Abuse Detection

Peter Gawronski
Betreuer: Marcel von Maltitz
Seminar Innovative Internettechnologien und Mobilkommunikation SS2014
Lehrstuhl Netzarchitekturen und Netzdienste
Fakultät für Informatik, Technische Universität München
Email: peter.gawronski@tum.de

KURZFASSUNG

Wenn Nutzerdaten in sozialen Netzwerken gesammelt existieren, entsteht ein Anreiz für Angreifer, teils aus kriminellen Kreisen, diese Daten auszulesen. Der Betreiber versucht dies durch geeignete Vorkehrungen zu verhindern, sodass ein Wechselspiel zwischen den zwei Kräften entsteht.
Zunächst werden mehrere Vorgehen, Typen und Ziele von Angriffen erläutert, wie Angreifer versuchen, in das Netzwerk einzudringen, um Daten aus ihm abzugreifen.
Im Bezug darauf werden drei verschiedene Methoden für das Aufspüren von unerwünschten Nutzern oder Verhaltens dargestellt, die aktuell von den Betreibern von sozialen Netzwerken verwendet werden (können) mit ihren Vor- und Nachteilen bei der Angriffsabwehr und Abwägungen zur Umsetzbarkeit. Konkrete Beispiele aus den zwei größten sozialen Netzwerken Facebook und Twitter zeigen den realen Einsatz solcher Mechanismen.

Schlüsselworte

Abuse Detection, Spam Detection, Phishing, Machine Learning

1. EINLEITUNG

Ziel beinahe aller Angriffe auf verschiedene soziale Netzwerke ist es, Nutzerinformationen aus diesen herauszuholen, um auf diesen Daten weiterzuarbeiten oder gleich als Rohdaten an Dritte weiterzugeben. Dahinter stehen häufig monetäre Interessen, da sich solch gewonnenen Daten auf Untergrundmärkten verkaufen lassen, dabei steigen die Preise für Datensätze mit detaillierter persönlicher Nutzerinformation. Doch auch neuere Angriffsmethoden müssen erkannt werden, wie Spear-Phishing, bei dem der User mit direkt auf ihn zugeschnittenen Botschaften kontaktiert wird und durch die personalisierte Form eher dazu bewegt wird, auf den Betrug hereinzufallen.

Daher ist es das Bestreben der Anbieter solcher Portale, unerwünschtes Verhalten einzudämmen und derartige Angriffe zu unterbinden. In der Realität wird eine schnelle Reaktion und zielgenaue Erkennung von Angreifern angestrebt, sodass der soziale Graph, also die Verknüpfungsstruktur zwischen den normalen Nutzern, nicht ausgespäht wird und weitgehend unangetastet bleibt.

Zunächst werden in Kapitel 2 verschiedene Angriffsziele und -strategien dargelegt, um dann in Abschnitt 3 auf die zugehörigen Abwehrmaßnahmen mit verschiedenen Konzepten einzugehen. Mögliche Reaktionen bei verdächtigen Nutzern oder eindeutigen Treffern werden in Kapitel 4 erläutert, reale Probleme der Defensivmaßnahmen in Kapitel 5.
Konkrete Umsetzungen solcher Defensivmaßnahmen finden sich in Kapitel 6 mit abschließender Zusammenfassung und Ausblick.

2. ANGREIFERSTRATEGIEN

2.1 Ziele des Angreifers/Typen von Abuse

Angreifer versuchen je nach Interessenslage andere Ziele zu erreichen und nutzen daher Angriffe vielfältiger Art, um erfolgreich zu sein. Dabei sind einige Typen miteinander verknüpft, so kann Spam Ausgangspunkt für weitere Angriffe sein, Phishing und Scam können massenweise wie Spam versendet werden. Dahinter stehen meist monetäre Interessen oder im nicht gewerblichen Bereich auch persönliche Schmähungen.

2.1.1 Spam

Spam ist die Verbreitung von unerwünschten Nachrichten wie Werbung, Kontaktgesuchen oder URLs zu unerwünschten Seiten. Diese Ausprägung des Angriffs existiert nun seit etwas mehr als 36 Jahren [26] und wächst beständig weiter. Das massenweise Auftreten kostet den Empfänger viel Zeit, und damit Geld, zum aussortieren und löschen, die niedrigen Eigenkosten des Senders durch die weltweite Vernetzung machen ihn fast omnipräsent. Daher ist es eines der Hauptziele von Abwehrmechanismen, Spam einzudämmen [16]

2.1.2 Scam

Scam ist wörtlich Vorschussbetrug, also das Versprechen von Leistungen oder Waren gegen Vorkasse, welche dann nicht eingehalten werden. Der Begriff wird häufig auch verwendet im Zusammenhang mit Schneeballsystemen, bei denen ein einfacher Geldverdienst angepriesen wird, dieser aber real nie eintritt [25]. Dies ist häufig eine Unterart von Spam, jedoch mit deutlich höherer krimineller Energie und steht im Zusammenhang mit vielerlei Betrugsfällen. Hier warnte sogar das BKA wegen solcher Fälle [21].

2.1.3 Phishing

Phishing ist der gezieltere Zugriff auf Zugangsdaten legitimer Nutzer, da diese bereits soziale Beziehungen im Netzwerk haben, die Angreifer damit unerkannt bleiben und ist eine Form des Identitätsdiebstahls.
Häufig ist hier ebenfalls Spam der Einstiegspunkt des Angriffs, dabei lässt sich durch das massenhafte versenden von

doi: 10.2313/NET-2014-08-1_21

Phishingnachrichten eine große Gruppe Nutzer erreichen, sodass der Phish trotz geringer Quote doch Erfolg hat.
Allgemeines Phishing kann gravierende Konsequenzen haben: Wenn ein legitimer Nutzer Opfer eines Phishings wird, ist er möglicherweise plötzlich Teil eines Botnetzwerks und gleichzeitig für die vorherigen Verfahren nur schwer ermittelbar.
Je nach abgegriffenen Logins können später weitere Zugangsdaten bezogen werden, z.B. bei der Übernahme eines Mailaccounts, auf den Passwort-Wiederherstellungsmails anderer Plattformen gesendet werden.
Aber auch direkter Schaden, wie Zugriff auf Bankkonten, Diebstahl sensitiver Daten oder Firmendokumenten etc. ist möglich.

