Analyse von Traceroutes und rDNS Daten im Internet Census 2012

Stefan Liebald, Stefan König
Email: stefan.liebald@web.de, s.koenig@tum.de
Betreuer: Ralph Holz
Seminar Future Internet SS2013
Lehrstuhl Netzarchitekturen und Netzdienste
Fakultät für Informatik, Technische Universität München

KURZFASSUNG

In dieser Arbeit untersuchen wir die Ergebnisse des 2012 durchgeführten Internet Census im Hinblick auf Traceroutes und reverse DNS Einträge. Die Untersuchung richtet sich dabei unter anderem auf geographische Aspekte. Zur Auswertung verwendeten wir die spaltenorientierte Datenbank MonetDB sowie Tools zur Umwandlung von IP Adressen in die zugehörigen Autonomen Systeme, beziehungsweise die zugeordneten Länder. Aus den so gewonnenen Daten konnten wir die betroffenen IP Adressen der Traceroutes geographisch zuordnen und Einblicke in die Verteilung der Start und Zielgeräte gewinnen. Desweiteren war es uns möglich Einblicke in die Struktur des Internets zu gewinnen und interessante Traceroutes zu analysieren. Mittels der rDNS Einträge konnten wir eine Übersicht über die Verteilung der einzelnen Domains auf verschiedene Länder gewinnen, sowie die theoretischen Möglichkeiten diverser Länder aufdecken, auf Domains anderer Länder Einfluss zu nehmen.

Schlüsselworte

Internet Census 2012, Carna Botnetz, MonetDB, Traceroutes, rDNS, Auswertung

1. EINLEITUNG

Das Internet ist ein weltweit zusammenhängendes Netzwerk von Rechnern und Rechnernetzwerken. Bedingt durch den dezentralisierten Aufbau dieses enorm großen Netzwerks ist es nicht mehr möglich kurzfristig weiterführende Aussagen über den aktuellen Zustand des Gesamtnetzwerkes zu geben. Der verwendete Adressraum des aktuell noch am gebräuchlichsten Kommunikationsprotokolls IPv4 wird dabei mit 32 Bit Adressen beschrieben, was rund 4,3 Milliarden möglichen verschiedenen Adressen entspricht. Eine Analyse dieser großen Anzahl an verschiedenen adressierbaren Geräten gestaltet sich in der Praxis, aufgrund mangelnder Netzwerk- und Rechenkapazitäten, schwierig. Aufgrund der hochdynamischen Struktur und Architektur muss hierzu ein in kürzester Zeit erstellter Schnappschuss verwendet werden, um die Anzahl von Duplikaten zu verringern. Im Rahmen des Internet Census 2012 wurde ein solche Kartografierung des Internets mit Hilfe eines Botnetzes (Carna Botnet), auf welches im folgenden Abschnitt detailliert eingegangen wird, vorgenommen.

Die mittels Carna gewonnenen Daten bieten eine sehr interessante und umfangreiche Möglichkeit einen Schnappschuss des Internets zu analysieren. Aus Datenströmen, geografischen Standorten und vorhandenen Services können anschließend Aussagen über verschiedenste Parameter, wie eventuelle politische Einflussnahme oder die geografische Verteilung der weltweit genutzten Infrastruktur, getroffen werden.

Ein weiterer wichtiger Punkt der mittels Analyse solcher Daten erreicht werden kann, ist die Analyse des Internets auf Sicherheitsgefährdungen, um diese möglichst zeitnah erkennen oder vorhersagen zu können.

Diese Arbeit konzentriert sich auf zwei der durch den Internet Census 2012 erfassten Datensätze und gliedert sich wie folgt:

Einführend stellt Abschnitt 2 klar, was ein Internet Zensus ist, was genau der Internet Census 2012[1] für Daten erfasst hat und wie dieser durchgeführt wurde. Daraufhin folgt ein Überblick über die von uns zur Analyse verwendete Software sowie über während unserer Arbeit aufgetretene Probleme. In den Abschnitten 4 und 5 gehen wir genauer auf die Datensätze der Traceroutes und rDNS ein. Dabei stellen wir zuerst allgemeine Charakteristika der Daten vor und beschreiben dann unserer Auswertungen. In Abschnitt 6 widmen wir uns schließlich noch verwandten Arbeiten, bevor unsere Arbeit mit einer kurzen Zusammenfassung endet.

2. INTERNET CENSUS 2012

In diesem Abschnitt klären wir einerseits die Frage, was ein Internetzensus überhaupt ist und welche Gründe es dafür gibt einen solchen Zensus durchzuführen. Zum anderen geben wir anschließend eine kurze Einführung speziell in den Internen Census 2012 und erläutern sein Zustandekommen, sowie seinen Inhalt.

2.1 Was ist ein Internet Zensus

Allgemein bekannter ist der Begriff Zensus im Zusammenhang mit der Volkszählung. Eine Volkszählung dient dazu statistische Bevölkerungsdaten zu erheben (z.B. Alter, Einkommen, Beruf,...), um auf deren Grundlage beispielsweise Trends erkennen zu können denen gegengewirkt werden sollte. Ein Beispiel für einen solchen Trend wäre Altersarmut in

[1]Da wir den Internet Census 2012 als Eigennamen dieses Speziellen Zensus ansehen, verwenden wir Census um auf den Census 2012 zu verweisen und das deutsche Zensus, wenn wir allgemein über einen Zensus sprechen.

doi: 10.2313/NET-2014-03-1_03

bestimmten Regionen. Der Begriff Internet Zensus ist nun ganz ähnlich zu Verstehen, allerdings werden in einem solchen Zensus statt Bevölkerungsdaten Daten bezüglich des Internets erhoben (wie es der Name bereits andeutet). Gesammelt werden können dabei beispielsweise IP-Adresse, offene Ports, laufende Services und vieles weitere (siehe Abschnitt 2.4). Eine offizielle Definition des Begriffs „Internetzensus“ existiert allerdings noch nicht.

Analog zum Volkszensus lassen sich aus den Daten Trends Ablesen, welche beispielsweise für Sicherheitsfragestellungen relevant sein können. Ein Beispiel hierfür wäre eine Auswertung des Zensus in Bezug auf die Verbreitung veralteter Softwareversionen (z.B. Webserver) oder, wie im Internet Census 2012 zum Erstellen des Zensus genutzt, die Verbreitung der Verwendung von unsicheren Standard-Passwörtern.

