Sammlung geleakter Identitätsdaten zur Vorbereitung proaktiver Opfer-Warnung

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

1 Universität Bonn, Bonn, Deutschland
{malderle|wuebbeling|mm}@cs.uni-bonn.de
2 Fraunhofer FKIE, Bonn, Deutschland
{vorname.nachname}@fkie.fraunhofer.de

Abstract. Identitätsdiebstahl ist eine häufige Folge digitaler Einbrüche bei Unternehmen und dem anschließenden Entwenden von Mitarbeiter- oder Kundendaten. Kriminelle erbeuten Identitätsdaten, um sie als Sammlungen an potenzielle Betrüger zu verkaufen oder aber sie selber zu benutzen. Die Warnung von Betroffenen nach einem Identitätsdiebstahl ist notwendig. Es gibt bereits verschiedene Dienste, bei denen sich Identitätsinhaber über den Status ihrer verwendeten Identitäten informieren können. Eine proaktive Warnung von potenziellen Opfern solcher Betrüger gibt es bisher jedoch nicht. Damit die Betroffenen zeitnah benachrichtigt werden können, müssen zunächst geleakte und verfügbare Daten vorliegen. Diese Arbeit präsentiert eine Methodik zur Sammlung aktueller Identitätsdatenleaks zur Vorbereitung proaktiver Opfer-Warnung.

Keywords: Identitätsdiebstahl, Identitätsdatenleaks, Cyber-Crime, Identitätsbetrug, proaktive Opfer-Warnung

1 Einleitung

Jeder Internetnutzer verwendet gewöhnlich mehrere Online-Dienste wie Foren, Socialmedia-Plattformen oder Online-Shops [1]. Bei diesen Diensten besitzt er zumeist ein eigenes Benutzerkonto, welches in den meisten Fällen eine E-Mail-Adresse und das gewählte Benutzerpasswort enthält [2]. Das Benutzerkonto stellt eine digitale Identität des Benutzers dar, wobei der Benutzer zum Identitätsinhaber dieser digitalen Identität wird. Jeder Benutzer kann bei unterschiedlichen Diensten verschiedene Identitäten besitzen. Digitale Identitäten sind für Kriminelle von großem Interesse, um sie für betrügerische Aktivitäten zu verwenden. In der Vergangenheit wurden immer wieder großen Mengen digitaler Identitätsdaten illegal kopiert und für jegliche Art von Betrug verwendet. In den meisten Fällen bemerken die betroffenen Identitätsinhaber erst sehr viel später, dass sie Opfer einer Straftat geworden sind. Häufig werden gestohlene Identitäten nicht nur für einen klassischen Betrug missbraucht, sondern verkauft oder für eine breite Masse zugänglich im Internet veröffentlicht. Dadurch wird das Risiko für den Identitätsinhaber deutlich erhöht, weil eine größere Anzahl potenzieller Betrüger Zugang zu den Identitätsdaten erhält. Es existiert nicht nur die Gefahr, dass

Multikonferenz Wirtschaftsinformatik 2018,
March 06-09, 2018, Lüneburg, Germany

gestohlene Zugangsdaten für kriminelle Zwecke missbraucht werden, sondern es bestehen zusätzlich datenschutzrechtliche Risiken. Verschiedene digitale Identitäten einer Person könnten unter Umständen miteinander verknüpft werden, um ein umfassendes Profil des einen Benutzers zu erhalten [3].

Deshalb ist es erstrebenswert, den betroffenen Identitätsinhaber über eine Entwendung seiner Identitätsdaten zu informieren, so dass präventive, bzw. reaktive Maßnahmen eingeleitet werden können. Da sich nicht jeder Webdienst neben seiner Haupttätigkeit mit der kontinuierlichen und aufwändigen Analyse von Identitätsdatenleaks und der Benachrichtigung seiner Benutzer beschäftigt und dies aus datenschutzrechtlichen Gründen vermutlich auch nicht darf [4], ist es sinnvoll, einen Dienstleister zu etablieren, der Privatpersonen und Unternehmen über Identitätsdatenleaks mit deren Identitäten informiert.

Für eine solche Benachrichtigung müssen mehrere Verarbeitungsschritte durchlaufen werden. Zunächst müssen gewöhnliche Quellen für Identitätsdatenleaks, sogenannte Datensenken, identifiziert werden. Anschließend werden die verfügbaren Daten aus den Datensenken geladen und ausgewertet, um auf dieser Grundlage gezielte Maßnahmen durchführen und Warnungen gegenüber dem Identitätsinhaber aussprechen zu können.

Die Basis dieser Warnungen bildet die Auswertung der Daten aus der Datensenke und somit das Wissen, welche Identitätsdaten entwendet wurden und öffentlich erreichbar sind. Problematisch hierbei ist, dass nicht alle Datenleaks ohne monetäre Gegenleistung veröffentlicht und herausgegeben werden. Beispielsweise werden manche Datenleaks auf dem Schwarzmarkt verkauft. Denkbar ist aber auch, das Datenleaks gar nicht veröffentlicht, sondern nur im Verborgenen missbraucht werden.