„It doesn't matter how many firewalls, encryption software, certificates, or two-factor authentication mechanisms an organization has if the person behind the keyboard falls for a phish.“ (Aus Hong, J. [8])

Phishing ist bei gut vorbereiteten Attacken für den Angreifer relativ teuer, verspricht jedoch deutlich höheren Gewinn als die anderen Methoden.
E-Mails mit Aufforderungen von Fremden, Webseiten aufzurufen oder Software zu installieren, werden von erfahrenen Nutzern ignoriert oder zumindest kritisch hinterfragt. Ist die Mail jedoch von einem (scheinbaren) Freund oder übergeordneter Stelle wie dem Administrator der Firma, steigt die Glaubwürdigkeit und damit die Erfolgswahrscheinlichkeit [8].
Solch personalisierten Phishingmails, die möglicherweise an einen einzelnen Nutzer gehen, sind dann auch schwerer zu filtern, da sie normaler Korrespondenz sehr ähnlich sehen können. Ein Webseitenbesuch kann mittels vielfältiger Exploits zu Malware auf dem Rechner des Nutzers führen, wenn dieser nicht gut gesichert ist. Dies kann dann Schadensszenarien wie Datendiebstahl, Spionage, Passwort- oder Bankdatendiebstahl nach sich ziehen und dem Angreifer weiter nutzen.
Die Aufforderung zum Download von Software wird von den meisten Nutzern sehr kritisch betrachtet und ist auffälliger als das ausnutzen von Exploits, wird diese scheinbar legitime Software installiert ergeben sich die selben Konsequenzen wie vorher.

2.1.4 Social Graph Crawling

Die persönliche Verknüpfung einer Person innerhalb eines Netzwerkes hat einen hohen Wert, denn aus diesen Angaben können vielfältige Rückschlüsse auf den Nutzer gezogen werden. Solche Information kann für andere Angriffe genutzt werden wie z.B. personalisiertem Spam oder auf den Nutzer zugeschnittene Phishingversuche. Dabei kann der Angreifer einen Lawineneffekt nutzen, indem er von den Ergebnissen weitersucht und erhöht mit jeder Stufe des Angriffs den Umfang der abgegriffenen Daten um ein Vielfaches.

2.1.5 Motiv: Monetäre Interessen

Bei fast allen zuvor genannten Angriffszielen stehen Finanzen im Hintergrund. Aus der Investition in Angriffe soll ein größerer Gewinn abfallen, kriminell oder nicht, oder es soll zumindest einem anderen geschadet werden, ohne selbst Gewinn abzuschöpfen.
Hierbei zeigen sich zwei relativ neue Felder: Schaden für den Anbieter und/oder Gewinn für den Angreifer durch Werbung[20, 4].
Werbung ist im Internet weit verbreitet, daher wird dies ebenfalls ausgenutzt:
Ein Angreifer versucht mittels künstlich erzeugten Klicks seine eigene Seite in Suchmaschinen höher zu platzieren (z.B. über +1 von Google+) und schaltet dort Werbung. Durch die bessere Positionierung erreichen mehr Nutzer seine Seite und sehen die Werbung, also verdient der Angreifer Geld.
Dies ist häufig im Rahmen von sog. Suchmaschinenoptimierung anzutreffen, bei denen Firmen ähnliches ganz öffentlich anbieten, Suchmaschinenbetreiber sich jedoch dagegen sträuben.
Eine neuere Methode ist, dass ein Angreifer Bots Werbung auf Seiten sehen öässt, die der Werbende bezahlen muss und erzeugt so finanziellen Schaden.

2.1.6 Stalking, Bullying, Mobbing

Stalking, Bullying, Mobbing gehören neben anderem zu den privaten Fällen des Missbrauchs. Dieser ist meist persönlich motiviert und nicht gewerblich, wie die anderen Ausprägungen. Diese Form ist sehr spezifisch und meist auf einen engen Kreis begrenzt, teils nur bestehend aus Täter und Opfer.
Hier spielt die Form der psychischen Verletzung eine größere Rolle, monetäre Hintergedanken sind eher selten.

2.2 Vorgehensweisen

Für die Vielzahl an Angriffen gibt es eine entsprechende Anzahl an Möglichkeiten, den konkreten Angriff auszuführen, die es einem Angreifer ermöglichen, seine Ziele zu erreichen.

2.2.1 Fake Accounts

Geläufigste Version des Angriffs sind gefälschte Accounts mit frei erfundenen Daten, um sich damit im Netzwerk anzumelden und von da aus weiterzuarbeiten.
Auf dieser Basis können alle hier erwähnten Angriffe aufbauen, daher ist es für den Betreiber wichtig, solche Nutzer zu erkennen und zu entfernen.

2.2.2 Botnetze

Ein Botnetz ist ein Zusammenschluss von mehreren Computern, entweder vom Angreifer selbst gestellt oder durch Malware ferngesteuerte Fremdrechner, die dann Angriffe auf das Netzwerk starten. Angriffe auf soziale Netzwerke zielen dann unter anderem auf das Erstellen von Fake Accounts oder Accountübernahme.

2.2.3 Accountübernahme

Tritt häufig im Zusammenhang mit Phishing (vgl. [3, 8]) auf, dort wird nach erfolgreichem Phishing der Account infiltriert und Nutzen daraus gezogen. Dies wird teilweise auch durch Datendiebstahl vom Server realisiert, wenn Accountinformationen nicht gesichert gespeichert wurden.
Wie in Abschnitt 2.1.3 beschrieben, ist dieser Angriff sehr gefährlich für das Opfer, da der Angreifer uneingeschränkten Zugriff auf dessen Daten und Zugänge hat und diese für weitere Angriffe und Manipulation nutzen kann.

3. ABWEHRSTRATEGIEN

3.1 Probleme, Beweggründe/Motive

Ein Anbieter sieht sich vor dem Problem, dass er Aufwand betreiben muss und damit Kosten hat, um Missbrauch innerhalb des Netzwerkes zu unterbinden. Mangelhafte Überwachung durch den Betreiber hat vielerlei Folgen:

3.1.1 Probleme

– Mehr ungewollte Werbung/Spam [6]:
Nutzer wollen Spam nicht sehen und werden unzufrieden, wenn sie zu viel erreicht. Großangelegte Kampagnen verschwenden überdies hinaus auch die Ressourcen des Anbieters und verursachen so Kosten.
– Falsch adressierte Werbung [nach [4]]:
Werbung an unerkannte Bots kostet den Werbetreibenden, dieser übt daher Druck auf den Betreiber aus, um auch nur für Werbung an reale Nutzer zu zahlen.
– Abgreifen des sozialen Graphen [1]:
Ein Angreifer versucht, die sozialen Kontakte eines Nutzers auszulesen. Je nach internen Richtlinien will der dies Betreiber verhindern oder zumindest soweit einschränken, dass es für einen Angreifer unattraktiv wird. Diese Information haben meist einen hohen Wert für Dritte, aber auch den Betreiber.