2.2 Was ist der Internet Census 2012

Beim Internet Census 2012 handelt es sich um eine Zusammenführung der Daten aus mehreren IP basierten Scans über das gesamte Internet, welcher im Jahr 2012 durchgeführt wurde. Zur Durchführung dieser Scans wurde auf ein umfangreiches Botnetz (näheres hierzu im Abschnitt 2.3) zurückgegriffen. Der (anonyme) Urheber des Census stellte die Daten anschließend auf verschiedenen Plattformen [26] zur Verfügung. Der zur Verfügung gestellte Download umfasst gepackt rund 1,5 TB (entpackt 9TB) und kann beispielsweise via Bittorrent heruntergeladen werden. Er beinhaltet gepackte, Komma separierte, Klartextdateien, welche leicht aufbereitet und in verschiedene Datenbanksysteme importiert werden können. Einige der Daten können auch direkt im Browser betrachtet werden.

Der Autor nimmt in der gesamten Veröffentlichung keine Interpretation der Daten vor, sondern stellt nur das gesamte Projekt an sich vor. Der gesamte Census bezieht sich nur auf IPv4 Adressen, IPv6 wurde nicht untersucht. Aus diesem Grund werden wir, wenn wir im folgenden von IP Adressen reden, immer von IPv4 Adressen sprechen.

2.3 Carna Botnetz

Das Carna[2] Botnetz stellte die Grundlage zur Erzeugung des Internet Census 2012 dar. Aufgrund der rechtlich problematischen Situation der Datengewinnung bleibt der Autor der Studie anonym, stellt die gewonnenen Daten jedoch der Allgemeinheit zur Verfügung. Das verwendete Botnetz beruhte zum Zeitpunkt des Census auf rund 420.000 Geräten, auf die der Autor über Telnet Zugriff erhalten konnte. Grundlage hierfür waren nicht geänderte Loginnamen und Standardpasswörter, wie zum Beispiel Benutzername: „root“, Passwort: „root“. Die verwendete Software musste, um ein möglichst nicht invasives Netzwerk zu erhalten[3], nach jedem Neustart des Gerätes erneut aufgespielt werden, da sie sich nur im Hauptspeicher installierte. Aus diesem Grund schwankte die Gerätezahl während der gesamten Messung, da neugestartete Geräte nicht mehr verwendet werden konnten bis der Bot gegebenenfalls neu aufgespielt wurde. Die Aussagen bezüglich der möglichst geringen Invasivität durch den Autor konnten mangels Details von uns jedoch nicht weiter geprüft werden. Auch über die verwendeten Geräte werden keine weiteren Aussagen getroffen, diese werden in der Arbeit nur als ressourcenstarke und -schwache Geräte bezeichnet. Die ressourcenarmen Geräte wurden dabei als Endpunkte genutzt, die ressourcenstärkeren Geräte wurden zusätzlich auch zum Einsammeln der gewonnenen Daten der schwachen Geräte genutzt. Viele der durchgeführten Scans basierten auf Möglichkeiten des freien Open Source Tools Nmap [16], einem bekannten Portscanner.

[2] Göttin der römischen Mythologie.

[3] Vom Ersteller des Census wird in seinem Paper klargestellt, das er es als eines seiner Hauptziele ansah, keinen Schaden an übernommenen Geräten anzurichten.

2.4 Messungen

Im Laufe des Census wurde eine Reihe unterschiedlicher Daten gesammelt und schließlich zur Verfügung gestellt. Eine Übersicht über die Struktur der gewonnenen Daten kann in Tabelle 1 eingesehen werden. Die Inhalte werden im folgenden kurz beschrieben:

ICMP Echo Requests: ICMP Echo Requests (Pings) wurden mehrfach über verschieden lange Zeiträume an den gesamten IP Adressraum gesendet, die Antworten wurden gespeichert.

Reverse DNS: Das Domain Name System ordnet jeder Domain eine oder mehrere zugehörige IP Adressen zu und ermöglicht so die Umwandlung von Domain Namen in IP Adressen. Allerdings ist es in vielen Fällen auch möglich einer IP Adresse eine Domain zuzuordnen, dies nennt man „reverse DNS“. Im Internet Census wurde versucht zu möglichst vielen IP Adressen (rund 86% Abdeckung) die zugehörigen reverse DNS Einträge abzurufen. Zu jeder abgefragten IP Adresse wurden die resultierenden Domains oder ein Fehlercode abgespeichert.

Serviceprobes: Die Serviceprobes sind der größte Datensatz des Internet Census, hierfür wurden verschiedene Ports angefragt (geprobt) und deren Antwort gespeichert. Die Ergebnisse bieten die Möglichkeit zur Analyse welche IP Adressen welche Ports geöffnet haben und welche Services auf diesen laufen. Hieraus lassen sich beispielsweise Prognosen über zukünftige Angriffsziele erstellen, ein Beispiel ist der Telnet Port 23, welcher von Carna relativ einfach ausgenutzt werden konnte.

Hostprobes: Mittels Hostprobes wurde geprüft ob ein Host auf Anfragen reagiert. Im Gegensatz zum ICMP Ping Scan wurde bei den Hostprobes zusätzlich auch ein TCP SYN Paket an Port 443, ein TCP ACK Paket an Port 80 und eine ICMP Timestamp Anfrage gesendet. Dies bietet mehr Erkennungsmöglichkeiten ob ein Host online ist als ein einfacher Ping, da die Antwort auf den Ping möglicherweise unterdrückt wird, während auf eine der anderen Anfragen eine Antwort erfolgt.

Syncscans: Bei einem TCP SYN Scan[4] wird versucht eine Verbindung zu einem bestimmten Port aufzubauen. Sollte eine Antwort erfolgen, wird der Aufbau allerdings sofort abgebrochen und die Verbindung kommt nicht zustande. Durch die Antwort kann aber trotzdem auf den Zustand

[4] Der Autor des Internet Census spricht in seiner Arbeit immer von Syncscans, gemeint ist damit aber der bekannte SYN Scan.

doi: 10.2313/NET-2014-03-1_03

Typ	Felder	Anzahl	Größe
ICMP Ping	IP, Timestamp, Ergebnis	52 Mrd.	1,8 TB
Reverse DNS	IP, Timestamp, Ergebnis	10,5 Mrd.	366 GB
Serviceprobes	IP, Timestamp, Status, Ergebnis	180 Mrd.	5,5 TB
Hostprobes	IP, Timestamp, Status, Grund	19,5 Mrd.	771 GB
Syncscans	IP, Timestamp, State, Grund, Protokoll, Ports	2,8 Mrd.	435 GB
TCP IP Fingerprints	IP, Timestamp, Ergebnis	80 Mio.	50 GB
IP ID Sequence	IP, Timestamp, Ergebnis	75 Mio.	2,7GB
Traceroute	Timestamp, Quell IP, Ziel IP, Protokoll, Route	68 Mio.	18 GB

Tabelle 1: Übersicht über durchgeführte Messungen

des Ports geschlossen werden (geschlossen, offen oder gefiltert). In diesem Datensatz werden die Ports aufgelistet, die auf einen SYN Request reagiert haben, bzw. nicht reagiert haben. Durchgeführt wurde der Scan nur für eine Auswahl an bekannteren Ports (z.B. 23 Telnet, 80 http, 443 https) von erreichbaren bzw. antwortenden IP Adressen.