Die vorliegende Ausarbeitung beziffert die Menge frei verfügbarer Identitätsdatenleaks und analysiert den Aufwand, der für die kontinuierliche Beschaffung solcher Daten notwendig ist. Dazu wird eine systematische Herangehensweise gewählt, so dass zunächst existierende Datensenken identifiziert und anschließend die darin enthaltenen Daten beschafft werden. Zusätzlich wird untersucht, wie der beschaffte Datenbestand auch zukünftig aktuelle Daten integrieren und vorhalten kann. Diese Methodik wird zusammenfassend dargestellt.

Die Sammlung und Auswertung von Identitätsdatensätzen ist aktueller Forschungsgegenstand des Forschungsprojektes Effektive Information nach digitalem Identitätsdiebstahl (EIDI), das vom Bundesministerium für Bildung und Forschung gefördert wird. Das nachfolgende Kapitel 2 beschreibt verwandte Arbeiten im Bereich der Aufklärung von Identitätsdatendiebstahl. In Kapitel 3 wird die Sammlung von Datenleaks beschrieben, die anschließend in Kapitel 4 analysiert werden. In Kapitel 5 wird die Aktualität von Datenleaks und Identitätsdaten thematisiert, während Kapitel 6 das Modell der perpetuellen Datensammlung präsentiert. Kapitel 7 fasst die Ergebnisse zusammen und gibt einen Überblick über weitere geplante Arbeiten.

2 Verwandte Arbeiten

Dienste, wie *have i been pwned* [5], *Vigilante.pw* [6], *Hacked-Emails* [7], aggregieren große Datenmengen von Identitätsdatenleaks, damit Personen sich über die Betroffenheit der Datenleaks informieren können. Diese Dienste legen allerdings nicht offen, wie sie vorgehen, um an Sammlungen geleakter Daten zu gelangen.

Ein anderer Dienst, der zum Erhalt von Sammlungen genutzt werden kann und der häufig von Diensten der vorgenannten Art genutzt wird, ist *dumpmon* von Wright [8]. Dumpmon durchsucht Paste-Websites wie *Pastebin* [9] nach bestimmten Mustern, beispielsweise Emailadressen und Passwort-Hashes, um so Einträge (Pastes) mit Identitätsdaten zu erkennen. Anschließend informiert der Twitter-Account des Dienstes über entdeckte Identitätsdaten-Sammlungen. Die genutzte Software ist Open-Source und kann für eigene Projekte verwendet werden. Dieser Ansatz wird im Rahmen des vorliegenden Projekts als Einstieg genutzt, um weitere Arten an Datensenken einzubinden.

Verwandte Arbeiten im Bereich der Opfer-Warnung in Deutschland ist der *Identity-Leak-Checker* des Hasso Plattner Instituts (HPI) [10] und der *BSI Sicherheitstest* des Bundesamts für Sicherheit in der Informationstechnik (BSI) [11]. Beide bieten ähnliche Funktionen wie die bereits erwähnte Website *have i been pwned*. Nach der Angabe einer E-Mail-Adresse werden Informationen über die Existenz dieser E-Mail-Adresse in den verfügbaren Datensammlungen per E-Mail zugesandt. An dieser Stelle lassen sich mehrere Schwächen dieser Dienste aufzeigen. Potenziell betroffene Personen müssen diese Dienste kennen und zusätzlich regelmäßig nutzen. Problematisch ist auch, dass diese Dienste die in ihrem Datenbestand existierenden Identitätsdatenleaks nicht nennen. So kann nicht überprüft werden, ob ein solcher Dienst über aktuelle Datenleaks informieren kann. Eine Erweiterung dieser Dienste mit proaktiver Opfer-Information ist den Autoren nicht bekannt.

3 Die Sammlung von Datenleaks

Zur umfassenden Warnung betroffener Identitätsinhaber müssen ausreichend Daten vorliegen. Um die Handhabbarkeit und Skalierbarkeit der Daten als auch des Systems zu gewährleisten, wird eine Automatisierung angestrebt. Zusätzlich soll in einem gewissen Maße dafür gesorgt werden, dass der Datenbestand aktuell gehalten wird, um bei neuen Bedrohungen schnellstmöglich zu reagieren. Dazu werden zunächst existierende (halb-) öffentlich Datensenken analysiert, um anschließend die darin enthaltenen Daten für weitere Analysen zu sammeln.

3.1 Kategorisierung von Datensenken

Eine große Anzahl digitaler Identitäten werden regelmäßig von kriminellen Angreifern entwendet. In manchen Fällen werden die entwendeten Daten unmittelbar von diesen Angreifern genutzt, um betrügerische Handlungen durchzuführen. Oftmals

werden die Daten jedoch nicht von den Angreifern selbst genutzt, sondern vor der Nutzung durch Datenhehler weitergereicht oder verkauft.

Eine weitere mögliche Verwendung dieser Daten durch Kriminelle ist die Veröffentlichung, um eine Steigerung der eigenen Reputation in (cyber-) kriminellen Kreisen zu erlangen.

