H. Langweg, M. Meier, B.C. Witt, D. Reinhardt (Hrsg.): Sicherheit 2018,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2018 43

Ein Werkzeug zur automatisierten Analyse von Identitätsdaten-Leaks

Timo Malderle[1], Matthias Wübbeling[1,2], Sven Knauer[1,2], Michael Meier[1,2]

Abstract: Schon vor den Leaks von Dienstleistern wie last.fm, Playstation-Network oder Ashley Madison war Identitätsdiebstahl ein relevantes Thema im Bereich IT-Sicherheit. Die deutsche Gesetzgebung fordert zumeist eine Veröffentlichung der Umstände in relevanten Medien. Trotz öffentlicher Bekanntgabe und Präsenz in einschlägigen Medien erreichen relevante Informationen oft nur wenige Betroffene. Durch solche Veröffentlichungen lässt sich der Missbrauch von personenbezogenen und persönlichen Daten durch Kriminelle weder verhindern noch kontrollieren. Individuelle Benachrichtigungen von Betroffenen können die Folgen von Identitätsdiebstahl abschwächen. Dabei sollten die Benachrichtigungen weiterführende Informationen über den Umfang des Leaks beinhalten, welche die Kritikalität der betroffenen Merkmale darstellen und auch über mögliche Maßnahmen informieren. Um eine individuelle Information auf Basis verfügbarer Identitätsdaten-Leaks zu gewährleisten, müssen diese normalisiert und analysiert werden. Aufgrund der großen Menge kursierender Identitätsdatensammlungen ist eine Automatisierung notwendig. Diese Arbeit dokumentiert eine Implementierung zur automatisierten syntaktisch-, semantischen Analyse und Normalisierung relevanter Merkmale öffentlich verfügbarer Identitätsdaten als Vorbereitung zur individuellen Benachrichtigung von Betroffenen.

Keywords: Identitätsdiebstahl; Identitätsdaten-Leaks; Cyber-Crime; Opfer-Warnung; Reaktive Sicherheit

1 Einleitung

Illegal kopierte Sammlungen von Identitätsdaten-Leaks, im Folgenden nur Leaks genannt, kursieren über unterschiedliche Medien in kriminellen Kreisen. Das Internet ist ein beliebter Platz zum Austausch dafür. Betroffene Personen erfahren häufig erst von der Existenz solcher Leaks, wenn deren eigene Identität illegal verwendet wird und es zu einem Schaden kommt. Selbst wenn Mainstream-Medien über solche Leaks berichten, erfahren viele Personen nicht von der eigenen Betroffenheit. Die Warnung dieser Personen ist daher ein Ziel, das Sicherheitsforscher seit geraumer Zeit untersuchen. Dazu wurden Dienste entwickelt, die jedoch auf das Mitwirken der Betroffenen selbst angewiesen sind. Diese müssen dort vorhandene Daten zum Abgleich angeben und erhalten anschließend Informationen über deren Existenz in vorhandenen Leaks. So ein Vorgehen kann für die

[1] Universität Bonn, Institut für Informatik 4, Friedrich-Ebert-Allee 144, 53113 Bonn
{malderle | matthias.wuebbeling | knauer | mm}@cs.uni-bonn.de
[2] Fraunhofer FKIE, Friedrich-Ebert-Allee 144, 53113 Bonn

doi:10.18420/sicherheit2018_03

breite Masse nicht zielführend sein, weshalb proaktive Warnungen durch geeignete Akteure dringend notwendig sind. Für die eindeutige Zuordnung zu einer betroffenen Person müssen Identitätsmerkmale aus vorhandenen Leaks ermittelt werden. Anschließend lassen sich Betroffene möglicherweise sogar über unterschiedliche, ermittelte Kommunikationswege benachrichtigen, um mögliche Schäden abzuwehren oder um die Betroffenen auf die Folgen von Identitätsdiebstahl vorzubereiten.
Die vorliegende Arbeit präsentiert einen Ansatz für die automatisierte syntaktische und semantische Analyse von Leaks und die Identifikation solcher Identitätsmerkmale, die mittelbar oder unmittelbar die Zuordnung zu einer Person erlauben. Dafür werden im folgenden Kapitel 2 zunächst verwandte Arbeiten vorgestellt, die sich der Analyse von Leaks und der Informierung von Betroffenen widmen. Im anschließenden Kapitel 3 wird die Sammlung von öffentlich verfügbaren Leaks im Internet skizziert, die bereits in einer anderen Arbeit umfangreich vorgestellt wurde. Kapitel 4 demonstriert anschließend die praktische Umsetzung der Analyse in einem Werkzeug und evaluiert die verwendeten Mechanismen zur Klassifikation der Identitätsmerkmale anhand großer öffentlicher Leaks. Zum Abschluss fasst Kapitel 5 die Inhalte zusammen und gibt einen Ausblick über weitere Arbeiten die im Rahmen des Projekts umgesetzt werden.