TCP IP Fingerprints: Für einige IP Adressen war es möglich einen TCP/IP Fingerabdruck zu ermitteln. Mit diesem Fingerabdruck ist es unter Umständen möglich detaillierte Eigenschaften, wie Hersteller, Betriebssystem o.Ä. des jeweiligen Geräts zu ermitteln. Grundlage hierfür ist die jeweils eigene Implementierung des TCP/IP Protokollstapels in verschiedenen Betriebssystemen, wodurch sich bestimmte Felder im TCP oder IP Header je nach Betriebssystem unterscheiden. Diese Fingerabdrücke sind allerdings nicht zwangsläufig korrekt, da sich die Felder auch manuell konfigurieren lassen.

IP ID Sequence: Analyse der von den Hosts genutzten Strategien zur Erzeugung der Identifikationsnummern innerhalb des IP Headers.

Traceroutes: In diesem Datensatz sind die Ergebnisse der durchgeführten Traceroutes abgelegt, unter Angabe der einzelnen Hops und deren Laufzeiten. Es ist jedoch keine Information vorhanden ob das Ziel der Traceroutes erreicht wurde.

3. TECHNIK

Bedingt durch die sehr großen Datenmengen und Aufbau der Daten ist es nur schwer möglich die Verarbeitung allein mit potenter Hardware zu ermöglichen, auch die verwendete Datenbanksoftware muss für entsprechend große Datenmengen konstruiert sein. Bedingt durch die schmalen, aber sehr hohen Tabellen (Teils mehreren Milliarden Zeilen bei nur maximal sechs Spalten) bot sich ein auf spaltenweise Verarbeitung spezialisiertes Datenbankmanagementsystem (DBMS) an.

Hierfür standen uns mehrere, erprobte Varianten (MonetDB [15], Greenplum [17][5]) zur Verfügung. Eine weitere Möglichkeit war die Verwendung des zeilenbasierten Datenbanksystems PostgreSQL [18]. MonetDB ist im Gegensatz zu Greenplum eine Open Source Lösung, weswegen wir diesem den Vorzug gaben. PostgreSQL ist zwar ebenfalls Open Source, wurde von uns aber aus Performancegründen hinten angestellt, da einige durchgeführte Benchmarks erhebliche zeitliche Vorteile bei der Auswertung von Anfragen durch MonetDB aufzeigten. Für diesen Test verwendeten wir einen kleinen Teil (1 Millionen Einträge) des in Abschnitt 4 vorgestellten Traceroute Datensatzes. Tabelle 2 zeigt einige Anfragen, sowie die von Monetdb und PostgreSQL benötigte Zeit für deren Abarbeitung[6].

Die Verhältnisse schwankten zwar je nach Anfrage relativ stark, allerdings hatte MonetDB immer einen relativ großen Vorsprung vor PostgreSQL. Bei den Werten gilt zu beachten, dass sich die von PostgreSQL ohne Optimierung durch Indize ergaben, welche MonetDB bei Bedarf von alleine anlegt. Allerdings brachte ein Test mit Indizen nur Geschwindigkeitszunahmen von ungefähr 10%, was immer noch wesentlich langsamer war als MonetDB. Unsere Wahl des DBMS fiel schließlich auf MonetDB, für den Falle des Scheiterns, bedingt durch den Beta Status von MonetDB, wurde PostgreSQL allerdings als Fallback Lösung berücksichtigt. Als Hardware stand uns ein Server mit 24 Kernen und 144 GB Arbeitsspeicher zur Verfügung.

3.1 MonetDB

MonetDB ist ein Open Source Database Managment System, welches speziell auf große Datenmengen und komplexe Querys optimiert wurde. Im Gegensatz zu bekannteren zeilenbasierten DBMS, handelt es sich bei MonetDB um eine spaltenoptimierte Datenbanksoftware. Ein weiteres, sehr wichtiges und performancerelevantes Feature ist die Speziaisierung durch optimale Ausnutzung von CPU Caches. Weiterhin arbeitet MonetDB RAM zentrisch, Änderungen werden zuerst im RAM abgelegt und dann über ein Log File später erst in die Datenbank persistiert. Der größte Nachteil von MonetDB ist der derzeitige Beta Status, die mangelhafte Dokumentation und die noch nicht vollständig unterstützte SQL Syntax.

3.1.1 Probleme

Aufgrund der großen Datenmengen von mehreren Terabyte in Verbindung mit mehreren Milliarden Zeilen kam während der Auswertung auch die leistungsfähige Datenbank MonetDB an die Grenzen der Technik bzw. des Arbeitsspeichers. Bedingt durch die RAM zentrierte Arbeitsweise von MonetDB werden bei Querys die Daten initial in den Arbeitsspeicher geladen, um anschließend ausgewertet zu werden. Bei kleineren Datensätzen stellt dies kein Problem dar.

[5]Greenplum bietet sowohl Zeilen- als auch Spaltenorientierte Datenverarbeitung.

[6]Jede Anfrage wurde von uns mehrfach ausgeführt und eine Durchschnittszeit ermittelt.

doi: 10.2313/NET-2014-03-1_03

Anfrage	Zeit MonetDB	Zeit PostgreSQL	Verhältnis Postgres/MonetDB
select count(*) from data	0,305ms	253ms	829
select count(distinct targetIP) from data	1,1s	7s	6,4
select targetIP, count(distinct targetIP) from data group by targetIP	0,7s	11,2s	16

Tabelle 2: Performanzvergleich verschiedener Anfragen an MonetDB und PostgreSQL

Bei großen Tabellen, wie der RDNS Datensatz aus dem Internet Census waren die vorhanden 144GB RAM jedoch nicht mehr ausreichend um eine gesamte Spalte in den Arbeitsspeicher zu laden. Dies führte zu einer enorm hohen IO Load des Systems. Abhilfe konnte nur durch einen Abbruch des Querys geschaffen werden, dies gestaltete sich als initial schwierig, da das Abbrechen von Querys in MonetDB zum Zeitpunkt der Durchführung nur unzureichend dokumentiert war.