Geleakte Daten werden meist über eine sogenannte Datensenke verbreitet. Eine Datensenke ist der Dienst oder Speicherort im Internet, an dem Datenleaks gehandelt oder hinterlegt, beziehungsweise allgemein verfügbar gemacht werden. Datensenken können (halb-) öffentlich oder auf einen bestimmten Personenkreis beschränkt sein. Da diese Definition von einer Datensenke sehr umfassend ist, bedarf es zunächst einer Kategorisierung, um die Eigenschaften der jeweiligen Datensenke besser darstellen zu können. Dazu können Datensenken in die folgenden zwei Kategorien unterteilt werden: Automatisiert auffindbare Datensenken und Datensenken, die manuell identifiziert werden müssen.

Automatisiert auffindbare Datensenken sind beispielsweise einzelne Einträge auf *Paste-Pages* [12], wie *Pastebin* [9], über die sämtliche Arten von Texten mit anderen Personen geteilt werden können. Diese Paste-Pages werden häufig von Kriminellen dazu genutzt, Identitätsdaten aus Datenleaks zu teilen. Dabei kann die URL, die auf den jeweiligen Eintrag einer solchen Paste-Page verweist, als Datensenke bezeichnet werden. Viele dieser Paste-Pages bieten die Möglichkeit, neueste Veröffentlichungen anderer Benutzer einzusehen. Diese Funktion wird genutzt, um durch regelmäßiges Abrufen des Dienstes die relevanten Inhalte mit Identitätsdaten zu erhalten. Dazu werden alle neuesten unbekannten Einträge abgerufen und nach bestimmten Mustern, wie E-Mail-Adressen, durchsucht. Das Softwareprojekt *dumpmon* setzt diesen Ansatz um [8].

Manuell zu identifizierende Datensenken zeichnen sich dadurch aus, dass diese mit deutlich höherem Aufwand aufzufinden sind. Um Zugang zu diesen Datensenken zu erhalten, muss beispielsweise zunächst ein Hackerforum besucht werden. Dazu müssen häufig soziale oder semantische Hürden überwunden werden. Eine soziale Anforderung wäre, dass man von einem existierenden Forums-Mitglied geworben wird oder man in der Szene generell bekannt ist. Eine semantische Hürde könnte das Lösen von Captchas während der Registrierung sein. Diese Form der Identifizierung kann nicht ohne weiteres mit einem automatisierten Ansatz gelöst werden. Ein Grund dafür sind die divergenten Speicherorte, an denen die Datensenken abgelegt werden.

Mögliche weitere Speicherorte sind existierende *File-Hosting-Provider* [12], die anonymes Veröffentlichen von Dateien ermöglichen. Es können auch andere Teile des Internets als das Web verwendet werden, um Datenleaks zu verbreiten. Beispielsweise werden auch *BitTorrent* [12] oder das sogenannte *Usenet* genutzt. Diese Speicherorte können aufgrund ihrer Struktur und Datenmenge nur mit einem noch höheren Aufwand automatisiert ausgewertet werden. Für die Identifizierung solcher Datensenken benötigt es eine URL, die den Speicherort eindeutig beschreibt. Die URLs zu solchen Datensenken werden auf *Leak-Announcement*-Pages [12] weitergegeben. Dabei gilt es verschiedene Arten der Hinweis-Websites zu unterscheiden: Hacker Foren, Leak-Monitoring-Pages und Social Media [12].

Die Inhalte von Leak-Announcement-Pages verweisen per URL auf Speicherorte von Datensenken oder integrieren die Inhalte der Datensammlungen direkt. Leak-Announcement-Pages verwenden dabei verschiedene Sicherungsmechanismen, um die Inhalte zu schützen. Für den Zugriff bedarf es häufig der Registrierung eines Benutzerkontos. Dieser Vorgang ist meist durch Captchas gesichert und benötigt zusätzlich die manuelle Freischaltung durch einen Administrator des Dienstes. Die für die Registrierung genutzten E-Mail- und IP-Adressen können zur Ablehnung der Registrierung führen. Die tatsächlichen Kriterien sind den Autoren nicht bekannt. Nach der Freischaltung kann man je nach Dienst nicht alle Inhalte sofort einsehen. Eventuell wird eine soziale Beteiligung an den Inhalten des Forums gefordert. Beispielsweise müssen dann Beiträge zu Diskussionen oder eigene Datensammlungen veröffentlicht werden. In dem hier beschriebenen Projektkontext wurde eine Freischaltung nur mit gefälschten Dummydaten erreicht. Häufig hat dies allerdings zum zeitnahen Ausschluss aus dem jeweiligen Forum geführt, wobei erneute Versuche auch zum Erfolg führen können.

3.2 Identifizierung der Datensenken

Datensammlungen werden auf unterschiedlichen Wegen und von verschiedenen Kriminellen veröffentlicht. So tauchen Datensammlungen an verschiedenen Orten auf, an denen sie geteilt oder verkauft werden. Eine manuelle Recherche in gängigen Suchmaschinen identifiziert potenzielle Datensenken aus der Kategorie der manuell zu identifizierenden Datensenken. Weiterhin werden geeignete *Leak-Announcement-Pages* ermittelt. Insgesamt führt die Recherche zu einer Liste mit 15 solcher Dienste. Diese Liste wird gern auf berechtigte Anfrage zur Verfügung gestellt.