2 Verwandte Arbeiten

Die von einem Identitätsdiebstahl betroffenen Personen haben nur wenige Möglichkeiten, um zeitnah von einem solchen Vorfall zu erfahren. Eine Möglichkeit ist, die eigene Betroffenheit bei einem geeigneten Identitäts-Informationsdienst zu überprüfen. Beispiele für solche Dienste sind *have i been pwned* [Hu17], der *BSI-Sicherheitstest* [Bu17a] oder der *HPI-Leak-Checker* [Ha17]. Diese Dienste ermöglichen die Überprüfung, ob eine E-Mail-Adresse in einem Identitätsdaten-Leak enthalten ist. Dazu übermittelt der Nutzer seine E-Mail-Adresse über ein Eingabefeld auf der Website des Dienstes. Die dem Dienst vorliegenden Leaks werden anschließend auf das Vorhandensein der eingegebenen E-Mail-Adresse überprüft. Bei einem positiven Ergebnis wird der Anwender informiert, dass Teile seiner Daten in einem dem Dienst vorliegenden Leak vorhanden sind. Gegebenenfalls liefert der Dienst dem Nutzer zusätzlich bekannte Hashes oder Passwörter, welche mit der E-Mail-Adresse gemeinsam geleakt wurden.
Problematisch ist hierbei, dass diese Dienste nur die Überprüfung von E-Mail-Adressen anbieten. Benutzernamen, Telefonnummern oder Zahlungsmitteldaten können damit beispielsweise nicht überprüft werden. Für den Nutzer ist sowohl die Aktualität der den Diensten vorliegenden Leaks undurchsichtig, als auch der Inhalt der Leaks.
Für Online-Dienste, wie Social-Media-Dienste, gibt es verschiedene Ansätze, um zu registrieren, dass von ihnen die Identitätsdaten der Nutzer entwendet worden sind. Methoden mit Watermarking [KK12, SER12] oder Honeypots [BD15] kommen dabei in Frage. Diese Ansätze gehören zu den präventiven Maßnahmen. Sollte es zu einem Identitätsdaten-Leak kommen, hat der Online-Dienst die Möglichkeit dies zu erkennen, um geeignete Maßnahmen einzuleiten. Ob ein Online-Dienst solche Schutzmaßnahmen einsetzt, ist für den Benutzer

nicht ersichtlich. Werden diese Maßnahmen gar nicht oder nur unzureichend eingesetzt, können Identitätsdaten unbemerkt abhanden kommen. Die betroffenen Identitätsinhaber wären einer erhöhten Bedrohung ausgesetzt.
Weitere Arbeiten untersuchen die Verbreitung von Leaks [OMS16], deren Auswirkung auf die Privatsphäre durch die Verkettbarkeit von Identitätsdatensätzen [HN17], sowie deren wirtschaftliche Folgen [Bu16]. Darüber hinaus befasst sich eine Arbeit mit der Verarbeitung von Leaks und der Erkennung von Identitätsmerkmalen [Gr16]. Es ist aber nicht erkennbar, wie das dort vorgestellte System arbeitet. Zusätzlich ist unklar, wie mit Problemen umgegangen wird, die aus der vorgeschlagenen Lösung resultieren.

3 Sammlung von Leaks

Für eine umfassende Analyse öffentlich verfügbarer Leaks müssen diese zunächst gefunden und gesammelt werden. Öffentlich verfügbar ist ein Leak, wenn ein Zugriff ohne Zugangsbeschränkung möglich ist. URLs, die einen randomisierten String im Pfad besitzen, stellen keine grundsätzliche Beschränkung des öffentlichen Zugangs dar. Diese Arbeit basiert auf der technischen Umsetzung zur Sammlung von Leaks aus einer vorangegangen Arbeit [MWM18]. Da die weitere Analyse auf den zuvor gesammelten Leaks basiert, wird im nächsten Abschnitt das Konzept von Datensenken im Allgemeinen vorgestellt. Daran anschließend wird die stichprobenartige manuelle Analyse zur Vorbereitung auf eine automatisierte Auswertung dokumentiert.

3.1 Datensenken

Datensenken sind Speicherorte (beliebiger) Daten im Internet, die über eine URL referenzierbar sind. Einige Dienstleister stellen kostenfrei und ohne weitere Beschränkung Speicherplatz für solche Datensenken zur Verfügung. Dieses Angebot nehmen *Hacker* und *Datenhehler* in Anspruch und verteilen darüber unter anderem auch Leaks mit Identitätsdaten. URLs solcher Leaks werden von Datenhehlern oder von Hackern selbst weitergegeben. Dafür nutzen diese weitere Dienste wie spezialisierte Websites, Forensysteme oder andere soziale Medien.
Für die Sammlung von Leaks müssen die URLs der Datensenken verwendet werden, von denen die Inhalte der Leaks geladen werden. Der Erhalt der URLs ist dabei häufig automatisiert möglich, beispielsweise wenn Datensenken eine API mit entsprechenden Funktionen anbieten. Spezielle *Leak-Monitoring-Pages* aggregieren URLs mit unterschiedlich hohem Aufwand, die nicht unmittelbar automatisiert ermittelt werden können. Für die manuelle Untersuchung wurden Leaks sowohl von automatisiert als auch von nicht automatisiert nutzbaren Datensenken geladen. [MWM18]