Ähnliche Probleme traten auch beim Einspielen der RDNS Einträge auf. Hier werden von MonetDB die Einträge erst im RAM zwischengespeichert um dann über das Log File persistiert zu werden. Dies führte ebenfalls zu einer permanent sehr hohen IO Load des Systems und konnte von uns nur durch einen Neustart des gesamtes DBMS behoben werden. Weitere gravierende Probleme gibt es bei der Vergabe eines Primärschlüssels während des Imports von großen Datensätzen (bereits im 8 stelligem Bereich). Mit zunehmender Anzahl von importierten Tupeln kann ein immer stärkerer Einbruch der Performance bemerkt werden. Die vermutete Ursache für dieses Verhalten liegt in der automatischen Erzeugung von Indizies. Dieser Einbruch konnte dabei sowohl unter Verwendung einzelner Transaktionen, als auch der Verwendung einer Gesamttransaktion festgestellt werden. Für einen fehlerfreien Import musste später die Datenbank in den Wartungsmodus geschalten werden.

Darüber hinaus gab es während der Entwicklung große Problem mit dem Datentyp für IP-Adressen (INET), dessen Performance bricht unter nicht reproduzierbaren Bedingungen sehr stark ein, sodass eine Nutzung für weiterführende Auswertungen nicht möglich ist. Auch war die MonetDB SQL Python API nicht in der Lage Anfragen nach Daten des Typs INET zu verarbeiten und erzeugte einen Fehler, wenn der INET Typ nicht zuerst auf VARCHAR konvertiert wurde. Es lässt sich jedoch abschließend nicht sagen, ob die aufgetretenen Probleme wirklich alle Probleme des DBMS sind, oder ob sie auf die noch nicht ausgereifte (Python) API zurückzuführen waren.

3.1.2 Fazit

Wenngleich es bei der Anwendung von MonetDB zu größeren Problemen und Unannehmlichkeiten kam, überwogen die Vorteile während der gesamten Auswertung deutlich, vor allem da wir in der Lage waren, die meißten Probleme zu lösen oder zu umgehen. Bereits bei kleineren Datensätzen und selektiven Abfragen konnte MonetDB, bedingt durch die spaltenbasierte Arbeitsweise, extreme Geschwindigkeitsvorteile erzielen. Dieses Verhalten wurde in Tabelle 2 für einen kleinen Datensatz anhand PostgreSQL und MonetDB dargestellt. Die Performance von MonetDB überstieg dabei die Performance von PostgreSQL um Größenordnungen. Wichtig ist dabei Abfragen gezielt auf einzelne Spalten zu beschränken, bei Abfragen auf die gesamte Zeile relativiert sich der Geschwindigkeitsvorteil. Aufgrund der spaltenbasierten Architektur ist MonetDB nur eine praktikable Lösung für spezielle Daten. Für quantitativ kleinere und weniger spezifische Datenmengen stehen hingegen besser geeignete Generalisten zur Verfügung, die dann unter anderem eine deutlich weiter entwickelte Implementierung des SQL Syntax zur Verfügung stellen.

3.2 Geographische Zuordnung von IP-Adressen

Um eine Lokalisierung der verwendeten IP-Adressen durchzuführen gibt es eine Reihe verschiedener Mögchkeiten. Beispielsweise kann die Lokalisierung einer IP-Adresse anhand des Whois Eintrags ihres Autonomen Systems (AS) bestimmt werden. Dabei tritt jedoch das Problem auf, dass so nur das Land in dem das AS registriert ist bestimmt werden kann, jedoch können einzelne Adressen dem AS zugeordnete auch in anderen Ländern liegen. Um eine genauere Geographische Zuordnung einzelner IP-Adressen durchführen zu können, bieten sich Dienste wie MaxMind [13] an, die Datenbanken mit geolokations Informationen einzelner IP-Adressen zur Verfügung stellen. Im Rahmen der Arbeit wurde dabei auf die kostenfrei erhältlichen GeoLite Datenbanken zurückgegriffen. Um IP-Adressen mit diesen Datenbanken abzugleichen steht eine Reihe verschiedener Wege zur Verfügung. Von uns wurden die MaxMind Datenbanken dabei mittels der Python Bibliothek `pygeoip` [19] verwendet, um IP-Adressen auf die zugehörigen Ländercodes abzubilden.

3.2.1 Probleme

Wenngleich die Genauigkeit der Datenbanken von MaxMind selbst mit 99,8 % [12] angegeben wird, gestaltet sich eine Evaluation dieser Angabe im Rahmen der Arbeit als nicht machbar und muss als Grundwahrheit akzeptiert werden. Davon abgesehen gibt es auch bei einer Fehlerquote von 0.02% bereits rund 86 Millionen fehlerhafte Zuordnungen. Laut Maxmind erhöht sich die Unschärfe bei den Städtedaten monatlich um 1.5%. Es stehen jedoch keine älteren Versionen der Datenbanken zum Download bereit. Somit sind zum Zeitpunkt des Census deutlich erhöhte Fehlerquoten im Bereich zwischen 14% (Messungen Dezember 2012) und 26% (Messungen Mai 2012) möglich. Die so entstehende Fehlerquote wird dabei jedoch nur für Städte angegeben, eine Angabe auf Länderebene fehlt dabei. Der länderübergreifende Fehler wird sich, bedingt durch die deutlich gröbere Auflösung, vermutlich deutlich unter 1,5% befinden. Unter der Hypothese eines Worst-Case Szenarios muss dabei von oben genannter Fehlerquote von 14-26% ausgegangen werden. Seitens MaxMind werden jedoch keine weiteren Details bekannt gegeben, weder wie eine Lokalisierung der Adressen vorgenommen wird, noch wie deren Genauigkeit ermittelt

wird.

3.2.2 Fazit

Prinzipiell stellt die Lokalisierung mittels der GeoIP Datenbanken im Vergleich zur Lokalisierung mittels AS ein genaueres, aber auch unabwägbareres Verfahren zur Verfügung. Bei der Lokalisierung mittels der autonomen Systeme ist eine Fehlerquote bereits prinzipbedingt gegeben. Die Lokalisierung mittels der verfügbaren GeoIP Datenbanken bietet im Gegensatz dazu eine Genauigkeit von bis zu theoretischen 98,2%. Faktisch gesehen ist diese Fehlerquote im hier untersuchten Anwendungsfall jedoch viel zu gering. Eine genauere Abschätzung des Fehlers ist, aufgrund vorher genannter Gründe, nicht möglich.