Gefundene Datensenken wurden in öffentliche, halböffentliche und geschlossene Datensenken eingeteilt. Öffentliche Datensenken sind solche, aus denen Datenleaks ohne die Überwindung technischer oder sozialer Sicherungsmaßnahmen entnommen werden können. Dagegen zeichnen sich halböffentliche Datensenken dadurch aus, dass sie eine zumeist soziale Sicherungsmaßnahme durchsetzen. Bei den gefundenen halböffentlichen Seiten war eine Registrierung notwendig, indem ein Benutzerkonto mit E-Mail-Adresse und Passwort angelegt wird. Weitere Angaben, zum Beispiel Name oder Anschrift, wurden nicht gefordert. Beide Formen erlauben eine (teil-) automatisierte Auswertung, wie in Abschnitt 3.1 dargestellt.

Geschlossene Datensenken sind solche Datensenken, die eine stärkere soziale oder technische Sicherungsmaßnahme umsetzen oder bei denen die Notwendigkeit eines finanziellen Aufwandes besteht. Natürlich wurden keine finanziellen Mittel oder realen personenbezogenen Daten als Tauschwährung aufgewendet. Eine automatisierte Auswertung ist dabei in der Regel nicht möglich.

Für die automatisierte Erfassung von Datenleaks werden folgende Dienste genutzt, von denen Identitätsdaten bezogen werden können: *pastebin.com*[8], *pastie.org*[8], *slexy.org* [8], *micropaste.com*, *siph0n.net*, *pastelink.net*. Dazu wurde das Programm *dumpmon* [8] angepasst und weiterentwickelt. Dieses Werkzeug ruft regelmäßig neueste Inhalte der genannten Dienste ab und speichert diese in einer Datenbank.

4 Analyse der Ergebnisse der gesammelten Daten

In diesem Abschnitt werden die erhaltenen Datensammlungen von den im vorigen Kapitel ermittelten Datensenken analysiert. Bei der manuellen Sammlung wurden insgesamt 520 Leaks aus den zuvor identifizierten Datensenken geladen. Insgesamt befinden sich in diesen Dateien 3.332.370.763 E-Mail-Adressen. Die Anzahl eindeutiger E-Mail-Adressen beträgt jedoch mit 1.563.002.958 etwas weniger als die Hälfte. Verwandte Dienste besitzen vergleichbar viele Daten. Der Dienst *have i been pwned* besitzt 4,72 Milliarden E-Mail-Adressen, wobei E-Mail-Spam-Listen ohne weitere Informationen wie Passwörter mit eingerechnet werden [5]. Der Dienst *Vigilante.pw* besitzt 3,56 Milliarden E-Mail-Adressen [6]. Dieser Vergleich zeigt, dass in diesem Projekt Identitätsdaten in der gleichen Größenordnung identifiziert und im weiteren Verlauf analysiert werden. Die mehrfache Speicherung von E-Mail-Adressen hat zwei Ursachen. Zum einen sind unter den Datenleaks auch sogenannte Combination-Lists, die eine Aggregation und Aufbereitung anderer Datenleaks darstellen. Zum anderen kann eine E-Mail-Adresse als Identitätsteil auch für mehrere Dienste verwendet werden und taucht somit in mehreren unterschiedlichen Datenleaks auf.

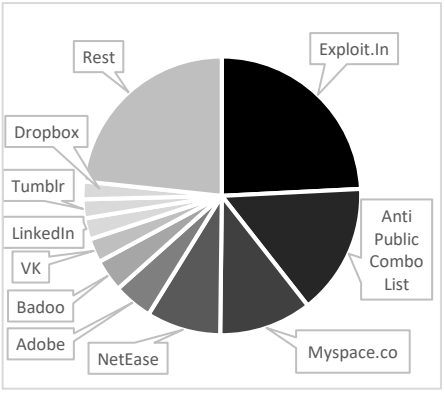

Abbildung 1. Anteile der einzelnen Leaks

In Abbildung 1 sind die Anteile der in der Datenbank enthaltenen E-Mail-Adressen dargestellt. Dabei werden die Top 10 Datenleaks getrennt und alle restlichen Datenleaks zusammengefasst dargestellt. Die Top 10 beinhalten zusammen 76,66 % aller gesammelten E-Mail-Adressen. Allein die drei Größten beinhalten mehr als 50 % der gesamten E-Mail-Adressen.

Bei der weiteren Analyse der einzelnen Datenleaks fallen stark unterschiedliche Formate der enthaltenen Daten auf. Es gibt offenbar kein einheitliches Format, das sich als De-facto-Standard durchgesetzt hat. Vielmehr stellt jeder Angreifer oder Datenhehler die Inhalte der Datensammlung in einem mehr oder weniger eigenen Format dar. Darüber hinaus gibt es Datensammlungen, die innerhalb einer einzigen Datei unterschiedliche Formate verwenden. Diese Tatsache deutet darauf hin, dass Inhalte verschiedener Sammlungen zusammengefasst wurden. Die verschiedenen Formate sind eine mögliche Hürde bei der automatisierten Syntaxanalyse.