3.2 Analyse der gesammelten Daten

Für die erste Auswertung der gesammelten Daten wurden 531 Leaks geladen. Zusammen verfügen diese Leaks über eine Sammlung von 3,33 Milliarden E-Mail-Adressen, wobei es nach Abzug der Duplikate noch 1,56 Milliarden E-Mail-Adressen sind. Vergleichbar ist die gesammelte Menge an E-Mail-Adressen mit der Menge an E-Mail-Adressen verwandter Dienste wie *have i been pwned* (4,72 Milliarden), wobei hier auch E-Mail-Listen (sogenannte Spam-Listen) ohne sonstige Informationen wie Passwörter verwendet wurden [Hu17]. Ein anderer Dienst namens *Vigilante.pw* besitzt ca. 3,56 Milliarden E-Mail-Adressen [vi17]. Die Größe einzelner Leaks variiert stark. Ebenso variabel gestaltet sich das Format der Daten. Im Laufe der Analyse wurde festgestellt, dass es keinen De-facto-Standard für das Format von Leaks gibt, sondern sich jeder Angreifer oder Datenhehler eines mehr oder weniger eigenen Formates bedient. Der Großteil der Leaks (394 von 533) wurde als CSV/TXT Datei veröffentlicht, gefolgt von (teils unvollständigen) SQL-Dateien (126). Aber auch andere Formate wie PHP (8), JSON (3), HTML (1) und XLS (1) waren vertreten. Zudem sind die Leaks häufig mit weiteren Daten wie ASCII-Art oder ähnlichen Szene-Merkmalen versehen. Da der Formatwechsel oftmals auch innerhalb einer Datei stattfindet, kann von einer Aggregation der Inhalte ausgegangen werden. Diese inhomogene Strukturierung der Daten ist für eine generische Syntaxanalyse ein mögliches Hindernis. Als Reaktion wird während der automatisierten Analyse jeder Datensatz aus einem Leak entsprechend seiner zeilenweisen Struktur in Blöcke unterteilt und anschließend weiter mittels Parser analysiert. Hierbei wird festgestellt, dass die Datensätze in den geleakten Dateien zumeist eine der folgenden Formen besitzen, welche in Teilen oder in einem gesamten Leak genutzt werden:

```
e-mail-adr.:passwort
e-mail-adr.:passwort-hash
e-mail-adr.:passwort-hash:klartext-passwort
e-mail-adr.:benutzername:paswort-hash:salt
nutzer-id:benutzername:e-mail-adr.:ip-adr:passwort-hash:salt
benutzername:e-mail-adr.:passwort-hash:salt:geburstag:klartext-passwort
Gesamte SQL-Tabelle als txt/csv
Ein kompletter SQL-Dump
```

List. 1: Beispiele möglicher Anordnungen von Identitätsmerkmalen in einem Leak

In List. 1 ist zu sehen, dass Trennzeichen dazu genutzt werden, um die in einem Datensatz vorliegenden Identitätsmerkmale voneinander zu trennen. Statt des in der vorherigen Auflistung genutzten Trennzeichens *Doppelpunkt* werden auch weitere Trennzeichen wie *Semikolon*, *Komma*, *Tab*, '\t', '\r', *Leerzeichen* oder ähnliche genutzt. Da nicht alle möglichen Trennzeichen zuvor bekannt sind, wird ein dynamischer Ansatz für die Erkennung genutzt. Dabei wird die syntaktische Analyse zusammen mit der semantischen Analyse durchgeführt.

Neben der Analyse der manuell gesammelten Daten wurden auch die automatisiert gesammelten Daten untersucht. Diese stammen aus einer Untergruppe der Datensenken, genannt Paste-Pages. Paste-Pages bezeichnen Websites, welche es ermöglichen, text-basierte Inhalte ohne Inhaltskontrolle öffentlich verfügbar zu verbreiten. Hierfür platziert der Nutzer seine

Inhalte auf der Website, wie beispielsweise Ausschnitte aus einer geleakten Datenbank, und teilt den generierten Link sichtbar für mögliche Interessenten. Die selektierten Paste-Pages lieferten täglich im arithmetischen Mittel 91 neue, relevante, aber nicht duplikatfreie Pastes für die Datenbank. Es konnten von April 2017 bis November 2017 insgesamt 16.000 Pastes gesammelt werden, die zusammen 14.168.206 Millionen E-Mail-Adressen enthalten.

4 Werkzeug zur automatisierten Analyse

Zur effizienten Benachrichtigung von Betroffenen nach einem Identitätsdatendiebstahl wird ein Werkzeug benötigt, welches gesammelte Leaks automatisch analysiert. Das Werkzeug soll unterschiedliche Detailtiefen bei der Analyse unterstützen. Beispielsweise kann ein Leak ausschließlich mit einem regulären Ausdruck nach E-Mail-Adressen durchsucht werden. Allerdings gehen zusätzliche Informationen aus dem Leak verloren. Der genannte oberflächliche Ansatz ist deswegen nicht zielführend. Ein erweiterter Ansatz ist, die Daten so zu analysieren, dass möglichst viele Identitätsmerkmale erkannt werden. Als Identitätsmerkmale werden alle Attribute einer Identität bezeichnet, welche dieser individuell zugerechnet werden können. Dies können neben persönlichen Merkmalen wie Name oder Geburtsdatum auch Identifikatoren wie Email-Adressen, Benutzernamen und zugehörige Passwörter sein. Ebenfalls zu den Identitätsmerkmalen zählen Attribute wie IBAN, Kreditkartennummern, Adresse oder Telefonnummern. Anhand der Analyse dieser Merkmale kann ein genaueres Risiko für diese Identität erhoben werden. Eine effektive und zeitnahe Benachrichtigung eines Betroffenen ist mit höherer Dringlichkeit zu betrachten, wenn Konto- oder Kreditkartendaten veröffentlicht wurden, als wenn Zugangsdaten eines Browsergames betroffen sind. Die detaillierte Analyse eines Leaks hat den Vorteil, dass sich weitere Kommunikationskanäle zur Warnung ergeben. Eine Anschrift etwa kann für eine postalische Benachrichtigung verwendet werden. Um dies umzusetzen, müssen Identitätsdaten syntaktisch geordnet werden, um sie anschließend semantisch zu analysieren. Dazu werden zunächst Trennzeichen erkannt, um anschließend die Identitätsmerkmale genauer zu untersuchen.

