestellte virtuelle Umgebung zur isolierten Ausführung von Prozessen. Ein Vorteil der Container gegenüber der herkömmlicher virtuelle Maschinen ist der vielfach geringere Ressourcenbedarf.[15]

2.3 Technologien und Frameworks der Tests

TODO Author: Daniel (Needs review)

In diesem Kapitel werden diejenigen Technologien und Frameworks erläutert, die zur Umsetzung der Sicherheitstests verwendet werden.

[13] **gradle**
[14] **rest**
[15] **docker**

TODO Author: Daniel (needs completion) Python ist eine Programmiersprache, die einen schnellen Projektstart ermöglicht und ist auf Integration von verschiedenen Systemen spezialisiert. Die Sprache wird von der Python Software Foundation nach Open Source Standards entwickelt. Die aktuellste Version ist Python 3.6.1, wobei bei der Implementierung der Tests keine einheitliche Version verwendet wirddiesen Nebensatz in Retrospektive, als Punkt zur Verbesserung?. Python zählt zu den dynamisch typisierten Programmiersprachen, was bedeutet, das es wie bei JavaScript**??** erst zur Laufzeit zu einer Typenprüfung kommt. Weiterhin werden Codeblöcke nicht durch Sonderzeichen (wie z.B. geschweifte Klammern in Java) gekennzeichnet, sondern definieren sich an der Einrückungstiefe.[16]

- Phantom JS
  TODO Daniel
- HTtrack
  TODO Samuel
- Resemble JS
  TODO Samuel

Um Webseiten mit allen ihren Ressourcen herunterzuladen wurde die freie Software *HTTrack* verwendet. Mit *HTTrack* können Webseiten in einem lokalen Verzeichnis gespeichert werden. Hierfür erzeugt das Programm rekursiv alle notwendigen Verzeichnisse und läd anschließend alle Ressourcen, wie HTML-, CSS- und JavaScript-Dateien, als auch Bilder und andere Dateien herunter. Außerdem ist es möglich automatisiert alle HTML-Links entsprechend zu modifizieren. Abschließend bietet HTTrack umfassende Konfigurationsoptionen um es für den optimalen Gebrauch anpassen zu können.[17]

Für die Analyse und den Vergleich von Bildern wurde auf die freie JavaScript-Bibliothek Resemble.js zurückgegriffen. Mit Resemble können jegliche Arten von Bildanalyse und Bildvergleich genutzt werden. Urspünglich wurde es für eine Bibliothek von Phanton JS entwickelt, kann aber inzwischen vielseitig eingesetzt werden. Resemble bietet einige Einstellungsmöglichkeiten um Bilder analysieren und miteinander vergleichen zu können. Als Resultat liefert es bei der Bildanalyse Helligkeits- und Farbwerte des Bildes. Beim Bildvergleich bekommt man den prozentualen Unterschied der

[16] **pythonHomepage**
[17] **httrack**

beiden Bilder, sowie einige zusatzinformationen. Außerdem ist es möglich mit Resemble.js ein Differenzbild mit der Hervorhebung der Unterschiede zweier Bilder zu erzeugen.[18]

Zu einer umfassenden Analyse gehört selbstverständlich auch die Analyse des Netzwerktraffics. Dazu wird ein entsprechendes Tool genutzt. Webifier nutzt für diesen Zweck den *Bro Network Security Monitor*. Bro ist ein Unix-basiertes Network Intrusion Detection System (NIDS).[19] Zudem ermöglicht Bro dem Nutzer den Netzwerktraffic zu loggen und mittels eigener Skriptsprache zu filtern.[20]

2.4 Angriffstypen

2.4.1 Malware

TODO Samuel

2.4.2 Request Header Investigation

TODO Daniel

2.4.3 JavaScript Port Scanning

TODO Jani

2.4.4 JavaScript IP Scanning

TODO Jani

[18] **resemblejs**
[19] **bro**
[20] **bro2**

2.4.5 Clickjacking

TODO Jani

2.4.6 Phishing

TODO Samuel

3 Konzept

3.1 Gesamtkonzept

3.1.1 webifier Tests

TODO Jani

3.1.2 webifier Tester

TODO Samuel

3.1.3 webifier Platform

TODO Daniel

3.1.4 webifier Mail

TODO Daniel

3.1.5 webifier Data

TODO Samuel

3.1.6 webifier Statistics

TODO Jani

3.2 Testarten

3.2.1 Virenscan

TODO Samuel

- Httrack (Umsetzung)
- Download aller Dateien der Webseite
- Scannen der Heruntergeladenen Dateien
  - Clamav (Umsetzung)
  - AVG (Umsetzung)
  - CAV (Umsetzung)

3.2.2 Vergleich in verschiedenen Browsern

TODO Daniel

3.2.3 Test auf Port Scanning

TODO Jani

3.2.4 Test auf IP Scanning

TODO Jani

3.2.5 Link Checker

TODO Daniel

- herausfiltern aller Links und nachgeladenen Ressourcen

3.2.6 Google Safe Browsing

TODO Daniel

3.2.7 Überprüfung des Zertifikats

TODO Samuel

- Auslesen der relevanten Informationen des Zertifikates der WEbseite
- Validierung des Zertifikates

3.2.8 Erkennung von Phishing

TODO Samuel

- Herausfiltern der Schlagwörter
- Finden möglicher Duplikate der Webseite
  - Erstes Schlagwort zu Top Level Domains
    - com
    - ru
    - net
    - org
    - de
  - Websuche nach den Schlagwörtern mittels Suchmaschinen

- DuckDuckGo
- Ixquick
- Bing

3.2.9 Screenshot

TODO Jani

4 Umsetzung

4.1 Gesamtanwendung

4.1.1 webifier Tests

TODO Jani

4.1.2 webifier Tester

TODO Samuel

4.1.3 webifier Platform

TODO Daniel

4.1.4 webifier Mail

TODO Daniel

4.1.5 webifier Data

TODO Samuel

4.1.6 webifier Statistics

TODO Jani

4.2 Tests

4.2.1 Virenscan

TODO Samuel

4.2.2 Vergleich in verschiedenen Browsern

TODO Daniel

4.2.3 Test auf Port Scanning

TODO Jani

4.2.4 Test auf IP Scanning

TODO Jani

4.2.5 Linkchecker

TODO Daniel

4.2.6 Google Safe Browsing

TODO Daniel

4.2.7 Überprüfung des Zertifikats

TODO Samuel

4.2.8 Erkennung von Phishing

TODO Samuel

4.2.9 Screenshot

TODO Jani

5 Analyse

6 Ausblick

6.1 Weitere Tests

6.2 Weitere Module

7 Fazit

7.1 Zusammenfassung

7.2 Bewertung der Ergebnisse