Für die automatische Auswertung wird jeder Datensatz in einem Leak in mehrere Bereiche eingeteilt, abhängig von der jeweiligen Zeilenstruktur. Bei einer Analyse der vorhandenen Leaks wurde festgestellt, dass die Datensätze in den geleakten Dateien zumeist eine der folgenden Formen besitzen:

- email:password
- email:hash
- email:hash:cracked-password

- email:user:hash:salt
- userid:username:email:ip:hash:salt
- Gesamte SQL-Tabelle als CSV
- Ein kompletter SQL-Dump

Diese Auflistung zeigt die am häufigsten gefundenen Formate. Anstatt des in der vorherigen Auflistung genutzten Trennzeichens *Doppelpunkt* kann auch ein *Semikolon*, *Komma*, *Tab*, *\r*, *Leerzeichen* oder Ähnliches genutzt werden.

In Tabelle 1 sind die Identitätsdatenleaks mit den meisten E-Mail-Adressen absteigend aufgelistet. Die zweite Spalte der Tabelle zeigt den Dateinamen der Datensammlung. Die dritte Spalte listet die Gesamtsumme der E-Mail-Adressen auf, die in dem jeweiligen Datenleak zu finden sind. Die Spalte *TZ* zeigt das ermittelte Spaltentrennzeichen des Datenleaks und die letzte Spalte der Tabelle zeigt die Anzahl mit dem Spaltentrennzeichen gefundener Spalten. Diese können unterschiedliche Datentypen wie E-Mail-Adressen, Passwörter, Postadressen, Benutzernamen usw. beinhalten.

Tabelle 1. Top 10 der gesammelten Leaks

	Name des Leaks	*Summe E-Mail-Adressen*	*TZ*	*Spalten*
1.	Exploit.In	806581759	:	2
2.	Anti Public Combo List	506952026	:	2
3.	Myspace.com	358801827	:	5
4.	NetEase	287616356	:	2
5.	Adobe	152462193	|	6
6.	Badoo	125352387	[TAB]	8
7.	VK	92735323	:	2
8.	LinkedIn	82208762	:	2
9.	Tumblr	73357032	:	2
10.	Dropbox	68678625	:	2

Nach der Analyse der manuell gesammelten Daten, werden die automatisiert gesammelten Daten ausgewertet. Bei den genutzten Diensten, auf denen automatisiert aufzufindende Datensenken gespeichert werden, gibt es zum Teil die Möglichkeit, über eine API alle zuletzt veröffentlichten Pastes abzurufen. Diese Funktion wird genutzt, um möglichst viele vorhandenen Pastes abzurufen. Da dort alle Arten von Texten geteilt und veröffentlicht werden, muss zunächst eine Auswahl der für dieses Projekt relevanten Texte erfolgen, da die Menge zu speichernder Daten sonst zu umfangreich ist. Die Auswertung und Auswahl relevanter Pastes erfolgt mittels vordefinierter Muster in entsprechenden Suchausdrücken. Konkret muss ein Paste mindestens drei E-Mail-Adressen enthalten, um als relevant zu gelten, da eine geringere Anzahl an E-Mail-Adressen auf eine E-Mail-Kommunikation zwischen zwei Personen hindeutet

und es sich hierbei um keinen Leak handelt. Alternativ zu den E-Mail-Adressen können auch jeweils drei IBAN- oder Kreditkartennummern auftauchen.

Die verwendeten Paste-Sites liefern im arithmetischen Mittel 91 neue Pastes pro Tag mit relevanten Informationen, die entsprechend gespeichert wurden. In dem Zeitraum von April 2017 bis Juli 2017 wurden so insgesamt 8.562.237 E-Mail-Adressen aus den 12.011 gefundenen Pastes importiert. Der größte Anteil, nämlich 85,3 %, der gespeicherten Daten wurde über den Dienst Pastebin, weitere 13,9 % über Slexy verbreitet. Die restlichen 0,8 % teilen sich auf die restlichen Dienste auf. Die meisten identifizierten Datensammlungen werden über Pastebin veröffentlicht. Das liegt vermutlich daran, dass dies der bekannteste Dienst dieser Art ist. Die Struktur der Paste-Inhalte gleicht der, die von *manuell identifizierten Datenleaks* bekannt ist. Allerdings fällt auf, dass Pastes häufig zu Beginn oder Ende des Textes noch kurze Fließtextabschnitt besitzen. In diesen finden sich Hinweise zu den enthaltenen Identitätsdaten, zum Beispiel ob es sich um einen Ausschnitt handelt, der aus einem größeren Datenleak stammt. Der gesamte Datenleak ist dann über eine angegebene URL käuflich zu erwerben. Häufig fehlen aber konkrete Angaben zu den geleakten Daten, weshalb es unbekannt ist, von welchen Diensten diese Zugangsdaten ausgeleitet wurden.

Beim Vergleich der Ergebnisse beider Herangehensweisen zur Sammlung von Identitätsdaten fällt ein deutlicher Unterschied auf. Die manuelle Sammlung ergab 3.332 Millionen E-Mail-Adressen, während die automatisierte Herangehensweise im gleichen Zeitraum nur 8,5 Millionen E-Mail-Adressen hervorgebracht hat.

5 Modell zur Identitätsdatensammlung

Dieser Abschnitt stellt die Methodik aus vorigen Erkenntnissen des Ablaufs zum Sammeln von Identitätsdatenleaks dar.