3.1.2 Konsequenzen

– Verfälschen von Ergebnissen [20, 14]:
Botnetze und Spammer können Sucheinträge verfälschen, so konnten bei Google Ergebnisse der Suche hervorgehoben werden durch Missbrauch der +1-Funktion oder bei Twitter unerwünschte Trendbegriffe entstehen. Derartigen Missbrauch von regulären Funktionen versucht der Betreiber zu vereiteln.
– Allgemein Schäden für den Betreiber, Nutzer oder legitimen Verwerter (z.B. legal Werbende):
Es gibt viele Arten des Missbrauchs, schlussendlich hat eine der Gruppen einen Schaden vielfältiger Art. Schäden für Nutzer oder Verwerter senken dabei die Zufriedenheit mit dem Anbieter, sie könnten abwandern. Daher ist es für den Betreiber erstrebenswert, dies effektiv zu verhindern.

Es ergibt sich eine weitere Notwendigkeit speziell bei persönlichen Angriffen aus Nutzer:
Die Plattform muss eine Möglichkeit besitzen, z.B. Mobbing oder Stalking oder Propaganda zu unterbinden. Dies ist mit den hier vorgestellten Methoden aber nur schwer möglich, da die Angreifer eine völlig andere Personengruppe darstellen und hier eher nur manuelle Überprüfung bleibt.

3.2 Abwehrtechniken

Drei verschiedene Techniken zur Erkennung bzw. Hervorhebung von Auffälligkeiten werden nachfolgend beschrieben und auf ihre Leistungsfähigkeit hin untersucht. Im Kapitel 6 werden hierzu real einsetzbare/eingesetzte Software gezeigt, die auf diesen Grundsteinen aufbaut.

Grundsätzlich ist die Abwehr ein Kreislauf, bei dem die Angreifer durch Verbesserung ihrer Angriffstechniken versuchen, mehr Daten abzugreifen, Spam zu verbreiten oder länger unerkannt zu bleiben und der Verteidigung, die diese Angriffsflächen abschirmt und schützt (vgl. Abb 1).

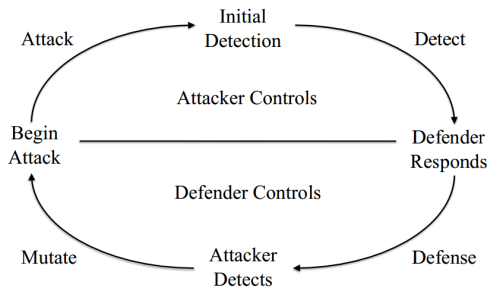


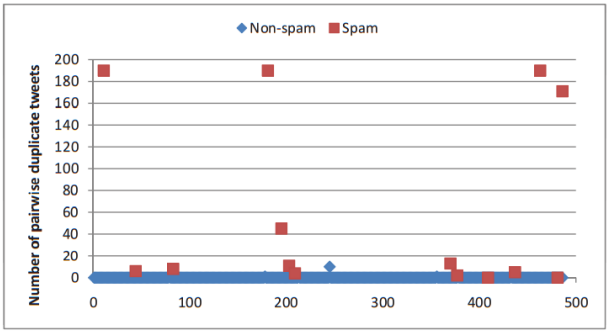
Abbildung 1: Der konfliktäre Zyklus, aus [10], S.1

3.2.1 Machine Based Learning

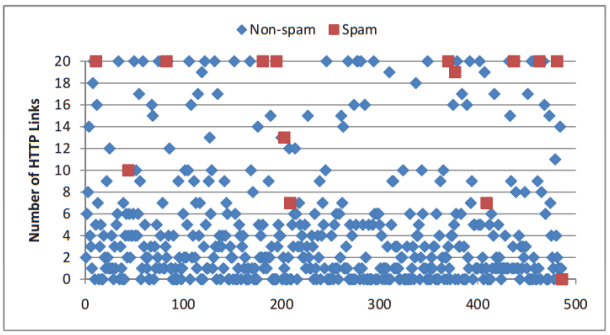
Ein häufiger Vertreter der Abwehr von unerwünschten Vorkommnissen ist das Maschinelle Lernen (ML) [10, 13]. Hierbei ist die Idee, dass man aus Erfahrungswerten lernen kann, welches Verhalten wahrscheinlich einem normalen Nutzer zuzuordnen ist und welches auf eine Maschine oder einen bösartigen Nutzer deutet. Im einfachsten Fall ist dies lediglich ein Schwellwert, meist jedoch mehrere gemessene Eigenschaften, die zusammen eine Wahrscheinlichkeit ergeben, mit der es sich um einen legitimen Nutzer handelt.
Dabei werden häufig Daten und Werte aus den jeweiligen sozialen Strukturen ermittelt, aber auch welche aus dem Verhalten der Nutzer.
So sind geläufige Metriken für Verknüpfungen die Anzahl der Freunde/Follower, meist unterschieden nach gestellten Anfragen/Abonnements und erhaltenen Anfragen/Followern, zusätzlich der Quotient aus den zwei Werten.
Hierbei zeigt sich, dass normale Nutzer häufig andere Werte erreichen als künstliche Accounts, so stellen Roboter bspw. viele Freundschaftsanfragen, erhalten selbst aber eher kaum welche oder sind nur innerhalb eines eigenen Netzes aus Bots verbunden.

Weitere Kriterien können aus dem Verhalten, insbesondere aus dem Inhalt ihrer Postings, gezogen werden (Abb. 2 für Twitter):
– Die Anzahl von duplizierten Posts (Abb. 2 (a)) ist ein sehr starker Indikator für Spam, da ein normaler Nutzer eher nie einen gleichlautenden Post verschickt. Ein spammender Account wird, bis auf in der Abbildung gefilterte Veränderungen im Hashtag oder @-Bezeichner, deutlich eher solch duplizierte Posts versenden.
– Zusätzlich zu diesem starken Indikator gibt es schwächere Indikatoren wie die Anzahl der Links und Hashtags in den Posts (c), hier gibt es jedoch auch normale Nutzer, die ähnliche Werte erreichen.
– Verbessert werden können diese Metriken, wenn man übergreifend nach Spamwörtern aus aktuellen Kampagnen sucht (Abb. 8): Es zeigen sich deutliche Unterschiede in der Häufigkeit von genutzten Spamwörtern, sodass sich Spamaccounts leichter filtern lassen.