3.3 Zerlegung von Domainnamen

Die Zerlegung von Domainnamen stellt ein Problem dar, da Top Level Domains oft nicht eindeutig zu identifizieren sind. So ist es nicht trivial aus den Domains „www.test.co.uk“ bzw. „www.test.com“ jeweils den Hostnamen zu identifizieren, da nicht bekannt ist welcher Teil der Domain der Hostname ist. Zum Extrahieren der verschiedenen Domainbestandteile gibt es aber die Möglichkeit auf bekannte Suffix Listen zurückzugreifen. In der hier vorliegenden Arbeit wurde die Python Bibliothek `tldextract` [10] von John Kurkowski zurückgegriffen.

3.3.1 Probleme

Auch wenn ein Großteil der Zerlegungen durch `tldextract` korrekt sind, ist bei Datensätzen mit mehr als 1 Milliarde Zeilen bereits eine Fehlerquote von 1% zu hoch. Bei einfacheren Adressen funktioniert `tldextract` einwandfrei, steigt die Anzahl der Ebenen innerhalb einer Adresse jedoch an, so steigt die Fehlerrate stark an. Im rDNS Datensatze existieren stellenweise auch Domains mit einer Länge von mehr als Hundert Zeichen, bei diesen ist ebenfalls eine sehr hohe Fehlerquote zu verzeichnen.

3.3.2 Fazit

`Tldextract` stellt eine gute Variante da um eine Vorverarbeitung von Adressen zu ermöglichen, aufgrund der hohen Anzahl von Adressen und der hohen Fehlerquote bei komplexeren Adressgebilden sind dennoch umfangreiche händische Nacharbeiten nötig um gut weiterverarbeite Resultate zu erhalten. Je nach gewünschten Ergebnisse kann hier jedoch auf weitere Hilfsmittel, wie beispielsweise Excel zurückgegriffen werden, um eine weitere Aufarbeitung der Daten vorzunehmen.

4. TRACEROUTES

Der erste Datensatz, den wir genauer untersuchen wollen, sind die Traceroutes. Traceroute ist ein Programm, welches es Nutzern ermöglicht den Weg nachzuvollziehen, den ein von dem Quellrechner gesendetes Paket auf dem Weg zu seinem Ziel nimmt. Dazu sendet der Host auf dem Traceroute ausgeführt wird zuerst drei Pakete[7] mit einer Time to Live (TTL) von eins aus, wodurch bei dem ersten Router auf der Strecke zum Ziel die TTL auf null dekrementiert wird und eine ICMP Fehler Nachricht[8] an den Traceroute Host gesendet wird. Traceroute sendet nun drei Pakete mit einer TTL von 2 aus, um so die Antworten des zweiten Routers auf der Strecke zum Ziel zu erhalten, aus denen er jeweils dessen IP-Adresse auslesen kann. Dieses Verfahren wird nun solange wiederholt bis die TTL der Pakete groß genug ist, um ihr Ziel zu erreichen. Der Zielrechner antwortet mit einer ICMP Nachricht [9] aus der erkennbar ist, dass das Ziel erreicht wurde. Die Rückgabe von Traceroute besteht aus allen IP-Adressen die sich auf der Strecke befinden und den zugehörigen Übertragungszeiten.

Nachdem wir einleitend allgemeine Eigenschaften und Probleme der Traceroute Daten betrachten, gehen wir dazu über, die Traceroutes auf IP-, AS- und Länderebene zu analysieren. So betrachten wir bei den IP-Traceroutes die Anzahl von genutzten Geräten, sowie die Ziele der Traceroutes. Desweiteren wollen wir herausfinden, welcher Anteil der IP-Adressen in den Traceroutes unbekannt sind, sowie ihre durchschnittliche Hoplänge betrachten. Bezüglich der Autonomen Systeme beschränken wir uns auf eine Erklärung der Differenz in der Anzahl AS-Nummern auf der Route selbst und als Ziel. Als Schwerpunkt dieses Teils der Arbeit betrachten wir schließlich die Traceroutes auf Länderebene und, unter anderem, klären, welche Länder weltweit am wichtigsten sind für die Internetinfrastruktur, wieviele Länder Traceroutes durchschnittlich auf dem Weg zu ihrem Ziel durchquert haben und ob sich interessante Routenverläufe erkennen lassen.

4.1 Eckdaten

Der Urheber des Internet Census erstellte innerhalb von 23 Tagen insgesamt 68.7 Millionen dieser Traceroutes. Als Quellrechner, von denen aus die Traceroutes gestartet wurden, dienten ihm hierzu rund 275.000 seperate Geräte[10], auf die er mittels Telnet zugreifen konnte, welche allerdings über begrenzte Ressourcen verfügten. So war es auf den so erreichbaren Geräten zwar möglich, das bereits installierte Traceroute Programm auszuführen, allerdings konnte nicht der eigentliche Carna Bot hochgeladen und ausgeführt werden, da Beispielsweise der Speicher nicht ausreichend war. Der Autor modifizierte Carna deswegen so, dass er sich über Telnet auf diese Geräte einloggte, die Traceroutes bei aufrechterhaltener Verbindung durchführte und das Ergebnis auf dem Carna Gerät in komprimierter Form speicherte.

Als Ziele für die Traceroutes dienten ihm dabei IP-Adressen, die bereits einmal auf seinen Telnet-Scanner angesprochen hatten und somit sehr wahrscheinlich noch in Verwendung waren. Insgesamt wurden Traceroutes zu rund 64 Millionen verschiedenen IP-Adressen gesendet. Die Ergebnisse der Traceroutes wurden, wie in Tabelle 1 bereits vorgestellt, abgespeichert.