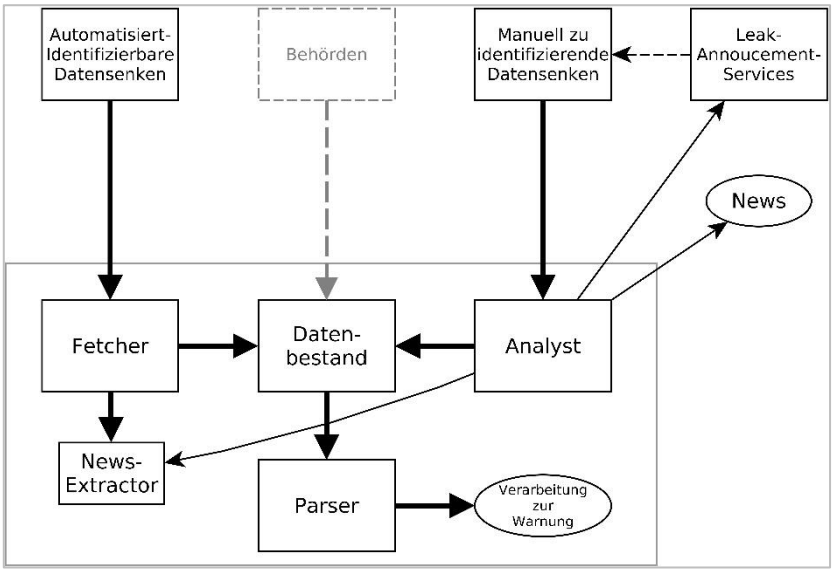

Abbildung 2. Modell zur Sammlung von Identitätsdatenleaks

Abbildung 2 zeigt die prototypischen Bestandteile des Prozesses zur Sammlung von Identitätsdatenleaks. Im oberen Teil der Abbildung sind die Quellen für die Identitätsdatenleaks dargestellt. Im linken Teil sind die *automatisiert identifizierbaren Datensenken* dargestellt, die regelmäßig von einem Softwaremodul mit der Bezeichnung *Fetcher* abgerufen werden. An dieser Stelle werden beispielsweise die letzten 100 Veröffentlichungen von Pastebin abgerufen und auf verwertbare Inhalte untersucht. Werden relevante Inhalte in einer Veröffentlichung gefunden, wird diese Veröffentlichung in der Projektdatenbank mit der Bezeichnug *Datenbestand* gespeichert.

Da *manuell zu identifizierenden Datensenken* nicht automatisiert gesammelt werden können, wird eine Person benötigt, die in dem Prozess als *Analyst* bezeichnet wird. Dieser recherchiert auf den *Leak-Announcement-Pages* relevante Datensammlungen. Werden relevante Inhalte gefunden, so kann der Analyst den Identitätsdatenleak über die gefundene URL aus der Datensenke herunterladen. Problematisch bei diesem Ansatz ist, dass aufgrund der manuellen Verarbeitung, die Daten in dem gepflegten *Datenbestand* nur schwer auf aktuellem Stand zu halten sind. Der *Analyst* muss entweder während der Recherche auf einer *Leak-Announcement-Page* einen neuen Datenleak ermitteln oder mit Hilfe einer externen Quelle über einen neuen Datenleak informiert werden. Dazu kommen beispielsweise Nachrichten-Websites, Blogs oder Twitter-Kanäle in Frage, die in der Abbildung mit *News* gekennzeichnet sind. Sollte

der *Analyst* durch *News* erfahren, dass ein neuer Identitätsdatenleak im Umlauf ist, so kann er gezielt nach diesem auf einer *Leak-Announcement-Page* suchen. Der *News-Extractor* hat die Funktion, automatisiert über neue Identitätsdatenleaks zu informieren. Dazu analysiert er den Fließtext der Pastes von Paste-Pages, falls dieser existiert. In diesem Fließtext wird, wie bereits erwähnt häufig auf weitere Daten des Identitätsdatenleaks verwiesen, da der vorliegende Paste nur eine Art Werbung bzw. Vorschau des gesamten Datenleaks darstellt. Diese Information kann genutzt werden, um den *Analysten* automatisch über neue Identitätsdatenleaks zu informieren.

Eine weitere Bezugsquelle für Identitätsdatenleaks kann auch durch Behörden dargestellt werden. Es ist denkbar, dass öffentliche Institutionen Interesse an der proaktiven Warnung betroffener Identitätsinhaber haben und deswegen Identitätsdatenleaks aus deren Besitz zur Warnung der Betroffenen zur Verfügung stellen.

6 Aktualität der Identitätsdatenleaks

Für die proaktive Warnung betroffener Identitätsinhaber ist es notwendig, die Aktualität eines vorliegenden Datenleaks zu bestimmen. Dabei muss die Aktualität differenziert betrachtet werden. Erstens ist die Aktualität des Datenleaks selbst von Interesse. Diese ist davon abhängig, wann die Daten von dem ursprünglichen legitimen Verwender kopiert wurden. Zweitens interessiert die Aktualität hinsichtlich der Verwendung durch Kriminelle, also Datenhehler oder Identitätsfälscher. Drittens ist die Identität der enthaltenen individuellen Datensätze relevant. Damit ist die Gültigkeit enthaltener Identitätsdaten gemeint, wie beispielsweise der E-Mail-Adresse oder eines Passworts. Die folgenden Abschnitte ermöglichen eine Unterteilung in Leak-Aktualität, Aktivitäts-Aktualität und Datensatz-Aktualität.