4.1 Erkennung der Trennzeichen

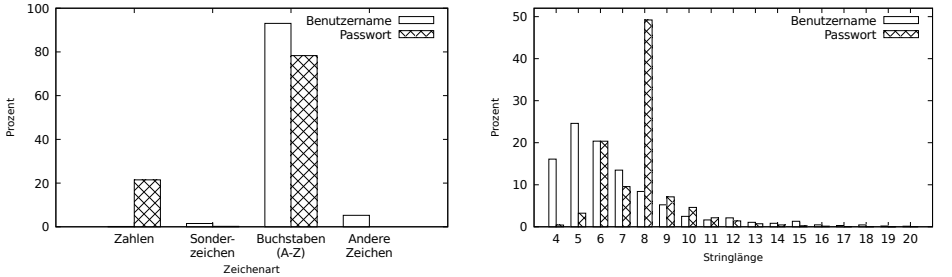
Die Identitätsmerkmale innerhalb eines Leaks sind mit einem Trennzeichen voneinander abgetrennt. Dabei können grundsätzlich beliebige Trennzeichen verwendet werden. Eine Möglichkeit zur automatisierten Analyse ist, die gängigsten Trennzeichen durch eine manuelle Analyse zu ermitteln. Wird in einem Leak allerdings ein anderes Trennzeichen verwendet, so können die Merkmale nicht mehr mittels Parser analysiert werden. Eine dynamische Erkennung der Trennzeichen ist demnach zielführender. Dazu werden zunächst reguläre Ausdrücke genutzt, die auch später zur Merkmalserkennung verwendet werden. Diese werden zeilenweise auf die Leaks angewandt, um die Zeichen vor und nach dem erkannten Ausdruck zu ermitteln. Aufeinander folgende Zeilen mit denselben syntaktischen

Strukturen werden über die Häufigkeit der ermittelten Trennzeichen zugeordnet. Alle zusammenhängenden Zeilen mit derselben Trennzeichen-Syntax werden zu einem Block zusammengefasst. Wird für eine Zeile ein anderes Trennzeichen als bei den vorangegangenen Zeilen ermittelt, beginnt das Verfahren für den neuen Block von vorne.

4.2 Erkennung von Identitätsmerkmalen

Für den Entwurf eines solchen Systems ist eine Identifizierung der Identitätsmerkmale notwendig, welche durch einen Parser automatisiert erkannt werden können. Durch eine manuelle Analyse der Daten kann festgestellt werden, dass folgende Merkmale in den Leaks vorkommen: *E-Mail-Adresse, Passwort-Hash, Salt, Passwort, Benutzername, Vorname, Nachname, Anschrift, Telefonnummer, IBAN, Kontonummer, Bankleitzahl, Kreditkartennummer, Geburtsdatum, ...* . Diese Merkmale besitzen verschiedene Eigenschaften. Beispielsweise lassen sich, wie schon zuvor gesagt, E-Mail-Adressen zuverlässig über die Syntax erkennen, da die Syntax vollständig definiert ist [Re01]. Kreditkartennummern besitzen auch eine erkennbare Syntax, allerdings existieren auch andere Zahlen von gleicher Länge, die als *False-Positive* erkannt werden. Kreditkartennummern besitzen eine Länge von 12 bis 19 Stellen und an der letzten Stelle eine Prüfnummer [Am16]. Um einen Großteil der *False-Positives* auszusortieren, kann bei jedem Wert, der über die Länge erkannt wurde, die Prüfnummer berechnet und kontrolliert werden. Eine effektive Erkennung der einzelnen Identitätsmerkmale anhand der Syntax ist abhängig von dem jeweiligen Identitätsmerkmal. Problematisch ist die Erkennung allerdings bei den Merkmalen *Passwort, Benutzername, Vorname, Name*. Ein Vor- und Nachname sollte in der Regel nur aus Buchstaben bestehen. Ob diese Einschränkung von dem jeweiligen Online-Dienst umgesetzt wurde ist offen. Generell bestehen diese vier Merkmale aus einem beliebig langen String, wobei dieser unter Umständen aus Buchstaben, Zahlen und Sonderzeichen bestehen darf. Eine Unterscheidung über die Syntax ist demnach nicht trivial.
Es wird ein weiterer Ansatz zur Erkennung der Merkmale benötigt. Dazu werden Listen recherchiert, welche die am häufigsten vorkommenden Elemente eines Merkmals beinhalten. Diese Listen werden mit den Elementen einer Spalte verglichen [Gr16]. Es werden Listen zur Erkennung der folgenden Merkmale genutzt: Passwort, Vorname [Bu17b, IA16], Nachname [Bu17b, Ve17], Bankleitzahl (vollständige Liste deutscher Banken) [De17]. Diese Listen bilden eine Sammlung von Vergleichslisten $L_1, ..., L_n$. Jede Liste enthält v Vergleichsbegriffe: $L_{1...n} = \{l \mid l \in [l_1, ..., l_v]\}$. Die zu analysierende Spalte M besitzt w Elemente: $M = \{m \mid m \in [m_1, ..., m_w]\}$. Gesucht ist die Liste L_t, die die größte Schnittmenge mit der gegebenen Spalte besitzt: $L_t = \{L_i \mid (|L_i \cap M|) \geq (|L_j \cap M|) \mid \forall i, j \in [1, ..., n]\}$. Im Anschluss muss ein Schwellwert y ermittelt werden, der von $(|L_t \cap M|)$ überschritten werden muss, um ein Ergebnis aufgrund des Grundrauschens auszuschließen: $\frac{(|L_t \cap M|)}{|M|} > y$. Zusammengefasst wird für eine bestimmte Spalte die Liste gesucht, die mit ihren Elementen am besten mit der Spalte übereinstimmt. Sollte die Übereinstimmung groß genug sein, so kann auf diese Weise auf den Inhalt der Spalte geschlossen werden.
Problematisch ist allerdings die Erkennung der Benutzernamen, da hierfür keine für das