Die Route setzte sich dabei aus einer Folge von Hop, IP-Adresse und Übertragungszeiten zusammen. Tabelle 3 zeigt einige Beispiele für die Daten. Aus Ressourcengründen mussten wir darauf verzichten, die Traceroutes getrennt nach

[7] Default Einstellung

[8] Typ 11 Time exceeded, Code 0 Time to live exceeded in Transit

[9] Typ 0, Code 0 Echo Reply bzw. bei UDP-basiertem Traceroute mit Typ 3 Destination Unreachable, Code 3 Port Unreachable

[10] Diese Geräte sind nicht Teil der 420.000 Geräte auf denen Carna lief.

doi: 10.2313/NET-2014-03-1_03

Timestamp	Quell-IP	Ziel-IP	Protokoll	Traceroute
1340158500	21.196.75.45	112.172.26.154	ICMP	
1340109900	201.200.74.1	125.99.70.79	ICMP	1:10.86.202.1:30ms,40ms,*;2:10.86.202.1:2540ms,2540ms,*;
1340158500	201.196.75.45	211.104.66.50	ICMP	1:10.39.155.45:40ms,30ms,50ms;2::*,*,*;3::*,*,*;

Tabelle 3: Beispiele für Einträge im Traceroutes Datensatz des Internet Census 2012

verwendetem Protokoll[11] zu untersuchen. Unsere folgenden Auswertungen beziehen sich nur auf Quelle, Ziel und den IP-Adressen auf dem Weg zwischen diesen.

4.2 Mapping auf Autonome Systeme und Länder

Um nicht nur die IP-Adressen der Traceroutes zu betrachten, erzeugten wir zwei weitere Datensätze. Als erstes übersetzten wir die IP-Adressen mittels des Pythonscripts `pyasn` [4] in die Nummern der Autonomen Systeme (AS), denen sie zugeordnet waren. Als Grundlage hierzu dienten uns drei Routing Information Base (RIB) Dateien. Diese spiegeln den Zustand von BGP Routern an 3 Zeitpunkten während des Erstellens der Traceroutes wieder. Dadurch dass wir zum Übersetzen der IP-Adressen in AS-Nummern jeweils die RIB Datei verwendet haben welche dem Timestamp der jeweiligen Traceroute am nächsten war, wollten wir eine möglichst große Genauigkeit gewährleisten.

Als zweites erstellten wir eine Übersetzung der IP-Adressen in Ländercodes, um mit den geographischen Standorten der IP-Adressen arbeiten zu können. Für diese Übersetzung nutzten wir das in Abschnitt 3.2 vorgestellte Verfahren mittels der Pure Python GeoIP API (`pygeoip`) und dem GeoLite Country Datensatz von Maxmind. Tabelle 4 zeigt einen kurzen Überblick darüber, wie viele verschiedene IP-Adressen, AS-Nummern und Ländercodes jeweils als Quelle, Ziel oder innerhalb der Traceroutes existieren, beziehungsweise wie viele insgesamt Vorkommen.

Ort	IPv4 Adressen	Autonome Systeme	Länder
Quelle	275.000	535	100
Route	1.900.000	12.921	199
Ziel	64.700.000	41.332	243
Gesamt	65.800.000	41.335	243

Tabelle 4: Überblick über die Anzahl der verschiedenen IP-Adressen, Autonomen Systeme und Ländercodes, gegliedert nach ihrem Vorkommen

Der enorme Unterschied zwischen der Anzahl der verschiedenen IP-Adressen, AS-Nummern und Ländercodes besonders bei den Zielen und der Route sticht hier besonders hervor. Wir waren in der Lage einige Erklärungen dafür zu finden und stellen diese unter anderem in den passenden Abschnitten 4.4 bis 4.6 vor. Dabei ist natürlich zu beachten, dass sich die Ursachen für die Abweichung auf IP-Ebene auch auf AS- und Länder-Ebene auswirken.

4.3 Probleme des Traceroute Datensatzes

[11]Die Traceroutes wurden per ICMP oder UDP erstellt.

Der Traceroutes Datensatz weist einige Probleme auf, welche die Interpretation der Daten wesentlich erschweren oder gar verhindern. Der Autor gibt so beispielsweise nur darüber Auskunft, welche Daten er gesammelt hat und in welchem Format sie gespeichert wurden. Genauere Details darüber, nach welchen Prinzipien beim Erstellen und Messen der Traceroutes vorgegangen wurde, nennt er allerdings nicht. Dies äußert sich beispielsweise darin, dass Traceroutes mit Quelle und Ziel gespeichert wurden, ohne sicherzustellen, ob sie ihr Ziel überhaupt erreicht hatten. Diese Vermutung wird durch Traceroutes untermauert, welche zwar mit Quelle, Ziel und Route abgespeichert wurden, bei denen aber eindeutig das Ziel nicht erreicht wurde. Erkennen konnten wir das durch die Tatsache, dass die von der entsprechende Traceroute gesendeten Pakete immer wieder zwischen den gleichen zwei verschiedenen IP-Adressen hin und her sprangen und somit zum Zeitpunkt der Messung in eine Routingschleife gerieten.

Eine weitere Auffälligkeit, für die wir ohne genauere Angaben seitens des Autors keine Erklärung finden konnten, ist der Fakt, dass manche Traceroutes mit ihrer Nummerierung nicht mit eins beginnen. Hier war es uns unmöglich zu sagen, welche Gründe das hat. Angesichts dieser Probleme sind die gewonnen Ergebnisse mit Vorsicht zu betrachten und eher als eine Art Trend zu verstehen.

4.4 Auswertung der IP-Adressen

Bevor wir uns der Analyse der Daten auf AS- und Länderebene widmen, betrachten wir einige Auswertungen auf IP-Ebene bezüglich fehlender Adressen und der Länge der Traceroutes.

4.4.1 Allgemeines

In seinem Paper zum Internet Census 2012 [26] nennt der Autor als Fazit seiner Arbeit eine Zahl von ungefähr 1.3 Milliarden genutzten IP-Adressen. Legt man diese als Basis zugrunde, so wurden als Quelle der Traceroutes 0,02% des Internets betrachtet, als Ziel immerhin rund 4,97%. Auf den Routen selbst lagen 0,15% der genutzten IP-Adressen, wenn man die unbekannten IP-Adressen nicht mitzählt, die im folgenden Abschnitt kurz erläutert werden.

4.4.2 Fehlende IP-Adressen auf der Route

Einer der Gründe warum, im Vergleich zu den Ziel-IPs, so viel weniger verschiedene IP-Adressen auf der Route selbst liegen, liegt in der Funktionsweise von Traceroute begründet. So ist es jedem Administrator eines Geräts möglich zu Unterbinden, dass dieses ICMP Fehlermeldungen aussendet, welche Traceroute verwendet um die IP-Adresse des Geräts auf dem Pfad zu bestimmen (siehe Abschnitt 4). In solchen Fällen bleibt die IP des entsprechenden Hops unbekannt, da Traceroute keine Rückmeldung erhält. Tabelle 3 zeigt in Beispiel drei bei den Hops 2 und 3 ein solches Verhalten, die IP-Adressen sind unbekannt und daher leer. Es war

doi: 10.2313/NET-2014-03-1_03

uns unmöglich, diese fehlenden IP-Adressen in die Anzahl der verschiedenen Geräte die sich auf den Routen befinden einzubeziehen. Durchschnittlich waren im Traceroute Datensatz 15,2% der IP-Adressen in den Routen[12] unbekannt, was jeder 6.-7. Adresse entspricht. Aussagen darüber, wieviele verschiedene reale IP-Adressen sich dahinter verstecken lassen sich leider nicht treffen. Bei den Zieladressen trat das Problem der unbekannten IPs nicht auf, da der Autor hier anscheinend einfach die Ziel-IP angegeben hat, ohne sicherzustellen, dass die Traceroute sie auch erreicht hat.