Das Gesamtkonzept des Maschinellen Lernens beruht darauf, gute Metriken zur Klassifizierung zu finden und regelmäßig anzupassen, da sich mit ihnen die Qualität des Ansatzes entscheidet.



(a) The Number of Pairwise Duplications



(c) The Number of Links

Abbildung 2: Metriken zur Erkennung von Spammern bei Twitter, aus [14], S.8

3.2.2 *Social Graph Properties*

Betrachten der Eigenschaften des sozialen Graphen oder Verknüpfungsgraphen zwischen den Nutzern stellen einen unabhängigen Ansatz zum Maschinellen Lernen dar, denn hier wird weder auf das Verhalten, noch die direkte Beziehung unter den Nutzern geachtet, sondern der Gesamtgraph der Verknüpfungen und Beziehungen der Nutzer untereinander.

Begründet wird diese Vorgehensweise mit der hohen Wahrscheinlichkeit, dass Fake Accounts nur lose über wenige Beziehungen oder gar nicht mit realen Nutzern verknüpft sind, vielmehr häufig nur unter sich [4]. Reale Nutzer hingegen sind meist stark untereinander vermascht [4]. Damit ergibt sich eine Möglichkeit aus der Graphentheorie, die eher isolierten Gruppen des sozialen Graphen zu finden und diese als wahrscheinliche Bots zu markieren.

Zur Analyse werden bei Cao [4] mehrere Startpunkte gewählt, von denen dann das Netz abgelaufen/besucht wird. Dabei können verschiedene Metriken zu einzelnen Knoten im Netzwerk bestimmt werden.
So können isoliertere Gruppen im Netzwerk in den einzelnen Verfahren unterschiedlich schnell und präzise erkannt werden, sodass hier ebenfalls die Wahl eines geeigneten Algorithmus der Schlüssel zum Erfolg bleibt.

3.2.3 *Honeypots*

Ein bereits aus Zeiten vor sozialen Netzwerken stammendes Verfahren sind Honeypots, also in diesem Fall seinerseits künstliche Accounts, die jedoch von der Plattform betrieben werden und selbst keine Interaktion anstoßen. Daher ist es ungewöhnlich und verdächtig, wenn ein anderer Account mit diesem Fake Account kommuniziert oder Freundschaftsanfragen stellt/folgt/erwähnt [11].

Honeypots sind deshalb mächtig, weil ihnen jegliche Form von Abuse auffallen kann und sie unabhängig von Metriken und ähnlichen arbeiten können. Stringhini et al. [11] untersuchten dabei u.a. Facebook und Twitter mit einer großen Anzahl Honeypots und erhielten so Daten über Spamkampagnen und typisches Spamverhalten.

Durch Honeypots wird auch ersichtlich, welche Arten und wie viel Spam derzeit gesendet wird. Aus diesen Daten können dann z.B. Kriterien/Suchbegriffe zur weiteren Analyse für ML-Systeme generiert werden oder auch direkt Nachrichten an Kontrollpersonen des Netzwerkes zur manuellen Überprüfung von Accounts gesendet werden.

Nachteilig ist hier, dass Honeypots nicht aktiv suchen, sondern passiv warten und im Normalfall keine Verknüpfungen ins Netzwerk haben. Dies kann von intelligenten Angreifern erkannt und betreffende Accounts ignoriert werden, wenn sie merken, dass es sich wahrscheinlich um Honeypots handelt.

3.2.4 *Flagging*

Beim Flagging werden Nutzer oder Beiträge nicht automatisch, sondern durch einzelne Nutzer manuell markiert. An der Effizienz in Treffern je Meldung gemessen ist es ein schlechtes Verfahren, in der Literatur werden Quoten von ca. 5% genannt [4]. Jedoch ist dies das einzige Verfahren, dass gezielte Angriffe auf einzelne Nutzer wie Stalking, Mobbing und ähnliches abdecken kann.
Derartige Systeme sind auch außerhalb sozialer Netzwerke fast immer vorhanden und ermöglichen Nutzern, etwas direkt beim Betreiber anzuzeigen, sodass dieser geeignete weitere Schritte vornehmen kann.
Dieses Verfahren sei hier jedoch als Sonderfall zu betrachten, da es nicht automatisiert erfolgt.

3.3 Abwehr von Phishing

Vorhergehende Abwehrmaßnahmen sind hauptsächlich auf das Eindämmen von Spam ausgelegt und versuchen, falsche Accounts zu enttarnen, den Haupteinstiegspunkt für Spam. Phishing kann, wenn es wie Spam verteilt auf die Nutzer einwirkt, ähnlich wie dieser mittels automatisierten Einstufens durch Maschinen gefiltert werden.

Es gibt jedoch Möglichkeiten, auch stark personalisiertes Phising einzuschränken und aufzudecken (nach [8]), wenn der erste Schritt der Mailfilterung nicht erfolgreich war:
– Sichere Browser: Wie in Abb. 3 zu sehen, unterstützen viele Browser Blacklisten von phishing-verdächtigen Seiten, z.B. nutzt Mozilla Firefox eine von Google verwaltete Liste [23, 24]. Der Browser weist den Nutzer aktiv darauf hin, dass die derzeitige Seite wahrscheinlich nicht die ist, für die er sie hält. Bei besonderer Schwere blockieren Browser den Zugriff zunächst sogar komplett, um Exploits auf der Webseite zu verhindern. Auch Antivirensoftware schlägt bei solchen Bedrohungen häufig Alarm.
Zu sicheren Browsern gehört auch, dass sie sichere Verbindungen deutlich kennzeichnen wie in Abb. 4: Die originale

doi: 10.2313/NET-2014-08-1_21

Webseite wird zusätzlich als überprüft und sicher markiert.
– Sensibilisierung der Nutzer: Bereits häufig ist in E-Mails der Betreiber zu lesen, dass ihre Webseite einen Nutzer nie per Mail nach ihren Accountinformationen fragen würden, aber auch auf Webseiten selbst finden sich dazu Hinweise, die häufig für Phishingversuche nachgebaut werden.
Dieser Punkt ist für den Angreifer die größte Hürde: Ein aufmerksamer Nutzer erkennt Phishing meist leicht und selbst auf ausgeklügelten Seiten wird der Nutzer eher darüber nachdenken, geheime Informationen einfach preiszugeben oder lieber nochmals zu verifizieren, ob der Grund und die Seite legitim sind.