Projekt ausreichende Liste gefunden werden kann. Aufgrund dessen wird ein weiteres Vorgehen benötigt. Eine weitere Möglichkeit zur Unterscheidung von Passwort-Spalten und Benutzernamen-Spalten ist es, dies anhand der Verteilung von Stringlängen und Zeichen der Merkmale aus den jeweiligen Spalten durchzuführen. Zu Untersuchung dieses Vorgehens wird der Leak *badoo* verwendet, da dieser Leak sowohl Klartextpasswörter als auch Benutzernamen enthält. Der Leak *badoo* enthält 112 Millionen Identitätsdatensätze. Teil eines solchen Datensatzes ist der Passwort-Hash, der Benutzername, die Telefonnummer und abhängig von der Version des Leaks auch das geknackte Passwort. Problematisch ist bei der Erkennung der Spalten bei diesem Leak lediglich die Passwort- und Benutzernamen-Spalte, da hierzu auf den ersten Blick ein semantisches Verständnis benötigt wird. Der Passwort-Hash, die E-Mail-Adresse, als auch die Telefonnummern können über die Syntax erkannt werden.
Vorstellbar ist, dass auch das Passwort und der Benutzername auf einer syntaktischen Ebene unterschieden werden können, indem die Verteilung von Passwortlängen und der Passwort-Buchstaben analysiert wird. Dies wird nun genauer betrachtet. In Abbildung 1 werden die Ergebnisse einer syntaktischen Untersuchung dargestellt. Es wurden aus dem genannten Leak die Spalten mit Benutzernamen und Passwörtern analysiert, indem die unterschiedlich oft vorkommenden Zeichen gezählt und die Längen der jeweiligen Merkmale festgehalten werden. Die Abbildung 1a zeigt die prozentuale Verteilung von Zeichen der jeweiligen Spalte in die Kategorien *Zahlen*, *Sonderzeichen*, *Buchstaben (A-Z)* und *andere*. Es ist zu erkennen, dass Passwörter einen deutlich höheren Anteil an Zahlen als Benutzernamen besitzen. In der Abbildung 1b ist die Häufigkeit der String-Längen von Passwörtern und Benutzernamen dargestellt. Auffällig ist, dass Passwörter mit der Zeichenlänge von acht Zeichen deutlich häufiger vorkommen, als Benutzernamen mit acht Zeichen. Diese Verteilungen können dazu genutzt werden, um Benutzernamen- von Passwort-Spalten zu unterscheiden.



(a) Prozentuale Verteilung von Zeichenarten

(b) Prozentuale Verteilung der Stringlänge

Abb. 1: Eigenschaften von Passwort und Benutzername

4.3 Anforderungen und Umsetzung

Um in allen vorliegenden Leaks die im vorigen Abschnitt beschriebenen Identitätsmerkmale möglichst eindeutig zu identifizieren, wird ein Framework für die Normalisierung der Daten erstellt. Ausgehend von einem einzelnen Leak zeigt Abbildung 2 die beteiligten Module und den Datenfluss der Normalisierung.
Das *Input-Modul* importiert Rohdaten von Leaks in unterschiedlichen Formaten (z.B. Archive, CSV- oder SQL-Dateien) und normalisiert diese, bevor die Daten an den *Separator* weitergereicht werden. Der *Separator* ermittelt die innerhalb des Leaks verwendeten vorhandenen Trennzeichen. Hierzu wird der Leak zeilenweise durchgegangen und per regulärem Ausdruck eindeutig erkennbare Felder, wie bspw. Email-Adressen, detektiert. Das Zeichen links und rechts vom gefundenen Feld wird extrahiert und als mögliches Trennzeichen vermerkt. Anschließend wird das gefundene Zeichen als Trennzeichen für die folgenden Zeilen verwendet und die Anzahl der so getrennten Spalten auf Gleichheit überprüft. So ermittelte Zeilen mit gleichem Trennzeichen und gleicher Spaltenanzahl werden als Block übergeben. Dieses Vorgehen ist notwendig, da viele Leak-Dateien eine Aggregation unterschiedlicher anderer Leaks darstellen. Das führt mitunter dazu, dass innerhalb einer Datei unterschiedliche Trennzeichen und auch eine unterschiedliche Reihenfolge der Spalten vorliegen können. Um diese Möglichkeit zu berücksichtigen, ermittelt der *Separator* möglichst große Blöcke, die dasselbe Trennzeichen verwenden. An diesem Punkt ist davon auszugehen, dass erkannte Blöcke zusammengehören und die Semantik der Spalten über den gesamten Block dieselbe ist.