4.4.3 Hops

Schließlich betrachteten wir noch die Anzahl der Hops, die die Traceroutes benötigten, um von der Quelle zu ihrem Ziel zu kommen. Mit einem Hop ist dabei die Anzahl der IP-Adressen gemeint, welche zwischen Quelle und Ziel liegen, inklusive dem Ziel. Der erste Eintrag in dem Traceroute Beispiel in Tabelle 3 hätte dementsprechend einen Hop, der zweite drei. Abbildung 1 zeigt wie viele Einträge im Traceroute Datensatz jeweils wieviele Hops benötigten. Die Anzahl Traceroutes ist hierbei in logarithmischer Form angetragen.

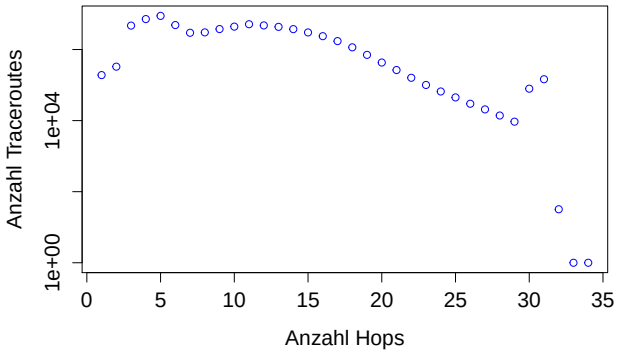

Abbildung 1: Anzahl der Traceroutes nach der Anzahl der Hops von Quelle bis zum Ziel

Aufgefallen ist uns bei diesem Bild vor allem der plötzliche Anstieg an Traceroutes mit 31 beziehungsweise 32 Hops, nachdem bis zu diesem Punkt ein abnehmender Trend in der Anzahl Traceroutes erkennbar war. Unsere Vermutung war, dass sich bei den größeren Hopzahlen vor allem Traceroutes befinden, welche in einer Routingschleife gefangen waren und somit immer zwischen den selben IP-Adressen hin und hersprangen, bis die Time to Live der Pakete 0 erreichte. Normalerweise beträgt die maximale TTL von Paketen (Default Einstellung) die Traceroute aussendet 30, was inklusive Ziel einem Maximum von 31 Hops entspricht. Warum es auch so viele Traceroutes mit 32 Hops (plus jeweils eine mit 33 und 34 Hops) gibt können wir nicht eindeutig erklären, da uns dazu Angaben des Autors des Census fehlen. Möglicherweise liegt der Grund aber in unterschiedlichen Implementierungen von Traceroute auf den übernommenen Geräten. Abbildung 2 zeigt, wie viel Prozent der Traceroutes mit einer bestimmten Hoplänge Routen aufweisen, auf denen die selbe IP-Adresse mehrfach vorkommt.

[12]Die Routen beinhalten nicht Start- und Ziel-IP einer Traceroute.

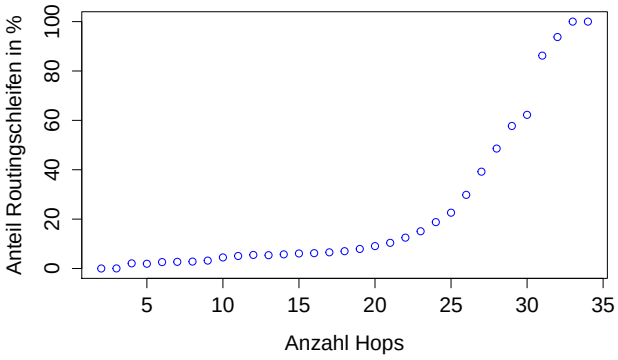

Abbildung 2: Anteil der Traceroutes mit Routingschleifen an gesamten Traceroutes nach Anzahl der Hops von der Quelle bis zum Ziel

Es ist gut zu sehen, dass mit steigender Hopzahl die Wahrscheinlichkeit zunimmt, dass ein Paket in eine Routingschleife gerät. Ab einer Länge von 21 Hops gerieten über 10% der Traceroutes in eine Routingschleife, ab 28 Hops sogar mindestens die Hälfte. Hierbei ist zu beachten, dass wir nur Routingschleifen feststellen konnten, bei denen zumindest eines der an der Schleife beteiligten Geräte auch eine ICMP Fehlermeldung bei ausgelaufener TTL zurück an die Quelle der Traceroute gesendet hat. Sollten alle an einer Schleife beteiligten Geräte ihre Fehlermeldung unterdrückt haben (siehe Abschnitt 4.4.2), konnten wir die Schleife nicht erkennen. Die Prozentwerte der Grafik sind folglich als Mindestwerte zu verstehen. Nutzt man zur Berechnung des Mittelwerts nur jeweils den Anteil an Traceroutes, der keine Schleife enthält, ergibt sich eine durchschnittliche Hopzahl von 9,17 Hops pro Traceroute.

Da eine Routingschleife ein starkes Indiz dafür ist, dass ein gesendetes Paket sein Ziel nicht erreicht hat, lassen sich anhand der Schleifenwahrscheinlichkeiten aus Abbildung 2 Rückschlüsse auf die Traceroutes ziehen, bei denen das im Datensatz angegebene Ziel nicht erreicht wurde. Wir schließen daraus, dass die Aussagekraft der Traceroutes mit steigender Anzahl Hops immer stärker abnimmt, da der Anteil der Routingschleifen für hohe Hopwerte stark ansteigt.

4.5 Auswertung der Autonomen Systeme

Aus Ressourcengründen konnten wir die Autonomen Systeme nicht genauso tiefgehend untersuchen wie die Länder, durch die die Traceroutes verliefen. Allerdings wollten wir zumindest eine Erklärung für die Differenz zwischen der Anzahl AS auf der Route selbst ($\approx$ 13.000) und der Anzahl AS als Ziel der Traceroutes ($\approx$ 41.000) finden.