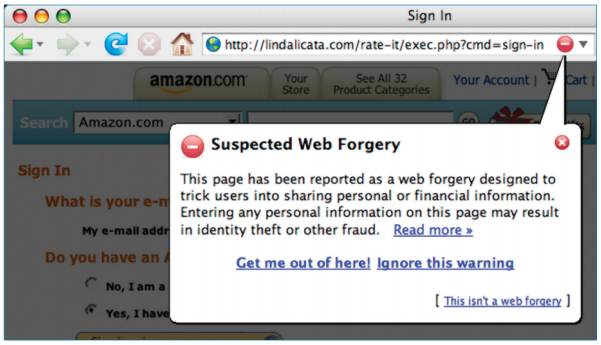


Abbildung 3: Aktive Warnung bei Firefox, aus [8], S.6

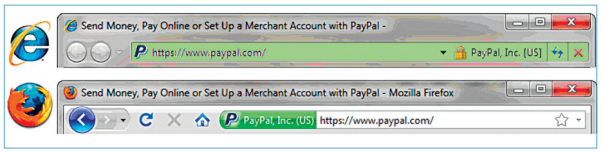


Abbildung 4: Validierung bei Internet Explorer und Firefox, aus [8], S.6

4. BEHANDLUNGSSTRATEGIEN

4.1 Fake Accounts

Wurde mittels eines der Verfahren aus Kapitel 3 mit hoher Wahrscheinlichkeit festgestellt, dass ein Account nicht einem echten Nutzer entspricht, sind mehrere Optionen möglich, diesen Account zu überprüfen:

4.1.1 Captchas

Captchas sind kurze Überprüfungswerkzeuge, die eine einfache Unterscheidung zwischen Mensch und Maschine ermöglichen sollen und sind weit verbreitet. Meist besteht ein Captcha aus einem kleinen Bild mit Textinhalt, der verzerrt/-verschleiert ist, sodass er für einen Menschen lesbar bleibt, eine automatische Texterkennung (sog. OCR: Optical Character Recognition) jedoch fehlschlägt. Dies soll automatisierte Anfragen an ein System verhindern, da für jede Anfrage im optimalen Fall nur ein Mensch die Aufgabe lösen kann, so die Kosten eines Angriffs steigen und damit vielleicht nicht mehr im akzeptablen/wirtschaftlichen Bereich liegen.

Bekanntestes Beispiel ist das 2009 von Google aufgekaufte reCAPTCHA, welchen dem Nutzer zwei Wörter oder Zahlen vorlegt und dieser diese in ein Textfeld tippt. Dabei ist lediglich ein Wort für die Erkennung relevant, das andere stammt aus nicht erkannten Worten aus der Digitalisierung von Büchern oder Straßenschildern von Google Streetview [18].
Vorteil ist die im Normalfall relativ einfache und kurze Überprüfungszeit für einen legitimen Nutzer, wohingegen ein Bot o.ä. scheitern wird und entfernt werden kann. Nachteilig ist die teils eingeschränkte Barrierefreiheit: Angreifer entziffern durch immer bessere OCR-Leser selbst schwierige Captchas, daher müssen diese wiederum komplexer werden. Schließlich werden sie aber auch für Menschen kaum lesbar, insbesondere bei eingeschränkter Sehfähigkeit.
Zu schwere Captchas erzeugen Frust bei legitimen Nutzern, daher ist die Anforderung an das System zweiteilig: Captchas sollen für Menschen einfach bleiben und nur Bots fernhalten und eine geringe Falschpositiv-Rate in Kapitel 3 gewünscht wird. Ein normaler Nutzer sollte im Normalfall nur bei der Registrierung ein einzelnes Captcha lösen müssen, um sich als Mensch zu identifizieren.

4.1.2 Überprüfung durch Menschen

Dies ist einer der am häufigsten eingesetzte Weg, Nutzer zu überprüfen und häufig auch die letzte bzw. endgültige Station bei der Bewertung, ob ein Account, Beitrag oder Verhalten um legitimen oder nicht legitim ist und welche Schritte eingeleitet werden müssen wie z.B. das Löschen des betreffenden Accounts oder Sperren/Verwarnungen gegen Nutzer, möglicherweise auch rechtliche Mittel.

Das spanische soziale Netzwerk Tuenti nutzte nach Cao [4] allein vierzehn Vollzeitbeschäftigte, um gegen Missbrauch vorzugehen. Facebook hat auch eine große Zahl Mitarbeiter, die sich ausschließlich mit der Bekämpfung von unerwünschtem Verhalten beschäftigen [10]. Abwehrstrategien haben auch immer das Ziel, diese Arbeitsleistung zu reduzieren bzw. die Trefferrate zu erhöhen, um die Effektivität des Aufwands zu steigern.

Vollständige Automatisierung ohne menschliche Hilfe lässt sich kaum erreichen, da es neben den Hauptangreifern wie Spammern oder Datensammlern auch die in Kapitel 2 genannten persönlichen Angriffe wie Mobbing oder Stalking gibt, welche durch Menschen von Hand gemeldet und von Menschen geprüft werden müssen. Auch anstößige Bilder müssen manuell von Menschen daraufhin geprüft werden, ob sie gegen die Nutzungsrichtlinien verstoßen und werden dann gelöscht.

4.1.3 Überprüfung befreundeter Accounts

Wie in Abschnitt 3.2.2 erläutert, sind Fake Accounts häufig gut untereinander vernetzt, haben jedoch nur sehr wenige Verbindungskanten zum Hauptgeflecht der Nutzer.
Daher erscheint es sinnvoll, ebenfalls die verknüpften bzw. befreundeten Nutzer zu überprüfen, ob sie nicht auch selbst nur künstliche Accounts sind, wenn sie nicht durch die Mechanismen sowieso schon markiert wurden [4].

4.2 Phishing

Wurde Phishing erkannt ist das Beseitigen der zugehörigen Webseite aus dem Netz von hoher Priorität, da ihr so nicht mehr Menschen zum Opfer fallen können [8]. Zwischenzeitlich sollte die Webseite von Browsern wie in 3.3 markiert

doi: 10.2313/NET-2014-08-1_21

und geblockt werden.
Ist ein Phishingversuch erfolgreich gewesen, sollte im Rahmen der Schadensbegrenzung der zugehörige Account gründlich untersucht und davon abhängige Dienste und Registrierungen ebenfalls überprüft werden und die zugehörigen Zugangsdaten geändert werden.