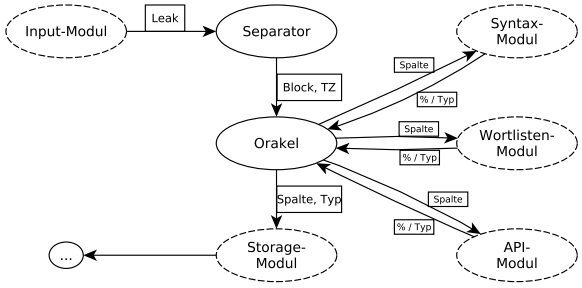


Abb. 2: Beteiligte Module und Datenfluss bei der Normalisierung und Speicherung von Leakdaten

Jeder Block wird mit dem ermittelten Trennzeichen an das *Orakel* übergeben. Das *Orakel* ermittelt die Semantik der Spalten eines Blocks. Dabei verwendet das *Orakel* in der aktuellen Ausprägung drei Module, die bei der Auswahl der Spalten-Semantik unterstützen: das *Syntax-Modul*, das *Wortlisten-Modul* und das *API-Modul*. Das *Orakel* teilt die Blöcke anhand des Trennzeichens in einzelne Spalten und übermittelt diese jeweils an die entwickelten Module. Als Rückgabe erhält das *Orakel* den Anteil der Übereinstimmung mit der Wortliste.

Das *Syntax-Modul* erlaubt die Zuordnung einer Spalte zu Bedeutungen, die anhand von Automaten ermittelt werden können. Dazu gehören vor allem solche Datensätze, die bei der Erstellung bestimmten Regeln folgen, die über reguläre Ausdrücke zu prüfen sind. Das *Syntax-Modul* prüft dabei mit regulären Ausdrücken auf die folgenden Typen: E-Mail, Telefonnummer, Bankleitzahl, Kontonummer, Kreditkartennummer, Hashwerte (z.B. gehashter Passwörter mit Salz, etc.), Datum, IP-Adressen. Dabei werden anschließend auch

Prüfsummen berechnet, wie etwa bei Kreditkartennummern, um nur Typ-konforme Daten zu berücksichtigen.

Das *Wortlisten-Modul* ermöglicht dem *Orakel* den Zugriff auf entsprechende Wortlisten. Dabei soll das gesuchte Merkmal durch einen Vergleich mit spezifischen Wortlisten identifiziert werden. Über das *Wortlisten-Modul* lassen sich die folgenden semantischen Typen von Spalten erkennen: Vornamen, Nachnamen, Klartextpasswörter. Um die Merkmale möglichst genau zu unterscheiden, werden spezifische Listen der entsprechenden Typen verwendet. Am Beispiel von Klartextpasswörtern lässt sich darstellen, weshalb umfangreiche Listen mit vielen Elementen nicht vorteilhaft bei der Merkmalserkennung sind. Passwörter besitzen durch die hohe Individualität bestenfalls eine hohe Entropie. Durch die große Menge individueller Passwörter würde bei der Berücksichtigung aller bekannten Passwörter eine hohe Anzahl an Zeichenkombinationen als Passwort erkannt, da diese auch als Passwort verwendet werden. Daher würde eine umfangreiche Wortliste aus Passwörtern mit hoher Wahrscheinlichkeit bei einer Spalte mit Vornamen positiv spezifizieren, möglicherweise sogar viel deutlicher, als eine Vornamens-Wortliste. Bei einer Beschränkung auf die gängigsten Namen und Passwörter, werden zwar die Trefferraten geringer, dafür werden die Aussagen genauer und einfacher differenzierbar, da viele False-Positives vermieden werden.

In Abbildung 3 ist die Evaluation der Wortlisten zu erkennen. Ausgeführt wird die Evaluation der genutzten Listen und des gedachten Vorgehens auf 100.000 Datensätzen des Leaks der *Modern-Business-Solutions*, welche Identitätsdaten aus der Automobilbranche und Personalvermittlung enthalten [Ch16]. Dieser Leak wird genutzt, da Vornamen und Nachnamen enthalten sind. Zur Evaluation der Passworterkennung werden 100.000 Datensätze eines Leaks einer Dating-Website verwendet, da dieser Leak Klartextpasswörtern beinhaltet [Wh16]. Beide Leaks wurden zunächst in einer manuellen Klassifikation begutachtet und die Spalten für Passwörter, Vornamen und Nachnamen identifiziert, alle anderen Daten wurden nicht weiter berücksichtigt. In der Evaluation wird nun getestet, wie signifikant sich Vornamen-, Nachnamen-, und Passwort-Spalten mit den ermittelten Wortlisten unterscheiden lassen. In Abbildung 3 sind auf der X-Achse drei Kategorien zu sehen: Passwort-, Vornamen-, und Nachnamen-Detektion. Jede dieser Kategorien besitzt eine Spalte mit 100.000 Attributen aus den zuvor genannten Leaks als Eingabe. Die verschiedenen Balken in der Abbildung stehen für unterschiedliche, getestete Wortlisten. Die Höhe eines Balkens gibt an, wie hoch die Übereinstimmung der jeweiligen Spalte aus dem Leak mit der Wortliste ist. Die Abbildung zeigt deutlich, dass die gewählten Listen eine gute Spezifizität besitzen, um die Semantik einer Spalte zu ermitteln. Bei dem Test der Passwort-Detektion ist zu erkennen, dass alle getesteten Passwort-Listen Übereinstimmungen mit der Passwort-Spalte des Leaks besitzen. Die anderen Wortlisten besitzen jedoch keine nennenswerte Übereinstimmung. Ein Passwort lässt sich somit effektiv erkennen. Bei den anderen Detektionen haben die Listen von der gleichen Kategorie wie die zu testenden Spalten auch eine deutlich höhere Übereinstimmung. Auch Vornamen und Nachnamen lassen sich somit effektiv erkennen.