6.1 Leak-Aktualität

Die Aktualität des Datenleaks selbst lässt sich in den meisten Fällen nicht exakt datieren. Dies liegt daran, dass eine öffentliche Aufarbeitung von Datenleaks nur in seltenen Fällen stattfindet. Sobald aber eine Zuordnung eines Datenleaks zu einem in der Öffentlichkeit diskutierten Vorfall möglich ist, zum Beispiel Playstation-Network [13] oder Adobe [14], lässt sich über die Berichterstattung oft manuell eine entsprechende Datierung vornehmen. Wenn die enthaltenen Daten auch weiterführende Informationen zu den beschriebenen Accounts beinhalten, lässt sich beispielsweise über das Datum der letzten Benutzer-Anmeldung aller enthaltenen Identitäten eine zeitliche Eingrenzung vornehmen. Dies ist zum Beispiel im Falle des Last.fm-Datenleaks [5] möglich. Die neueste Anmeldung eines Benutzers ist in dem Datenleak auf den 22.03.2012 datiert, vermutlich wurden die Daten in unmittelbarer zeitlicher Nähe dazu entwendet. Somit ist eine näherungsweise Bestimmung der *Leak-Aktualität* möglich.

Die Information über die Aktualität des Datenleaks selbst lässt zunächst keine Rückschlüsse auf die Aktualität der enthaltenen individuellen Datensätze zu. Dafür ist eine weitere Betrachtung dieser notwendig.

6.2 Aktivitäts-Aktualität

Zur Bewertung des individuellen Risikos für Identitätsinhaber ist es notwendig, Informationen über die tatsächliche Nutzung von Identitäten aus einem Datenleak abzuschätzen. Nur, wenn die Daten tatsächlich noch aktiv durch Kriminelle genutzt werden, ist eine Information der betroffenen Identitätsinhaber sinnvoll. Wenn Datensammlungen beispielsweise im Rahmen einer Beschlagnahmung durch Strafverfolgungsbehörden analysiert werden, besteht theoretisch die Möglichkeit, dass die gefundenen Daten noch nicht an weitere Kriminelle weitergereicht wurden. Das Missbrauchspotenzial ist daher unmittelbar abhängig von der Weitergabe der Datensätze eines Datenleaks.

Innerhalb der vorgestellten Arbeiten werden ausschließlich aktiv geteilte Daten für die weitere Analyse verwendet. Es ist daher davon auszugehen, dass die enthaltenen Datensätze zu diesem Zeitpunkt auch für Kriminelle von Interesse sind und von diesen aktiv genutzt werden. Das Datum des Pastes, beziehungsweise des Forum-Beitrags auf dem ein Datenleak beworben wird, lässt sich daher als *Aktivitäts-Aktualität* verwenden.

6.3 Datensatz-Aktualität

Die Nutzbarkeit eines individuellen Datensatzes ergibt sich aus der tatsächlichen Aktualität der enthaltenen Informationen. Dabei wird zwischen *aktiver Nutzbarkeit* und *passiver Nutzbarkeit* unterschieden. Aktive Nutzbarkeit beschreibt die Möglichkeit für einen Identitätsbetrüger unmittelbar die beschriebene Identität missbräuchlich zu verwenden. Dafür sind Informationen über die ursprüngliche Verwendung der beschriebenen Identität sowie aktuelle Account-Informationen wie E-Mail-Adresse und Klartextpasswort notwendig. Es ist somit für den Angreifer möglich, sich bei dem Dienstanbieter anzumelden und den Dienst zu verwenden, was zu weiteren Möglichkeiten für den Identitätsbetrüger führen kann.

Die passive Nutzbarkeit ermöglicht einem Angreifer die Verwendung der enthaltenen Informationen zur mittelbaren Nutzung der dort beschriebenen Identität. Der Angreifer kann die enthaltenen Informationen nutzen, um im Namen des Identitätsinhabers weitere Accounts bei anderen Diensten zu registrieren und zu nutzen. Eine weitere Möglichkeit ist die Nutzung zur Passwort-Wiederherstellung. Häufig ist es dafür nötig, zusätzliche Informationen wie eine Kundennummer, Name und Vorname oder ein Geburtsdatum zu liefern, um ein neues Passwort setzen zu können. Auch die Antworten von zusätzlichen Sicherheitsfragen können grundsätzlich in Datenleaks enthalten sein und zum Account-Zugriff verwendet werden.

Um die Datensatz-Aktualität eines individuellen Datensatzes zu ermitteln, ist es notwendig, die aktive und passive Nutzbarkeit jedes Datensatzes weiter zu untersuchen. Eine Betrachtung der rechtlichen und technischen Möglichkeiten sind leider im

Rahmen dieses Beitrags aus Platzgründen nicht möglich, werden aber im weiteren Verlauf des Projekts intensiv betrachtet.