5. PROBLEME UND BESCHRÄNKUNGEN

Jeder der Ansätze aus Abschnitt 3 hat seine Beschränkungen und auftretenden Probleme, die für alle in ähnlichem Umfang gelten:

5.1 False Positives

Falsch positive Treffer sind legitime Nutzer, deren Verhalten dem eines Roboters zu sehr ähneln und daher auf der Liste der Verdächtigen landen. Soll nur ein Captcha gelöst werden, sind die meisten Nutzer noch bereit, dieses zu lösen. Wird der Aufwand größer, kann es jedoch schnell zu Ärger und Frust kommen, der sich dann negativ auf den Betreiber auswirkt. Dies passierte 2011, als Google seine internen Suchkriterien verschärfte [22] und zu viele falsch positive Treffer auftraten.
Daher müssen beim Entwurf und Test der Metriken beide Fehlerarten in einer Optimierungsphase berücksichtigt werden, damit weder zu viele Bots durchs Raster fallen, aber auch nicht viele Nutzer gestört werden.

5.2 Änderung des Botverhaltens

Dies ist ein direkter Teil des Zyklus aus Abbildung 1, in dem der Angreifer seine Methoden stetig nachbessert, um an den Abschirmmaßnahmen vorbeizukommen. Auch können atypische Bots für Spezialaufgaben genutzt werden (insbesondere personalisiertes Phishing), die dann von einer bestimmten Metrik schlicht nicht erkannt werden. Daher ist es für die Betreiber notwendig, mehrfache Absicherungen zu haben, sodass die Erkennung auch das neue Verhalten erkennt..

6. KONKRETE BEISPIELE

6.1 Facebook Immune System

Das FIS ist das Verfahren, welches von den Betreibern der Plattform selbst angewandt wird [10]. Dabei wird ein Ansatz des Maschinellen Lernens verwendet, der sich auf schnell ändernde Bedingungen anpassen kann und dabei quasi in Echtzeit auf das Geschehen reagiert, was bei einer Größe wie Facebook täglich ca. 25 Milliarden nutzergenerierten Aktionen entspricht.
Der Versuch des Auslesens des sozialen Graphen ist für die Entwickler ein Sonderfall, da der Angreifer, der ein Angriffsmuster erstellt, dieses möglichst lange unentdeckt lassen will, um weiter in den Graphen eindringen zu können.
Aufgebaut ist die FIS in verschiedene Teilbereiche, die ineinandergreifen: Klassifikatoren überwachen den Netzwerkverkehr und interagieren eng mit einem Regelwerk und auf der anderen Seite mit einer speziellen formalen Sprache zur Merkmalsfindung. Zusammengefasst entstehen so Regeln und Eigenschaften für die Kategorisierung von Verhalten und Beiträgen. Diese werden über Rückkopplungsschleifen zurückgespeist und rufen schlussendlich eine Reaktion, wie z.B. geeignete Maßnahmen bei Verdacht auf Fake Accounts.
Eines der Hauptaugenmerke der FIS ist das schnelle Anpassen der Umgebung auf eingehende Daten, um den Graph gegen alle Arten von Angriffen abzudichten und sich nicht auf ein spezielles Detail festzulegen. Dabei werden die häufiger auftretenden falsch positiven Treffer in Kauf genommen, denn eine leicht höhere Fehlerrate bei sehr wenig Nutzern trifft effektiv weniger Nutzer als eine niedrige Rate bei einer großen Masse an Betroffenen [10].

6.2 Automatisierter Angriff auf soziale Netzwerke

Balduzzi et al. [1] beschreiben einen einfachen, aber sehr effektiven Angriff auf verschiedene, große soziale Netzwerke: Diese erlauben es dem Nutzer, mittels der E-Mailadresse nach Freunden zu suchen und zeigen dann an, ob ein passender Account existiert.
Dieses Verhalten ist für den normalen Nutzer von Vorteil, da er so schnell viele Freunde und Bekannte erreicht, kann jedoch auch verschieden missbraucht werden:
- Gestohlene oder erfundene Mailadressen können verifiziert werden, ob eine reale Person dahintersteckt: Spamattacken treffen nun eher echte Nutzer.
- Werden auf verschiedenen Portalen die gleichen Adressen genutzt, handelt es sich wahrscheinlich um dieselben Nutzer.

Viel gravierender ist die lose Privatsphäreneinstellung vieler Portale: Die meisten zeigen öffentlich ein Foto, Freunde (und damit einen Teil des sozialen Graphen), Homepage, Geburtsdatum, Hobbies etc. Diese Information kann dann für gezielte Werbung oder auch für Phishingattacken ausgenutzt werden, die stark personalisiert durch die erhaltenen Informationen.
Auch der Weiterverkauf solcher Daten kann für den Angreifer lukrativ sein.
In dem Artikel zeigt sich ein Unbewusstsein der Plattformbetreiber gegen derartige Angriffe: Ihre Anfrage-APIs reagierten ohne Einschränkungen und sehr schnell auf E-Mail Abfragen, beispielsweise Facebook mit bis zu zehn Millionen Datensätzen am Tag von einem einzelnen Rechner aus.

Balduzzi et al. [1] zeigen auch Gegenmaßnahmen auf, wie die Nutzung von Captchas, die Beschränkung der Anzahl an Abfragen eines einzelnen Nutzers, eine Anfrageraten-Limitierung (z.B. wenige pro Woche) oder einer Verschleierung der Verknüpfung der Mailadresse zu Account.

6.3 SybilRank mit Random Walk

Ein von Cao et al. [4] entwickeltes Verfahren SybilRank nutzt zur Analyse des Netzes zufällige Einstiegspunkte in den (ungerichteten) sozialen Graphen und führt dann einen gekürzten Random Walk auf diesem aus, baut dabei ein Netz des Vertrauens auf und errechnet danach, wie eng ein Nutzer an diesem Netz hängt. Wie in Abb. 5 ersichtlich, haben echte Nutzer fast immer deutlich kürzere Wege, damit lassen sich Bots gut identifizieren.

SybilRank soll dabei laut Cao [4] deutlich effizienter sein als vergleichbare, ähnliche Verfahren und hat darüber hinaus in der Testumgebung eines spanischen sozialen Netzwerkes (Tuenti) sehr hohe Trefferquoten erzielt: 100% aus 50.000 sicheren Treffern und 90% aus 200.000 wahrscheinlichen Treffern, was über den Werten vergleichbarer ML-Techniken liegt und zeitlich effizienter ist als die verglichenen ähnlichen Ansätze.