Es wird ebenfalls aus Abbildung 3 deutlich, dass bereits kleine Listen ausreichen, um sich von den anderen Wortlisten-Typen zu unterscheiden. Obwohl die Übereinstimmung der

überprüften Passwort-Listen im einstelligen Bereich liegt, sind die Werte deutlich höher als die Rate der Vor- beziehungsweise Nachnamenslisten. Ebenfalls bemerkenswert ist die Tatsache, dass die Top-5.000 und Top-10.000 Passwort-Listen viele Vornamen enthalten. Die Trefferrate ist sogar deutlich höher als bei den Klartextpasswörtern selbst. Dies stellt an dieser Stelle kein Problem dar, weil die Vornamen-Listen deutlich höhere Trefferraten liefern und die Semantik den Spalten damit gut zugeordnet werden kann. Um die geringen Trefferraten der Passwort-Liste gegen False-Positive abzusichern, wird die im vorigen Abschnitt erwähnte Entropieanalyse im Anschluss durch das *Orakel* durchgeführt.

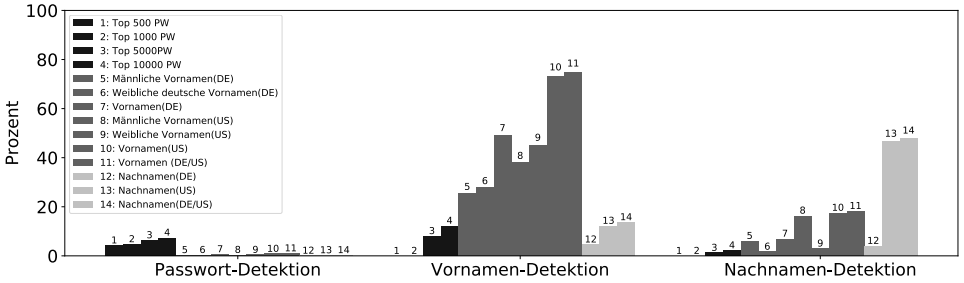


Abb. 3: Detektion von Merkmalen mit dem Wortlistenmodul

Als drittes Modul wird das *API-Modul* genutzt, um die Werte der jeweiligen Spalte gegen existierende Web- oder Offline-APIs zu testen. In der aktuellen Version soll die API eines großen Karten-Dienstleisters verwendet werden, um Postleitzahlen sowie Adressdaten zu erkennen. Auch für Telefonnummern lassen sich entsprechende APIs verwenden, wenn auch nur die Nummern erkannt werden, die im Telefonbuch eingetragen sind.

Nachdem das *Orakel* alle Spalten semantisch zugeordnet hat, werden die Daten der Spalte gemeinsam mit dem semantischen Typ der Spalte an das *Storage*-Modul weitergereicht. Das *Storage*-Modul unterstützt verschiedene Backends zur Speicherung der Leakdaten. Die aktuelle Implementierung speichert die Daten in einer MongoDB-Instanz. Aufgrund der geltenden Datenschutzgesetze sollen an dieser Stelle die Maßgaben der datenschutzrechtlichen Begleitung beachtet und umgesetzt werden. Insbesondere betrifft dies Datenreduktionsverfahren, beispielsweise zum Umsetzen von Sperrlisten auf Basis von Prüfsummen.

5 Zusammenfassung

Veröffentlichungen von Leaks mit Identitätsdaten sind heutzutage fast schon an der Tagesordnung. Die Dunkelziffer existierender Leaks ist vermutlich sehr viel höher. Es wird davon ausgegangen, dass viele Betroffene tatsächlich erst durch einen Schadenseintritt, also den tatsächlichen Identitätsdiebstahl, davon erfahren, dass ihre Identitätsdaten zuvor von Kriminellen kopiert wurden. Eine frühzeitige Benachrichtigung und Warnung der Betroffenen bereits nach dem Kopieren der Identitätsdaten durch Unbefugte ist daher wünschenswert.

Um Leaks zu verkaufen oder um die eigene Reputation zu erhöhen, veröffentlichen Kriminelle regelmäßig vorhandene Leaks ganz oder teilweise. Die weite Verbreitung durch solche Veröffentlichungen erhöht das Missbrauchspotential deutlich. Aus der Verfügbarkeit ergibt sich aber eine Möglichkeit für geeignete Akteure, Betroffene zeitnah zu ermitteln und über Risiken und Vorsichtsmaßnahmen zu informieren. Eine möglichst automatisierte Umsetzung zur Analyse von Leaks und Warnung von Betroffenen ist für die Verarbeitung der großen Datenmengen geboten.
Diese Arbeit demonstriert die automatisierte Ermittlung relevanter Identitätsmerkmale öffentlich verfügbarer Leaks. Dabei ist davon auszugehen, dass der Erhalt der Leaks von entsprechenden Datensenken technisch bereits einsatzbereit ist. Die Sammlungen von Identitätsdaten besitzen meist eine unbekannte und oftmals inhomogene Struktur, sowohl bezogen auf die Syntax als auch auf die Semantik der enthaltenen Daten.
Basierend auf regulären Ausdrücken, Prüfsummen, Wortlisten und öffentlich verfügbaren APIs wurde die Erkennung von Syntax und Semantik an öffentlich verfügbaren Leaks demonstriert. Die Zuverlässigkeit regulärer Ausdrücke wird dabei ergänzt um die Heuristik von Wortlisten und den Zugriff auf große Datensammlungen über API-Anbieter wie Geolocation-Dienste. Die Treffer-Quoten der jeweiligen Wortlisten wurden evaluiert und hinsichtlich falsch-positiver Zuordnungen optimiert.
Insgesamt bieten die in dieser Arbeit vorgestellten Ansätze zur automatisierten Analyse von Identitätsdaten-Leaks die Grundlage für die weitere Identifikation der betroffenen Personen auf Basis der enthaltenen Persönlichkeitsmerkmale. Zur Erreichung des intendierten Ziels einer effektiven Warnung Betroffener verbleiben jedoch offene Herausforderungen für künftige Arbeiten. Hierzu zählt insbesondere die Validierung von Identitätsdaten (Gültigkeit der Zugangsdaten) und im Zusammenhang damit eine Quantifizierung des für Betroffene bestehenden Risikos. Darüber hinaus ist zu untersuchen, welcher Art die Kommunikation und der dazu gewählte Kommunikationskanal zum Betroffenen entsprechen muss, sodass beim Opfer eine angemessene Wahrnehmung erreicht wird. Schließlich bleibt die Frage, wem die Anwendung der vorgeschlagenen Vorgehensweise zur Warnung Betroffener obliegen soll; fällt dies in den Aufgabenbereich öffentlicher Einrichtungen, gehört es zum Verantwortungsbereich der Dienstbetreiber oder sind zusätzliche Akteure gefordert.