7 Zusammenfassung

Benutzerkonten und Zugangsdaten repräsentieren dienstspezifische digitale Identitäten der Nutzer von Internetdiensten. Über unterschiedliche Wege, z.B. durch Einbrüche in diensterbringende IT-Systeme, gelangen diese Identitätsdaten in unautorisierte Hände und werden für kriminelle Aktivitäten missbraucht oder von Datenhehlern handelsmäßig vertrieben. Im Zuge dieser Aktivitäten gelangen umfangreiche Sammlungen von Identitätsdaten oft für die breite Masse zugänglich ins Internet. Beispielsweise werden frei zugängliche Internetdienste zur Speicherung oder zum Austausch von Identitätsdaten als Datensenken für Identitätsdatensammlungen verwendet und missbraucht. Typischerweise erhalten betroffene Identitätsinhaber von diesen Vorgängen keinerlei Kenntnis und sind den resultierenden Bedrohungen weitgehend wehrlos ausgesetzt.

Mit der Vision diesen Missstand zu beseitigen systematisiert der vorliegende Beitrag die Abläufe und die beteiligten Dienste als auch Komponenten im Umfeld dieser Aktivitäten und schlägt einen Prozess zur systematischen Sammlung und Auswertung von Identitätsdatensammlungen zum Zweck der proaktiven Information betroffenen Identitätsinhaber vor. Dazu wurden unterschiedliche Arten von Datensenken und mögliche Strategien zur Datenauffindung und -sammlung untersucht sowie Herausforderungen beim Umgang mit variierenden Formaten der Inhalte von Identitätsdaten(-sammlungen) diskutiert. Die konzeptionellen Betrachtungen wurden mit der exemplarischen Analyse der Inhalte bereits aufgefundener Identitätsdatensammlungen untermauert. Gleichzeitig wurden offene Fragen zur Beurteilung der Aktualität von Identitätsdaten und damit der daraus resultierenden Bedrohung thematisiert.

Eine Schwierigkeit bei diesem hochgradig automatisierten Ansatz sind die Datenformate der Sammlungen, da sich diese stark voneinander unterscheiden. Es gibt keinen Standard, der sich für solche Datensätze durchgesetzt hat. Ein weiteres Problem ist, dass ohne weitere Informationen nichts über die Qualität der Daten ausgesagt werden kann, da es keine Anhaltspunkte über die Menge an gefälschten und nicht validen E-Mail-Adressen in den einzelnen Sammlungen gibt.

Zukünftige Arbeiten werden die inhaltliche Analyse der gesammelten Datenleaks thematisieren und die Problematik der Betroffenen-Identifikation adressieren. Dabei stehen anschließend die möglichen Kommunikationswege und das Design der Warnungen im Fokus der weiteren Untersuchungen. Die Autoren danken dem Bundesamt für Bildung und Forschung (BMBF) für die Förderungen des Projekts Effektive Information nach digitalem Identitätsdiebstahl (EIDI) unter dem Förderkennzeichen 16KIS0696K.

Literatur

1. Flo Encio, D., Herley, C.: A Large-Scale Study of Web Password Habits. WWW ’07. 16, 657–666 (2007).
2. Bonneau, J., Van Oorschot, P.C., Herley, C., Stajano, F.: Passwords and the Evolution of Imperfect Authentication. Commun. ACM. 58, 78–87 (2015).
3. Heen, O., Neumann, C.: On the Privacy Impacts of Publicly Leaked Password Databases. In: Polychronakis, M. and Meier, M. (eds.) DIMVA 2017. pp. 347–365. Springer, Cham (2017).
4. ULD - Unabhängiges Landeszentrum für Datenschutz Schleswig Holstein: Kurzpapier Nr. 8 Maßnahmenplan „ DS-GVO “ für Unternehmen. In: DSK. pp. 1–3 (2017).
5. Hunt, T.: have i been pwned?, https://haveibeenpwned.com. (Sichtung: 28.11.17)
6. @vigilante: Vigilante.pw - The Breached Database Directory, https://vigilante.pw/. (Sichtung: 28.11.17)
7. Chia, J.M.: Hacked-Emails, https://hacked-emails.com/. (Sichtung: 28.11.17)
8. Wright, J.: Dumpmon, https://github.com/jordan-wright/dumpmon. (Sichtung: 28.11.17)
9. Pastebin: Pastebin, https://pastebin.com, (2017).
10. Hasso-Plattner-Institut für Digital Engineering gGmbH: HPI Leak Checker, https://sec.hpi.de/leak-checker, (2017). (Sichtung: 28.11.17)
11. Bundesamt für Sicherheit in der Informationstechnik: BSI-Sicherheitstest, https://www.sicherheitstest.bsi.de/. (Sichtung: 28.11.17)
12. Jaeger, D., Graupner, H., Sapegin, A., Cheng, F., Meinel, C.: Gathering and Analyzing Identity Leaks for Security Awareness. In: Mjølsnes, S. (ed.) PASSWORDS 2014. pp. 102–115 (2015).
13. Arthur, C., Stuart, K.: PlayStation Network users fear identity theft after major data leak, https://www.theguardian.com/technology/2011/apr/27/playstation-users-identity-theft-data-leak. (Sichtung: 28.11.17)
14. Hern, A.: Did your Adobe password leak? Now you and 150m others can check, https://www.theguardian.com/technology/2013/nov/07/adobe-password-leak-can-check. (Sichtung: 28.11.17)