Da die Annahme lautet, dass Bots nur schlecht zu realen Nutzern vernetzt sind, fallen prinzipbedingt diejenigen Bots durchs Raster, die bereits länger operieren und daher relativ gut immersiert sind. Dies bemängelt Yang et al. [15] und stellt seinerseits ein ML-Ansatz mit neuen Metriken vor, die auch solche Nutzer noch erkennen kann.

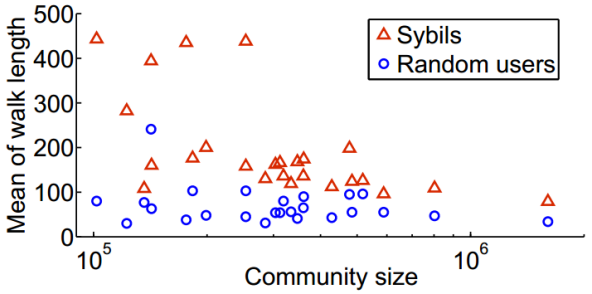


Abbildung 5: Mittlere Länge der Random Walks bei normalen Nutzern und Bots, aus [4] S.12

6.4 Maschinelles Lernen mit neuen Metriken

Yang et al. [15] stellen neue Metriken vor, die bereits existierenden deutlich überlegen sein sollen und sogar bessere Erkennung bieten als die zuvor genannten Eigenschaften des sozialen Graphen, Beispiele in Abb. 6 und 7. Beide Abbildungen zeigen die Verteilung (CDF) der jeweiligen Metriken, bezogen auf die Anzahl der versendeten Einladungen je Zeit, angenommener ausgehender und eingehender Freundschaftsanfragen und Verknüpfungsgrad des Nutzers. Auffällig sind die deutlichen Unterschiede der Anzahl angenommener ausgehender und eingehender Freundschaftsanfragen, dazu der Faktor der Verknüpfung der Freunde untereinander. Solche Differenzen eignen sich sehr gut für Klassifizierung mittels maschinellen Lernens, sodass laut Yang werden so auch bereits eingenistete Bots erkannt, die möglicherweise länger im Netzwerk sind und daher genügend Freunde haben, um nicht für Außenstehende gehalten zu werden.

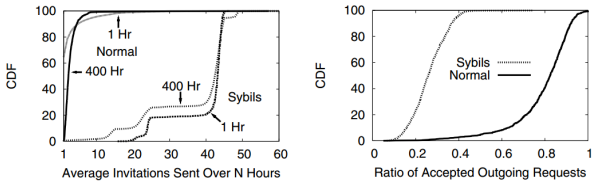


Abbildung 6: Vergleich zwischen Nutzer und Bot, aus [15], S.2

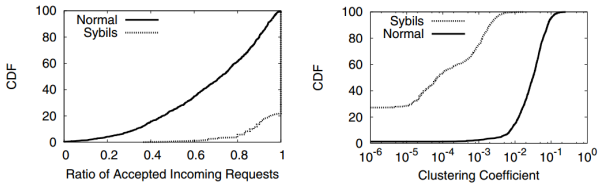


Abbildung 7: Vergleich zwischen Nutzer und Bots, aus [15], S.3

6.5 Honeypots: Analyse von Spam

Stringhini et al. [11] setzten eine Vielzahl von Honeypots in sozialen Netzwerken ein und stellten ihre Ergebnisse und Häufigkeiten der Spamkampagnen graphisch dar (Abb. 8). Dabei stellt der Durchmesser der Kreise das Volumen dar, aufgeteilt nach Kampagnentyp über die Zeit. Deutlich sichtbar ist die Verschiedenheit der einzelnen Kampagnen: So gibt es mehrere, die kontinuierlich laufen(1,5,6,7), zeitlich begrenzt sind (2,3,4,8) und teils sehr große Intensitäten erreichen können (4,6). Honeypots sind hier also sinnvoll, um das Ausmaß aktueller Angriffe auf verschiedenen Bereiche des Netzwerks zu erhalten, aber auch um z.B. die Effektivität von Abwehrmaßnahmen zu veranschaulichen.

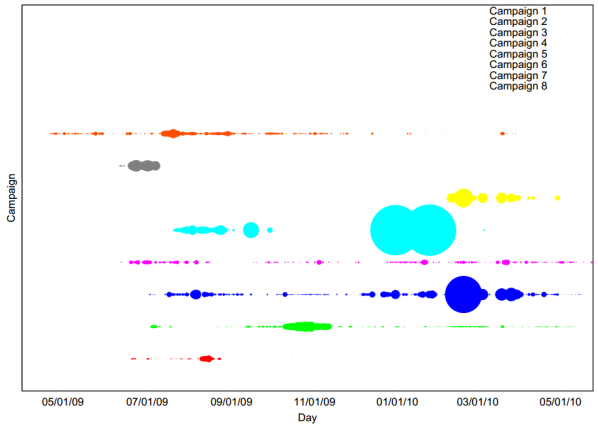


Abbildung 8: Spamkampagnen über Zeit, aus [11], S.8

7. ZUSAMMENFASSUNG/AUSBLICK

Insgesamt zeigt sich hier das Wechselspiel zwischen dem Angreifer und dem betroffenen Inhaber, der sein Netzwerk verteidigen muss. Die verschiedenen Beweggründe auf beiden Seiten werden den Wettlauf auch in Zukunft immer weiter treiben mit besseren Eindringstrategien auf der einen Seite und besseren Aufdeckmechanismen auf der anderen. Die Verteidigerseite versucht dabei den Großteil der Zeit die Oberhand zu behalten, da diese die Nutzerdaten schützt (vgl. Abb. 1) [10].

Maschinelles Lernen ist ein bewährtes Konzept, welches bereits länger aktiv eingesetzt wird und es daher viele Erfahrungswerte und Tauglichkeitsnachweise gibt. Es ist ein klassisches Verfahren, welches aus Eigenschaften eines Datensatzes Rückschlüsse auf die Eigenschaften zieht. Je nach Eignung und Auswahl der Parameter werden unterschiedliche Ergebnisse bei der Entdeckung gemacht, sodass diese Wahl essenziell ist und bei falscher Einstellung die Leistungsfähigkeit zunichte machen kann.