Die Autoren danken dem Bundesamt für Bildung und Forschung (BMBF) für die Förderung des Projekts *EIDI* unter dem Förderkennzeichen 16KIS0696K.

Literaturverzeichnis

[Am16] American National Standards Institute: Announcing Major Changes to the Issuer Identification Number (IIN) Standard. `https://www.ansi.org/news_publications/news_story?articleid=da7bcb04-0654-4e03-af54-0e55d50b93a8`, 2016. Sichtung: 11.12.2017.

[BD15] Baykara, M.; Daş, R.: A Survey on Potential applications of Honeypot Technology in Intrusion Detection Systems. International Journal of Computer Networks and Applications (IJCNA), 2 Issue 5:202–211, 2015.

[Bu16] Bundeskriminalamt: Bundeslagebild Cybercrime 2016. https://www.bka.de/SharedDocs/Downloads/DE/Publikationen/JahresberichteUndLagebilder/Cybercrime/cybercrimeBundeslagebild2016.html, 2016. Sichtung: 11.12.2017.

[Bu17a] Bundesamt für Sicherheit in der Informationstechnik: BSI-Sicherheitstest. https://www.sicherheitstest.bsi.de/, 2017.

[Bu17b] Butler, R.: Most Common First Names and Last Names in the U.S. https://names.mongabay.com/, 2017. Sichtung: 11.12.2017.

[Ch16] Christian Hirsch (Heise Online): Offene Datenbank: 58 Millionen Datensätze im Umlauf. https://www.heise.de/newsticker/meldung/Offene-Datenbank-58-Millionen-Datensaetze-im-Umlauf-3351104.html, 2016. Sichtung: 11.12.2017.

[De17] Deutsche Bundesbank: Download - Bankleitzahlen. https://www.bundesbank.de/Redaktion/DE/Standardartikel/Aufgaben/Unbarer_Zahlungsverkehr/bankleitzahlen_download.html, 2017. Sichtung: 11.12.2017.

[Gr16] Graupner, H.; Jaeger, D.; Cheng, F.; Meinel, C.: Automated Parsing and Interpretation of Identity Leaks. S. 127–134, 2016.

[Ha17] Hasso-Plattner-Institut für Digital Engineering gGmbH: HPI Leak Checker. https://sec.hpi.de/leak-checker, 2017. Sichtung: 23.08.2017.

[HN17] Heen, O.; Neumann, C.: On the Privacy Impacts of Publicly Leaked Password Databases. In (Polychronakis, M.; Meier, M., Hrsg.): DIMVA 2017, S. 347–365. Springer, C., 2017.

[Hu17] Hunt, T.: have i been pwned? https://haveibeenpwned.com, 2017. Sichtung: 11.12.2017.

[IA16] IA7.de - InternetAgentur: Die schönsten Vornamen. http://www.mybabysitter.de/extras/vornamen/, 2016. Sichtung: 11.12.2017.

[KK12] Kale, S.A.; Kulkarni, S.V.: Data Leakage Detection: A Survey. IOSR Journal of Computer Engineering (IOSRJCE)), 1 Issue 6:32–35, 2012.

[MWM18] Malderle, T.; Wübbeling, M.; Meier, M.: Sammlung geleakter Identitätsdaten zur Vorbereitung proaktiver Opfer-Warnung. In: MKWI 2018. 2018. Wird auf Anfrage zur Verfügung gestellt.

[OMS16] Onaolapo, J.; Mariconti, E.; Stringhini, G.: What Happens After You Are Pwnd: Understanding the Use of Leaked Webmail Credentials in the Wild. IMC '16 Proceedings of the 2016 Internet Measurement Conference, S. 65–79, 2016.

[Re01] Resnick, P.: Internet Message Format. RFC 2822, RFC Editor, April 2001.

[SER12] Shabtai, A.; Elovici, Y.; Rokach, L.: A Survey of Data Leakage Detection and Prevention Solutions, 2012.

[Ve17] Verein für Computergenealogie e.V. : Die 1000 häufigsten Familiennamen in Deutschland. http://wiki-de.genealogy.net/Die_1000_häufigsten_Familiennamen_in_Deutschland, 2017. Sichtung: 11.12.2017.

[vi17] vigilante: vigilante.pw. https://vigilante.pw/, 2017. Sichtung: 11.12.2017.

[Wh16] Whittaker, Z. (ZDNet): A dating site leaked over a million accounts because of shoddy security. http://www.zdnet.com/article/dating-site-leaked-one-million-accounts-because-of-shoddy-security/, 2016. Sichtung: 11.12.2017.