Hierbei sind Entwicklungen, wie die des Facebook Immune System [10] leistungsstarke Konzepte, wie aus einer großen Flut an eingehender Daten unerwünschte Nutzer oder Verhalten herausgefiltert werden können, um weitere Maßnahmen zu treffen. Das FIS nutzt dabei verhaltens- und verknüpfungsbasierte Parameter, um auffällige Nutzer zu zeigen.

doi: 10.2313/NET-2014-08-1_21

Echt graphbasierte Verfahren wie SybilRank [4] nutzen keine inhaltsbasierten Daten sondern beschränken sich auf das Auffinden von isolierten Gruppen innerhalb von Netzwerken, um künstliche Accounts zu identifizieren. Der soziale Graph als Ganzes wird hier statt eines Features des Accounts genutzt, was mit einigen Detaillösungen besser funktioniert als rein maschinelles Lernen.

Das 2011 vorgestellte Konzept von Yang et al. [15] erzielt mit einem schwellwertbasierten Verfahren, also eigentlich einem Prinzip des maschinellen Lernens, bessere Trefferquoten als zuvor. Dabei erkennt es vor allem auch für graphbasierte Verfahren unauffällige, da stark vernetzte, Accounts. Dies wird nur durch neue Metriken realisiert, die scheinbar besser geeignet sind als die zuvor genutzten.

Daher ist es schwer, einen klaren Favoriten aus den Verfahren zu wählen. Die Kombination aus mehreren Verfahren mit besserer Prävention der ersten Kompromittierung von Accounts ist dabei das Beste, der Aufwand dieser Maßnahmen muss dabei auch immer berücksichtigt werden.

In Zukunft wird der Kreislauf des Angriffs und der Abwehr weitergehen und es müssen neue Verfahren, Metriken und Konzepte entwickelt werden, um die Angreifer zu schwächen. Die hier vorgestellten Möglichkeiten sind dabei schon sehr mächtig und derzeit als Hilfsmittel stark genug, um sich gegen Angreifer zu wehren. Dabei dient der Mensch als Endkontrollorgan, um die wahrscheinlichkeitsbasierten Treffer zu bestätigen oder abzulehnen. Die Ergebnisse können dann in die Bewertungskriterien zurückfließen, um die Erkennung zu verbessern.

8. LITERATUR

[1] Balduzzi et al.: *Abusing social networks for automated user profiling*, In Proceedings of the 13th International Symposium RAID, Ottawa, Ontario, Canada, Seiten 422-441, Springer-Verlag Berlin Heidelberg, 2010

[2] Benevenuto, Fabricio, et al.: *Detecting spammers on twitter*, Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), Artikel 12, 2010

[3] Bonneau, Anderson, Danezis: *Prying data out of a social network*, ASONAM'09. International Conference on Advances in. IEEE, Seiten 249-254, 2009

[4] CAO, Qiang, et al.: *Aiding the detection of fake accounts in large scale social online services*, In Proc. of NSDI , 2012

[5] Danezis, George, and Prateek Mittal: *SybilInfer: Detecting Sybil Nodes using Social Networks*, In NDSS, 2009

[6] Gao, Hongyu, et al: *Detecting and characterizing social spam campaigns*, In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, Seiten 35-47, ACM, 2010

[7] Ghosh, Saptarshi, et al: *Understanding and combating link farming in the twitter social network*, In Proceedings of the 21st international conference on World Wide Web , Seiten 61-70, ACM, 2010

[8] Hong, J.: *The state of phishing attacks*, In Communications of the ACM, Seiten 74-81, ACM, 2012

[9] Irani, D.,Balduzzi, M., et al.: *Reverse social engineering attacks in online social networks*, In Detection of Intrusions and Malware, and Vulnerability Assessment, Seiten 55-74, Springer Berlin Heidelberg, 2011

[10] Stein, T., Chen, E., Mangla, K.: *Facebook immune system*, In Proceedings of the 4th Workshop on Social Network Systems, Atrikel 8, ACM, 2011

[11] Stringhini, G., Kruegel, C., Vigna, G.: *Detecting spammers on social networks*, In Proceedings of the 26th Annual Computer Security Applications Conference, Seiten 1-9, ACM, 2010

[12] Tan, Enhua, et al.: *UNIK: unsupervised social network spam detection*, In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, Seiten479-488, ACM, 2013

[13] Wang, Alex Hai: *Detecting spam bots in online social networking sites: a machine learning approach*, In Data and Applications Security and Privacy XXIV, Seiten 335-342, Springer Berlin Heidelberg, 2010

[14] Wang, Alex Hai: *Don't follow me: Spam detection in twitter*, In Proceedings of the 2010 International Conference on Security and Cryptography (SECRYPT), Seiten 1-10, IEEE, 2010

[15] Yang, Zhi, et al.: *Uncovering social network sybils in the wild*, In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, Seiten 259-268, ACM, 2011

[16] Zhang, X., Zhu, S., Liang, W.: *Detecting Spam and Promoting Campaigns in the Twitter Social Network*, In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, Seiten 1194-1199, IEEE, 2012

[17] Google Inc. reCAPTCHA, www.google.com/recaptcha/, Stand 14.06.2014

[18] Google Watch Blog Google nutzt Streetview-Fotos für reCAPTCHA, `http://www.googlewatchblog.de/2012/03/google-streetview-fotos-recaptcha/`, Stand 14.06.2014

[19] Bundesministerium für Arbeit und Soziales: *CAPTCHAs und Barrierefreiheit*, `http://tinyurl.com/ngmfz8c`, Stand 14.06.2014

[20] *Google Explores +1 Button To Influence Search Results*: `http://tinyurl.com/7g927oy`, 2011, Stand 14.06.2014.

[21] *Nigerianische Betrüger nutzen Internet*: `http://tinyurl.com/n9hqg2l`, 2006

[22] *Google+ Account Suspensions Over ToS Drawing Fire*: `http://tinyurl.com/5vrt524`, 2011, Stand 14.06.2014

[23] *Google-safe-browsing: Protocolv2Spec*: `http://tinyurl.com/64rssm`, 2009, Stand 20.07.2014

[24] *How does built-in Phishing and Malware Protection work?*: `http://tinyurl.com/q8oxvcf`, Stand 20.07.2014

[25] Polizei-beratung.de: *Vorauszahlungsbetrug*: `http://tinyurl.com/mu9xqyz`, Stand 20.07.2014

[26] heise.de: *Spam feiert 30. Geburtstag*: `http://tinyurl.com/mu9xqyz`, 2008, Stand 20.07.2014