Aus unserer Sicht deutet diese Differenz darauf hin, dass viele der betroffenen Autonomen Systeme sogenannte Stub- oder Multihomed-AS sind[28]. Das bedeutet, dass diese autonomen Systeme Endpunkte sind, die keinen Verkehr von externen AS an andere externe AS weiterleiten (Transit-AS

doi: 10.2313/NET-2014-03-1_03

). Diese können somit nur als Ziel- oder Quell-AS einer Traceroute vorkommen, nicht aber auf der Route selbst. Die Autonomen Systeme die auf der Route selbst vorkommen sind daher auf jedenfall sogenannte Transit Systeme, die Verkehr von und an andere AS weiterleiten. Ein Beispiel dafür sind Internet Service Provider, wie zum Beispiel die Deutsche Telekom oder AT&T. Der Anteil von Transit-AS an den gesamten AS beträgt somit mindestens 31%. Mindestens deswegen, da wir nicht feststellen können, wie viele der Ziel- und Quell-ASe der Traceroutes wirklich keinen Verkehr weiterleiten.

4.6 Auswertungen auf Länderebene

Von besonderem Interesse für uns waren Auswertungen auf Länderebene. Auf dieser Ebene lassen sich, unter anderem, Schlüsse über die Wege ziehen die Pakete nehmen um von einem Land in ein anderes zu gelangen. Zu beachten gilt hier, dass wen wir von Ländern sprechen alles gemeint ist, was einen Ländercode besitzt. Der Großteil sind durchaus richtige Länder wie Deutschland oder die USA, allerdings haben beispielsweise auch Kontinente Ländercodes.

4.6.1 Herkunft und Ziele

Zur Einführung betrachten wir die Herkunft und die Ziele der Traceroutes des Internet Census 2012. Abbildung 3 veranschaulicht die Länder, aus denen die Traceroutes gestartet wurden.

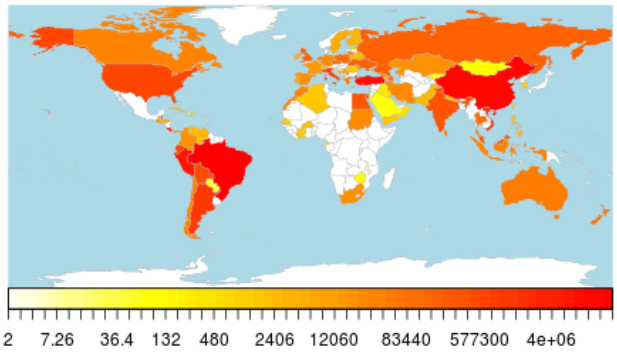

Abbildung 3: Anzahl der Traceroutes die von einem bestimmten Land aus gestartet wurden

Insgesamt wurden für das Erstellen der Traceroutes Geräte aus 100 Ländern genutzt. Tabelle 5 macht hierzu deutlich, dass knapp 50% der im Rahmen des Census erstellten Traceroutes allein aus Brasilien und Costa Rica kommen, beides südamerikanische Länder. Zusammengerechnet sind die Top 5 Länder sogar für knapp 86% der gestarteten Traceroutes verantwortlich.

Land	Anzahl	Anteil Gesamt
Brasilien	20.000.000	29,1%
Costa Rica	13.900.000	20,1%
China	10.600.00	15,4%
Türkei	8.500.00	12,4%
Peru	6.100.00	8,9%

Tabelle 5: Top 5 der Quellenländer für Traceroutes

Abbildung 4 zeigt die Zielländer der Traceroutes und analog dazu Tabelle 6 die Top 5.

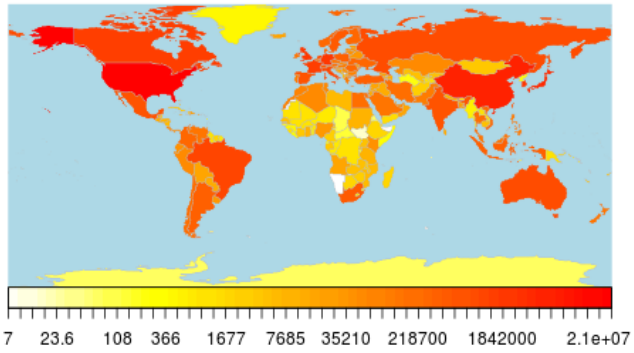

Abbildung 4: Anzahl der Traceroutes die an ein bestimmtes Land gesendet wurden

Land	Anzahl	Anteil Gesamt
USA	28.500.000	40,1%
China	6.200.000	9%
Japan	3.800.000	5,5%
unbekannt	3.500.000	5,1%
Großbritannien	2.300.000	3,3%

Tabelle 6: Top 5 der Zielländer für Traceroutes

Hier sticht einerseits der hohe Anteil an Zielen in die USA ins Auge, andererseits die „unbekannten“ Ziele. Mit unbekannten Zielen beziehen wir uns auf IP-Adressen in den Zielen der Traceroutes, welche mittels der Maxmind GeoIP Daten nicht einem Land zugeordnet werden konnten. Auf die vielen Ziele in den USA werden wir im Folgeabschnitt kurz eingehen.

4.6.2 Anteil betroffener IP-Adressen pro Land

Wesentlich interessanter als die Anzahl der gesamten gestarteten und ankommenden Traceroutes von, beziehungsweise in ein Land erschien uns die Frage, welcher Anteil von verschiedenen IP-Adressen pro Land von den Traceroutes genutzt wurden. Hierzu betrachteten wir zuerst jede IP-Adresse nur einmalig, unabhängig wie oft sie tatsächlich als Quelle oder Ziel einer Route vorkam. Die daraus gewonnene Anzahl betroffener IP-Adressen pro Land normalisierten wir dann mithilfe der Anzahl der dem Land zugeordneten IP-Adressen. Als Quelle der jedem Land zugeordneten IP-Adressen verwendeten wir wieder einen von Maxmind zur Verfügung gestellten Datensatz[14]. Abbildung 5 zeigt dazu, wieviel Prozent der IP-Adressen pro Land als Quelle für Traceroutes dienten.

Dies ist insbesondere deshalb interessant, da der Autor des Census die Traceroutes auf Geräten ausgeführt hat, die einerseits über relativ beschränkte Ressourcen verfügten und andererseits durch einen einfachen Telnet Login benutzt werden konnten. Die Abbildung gibt somit nicht nur einen Überblick über die betroffenen Geräte, sondern zeigt darüber hinaus den Anteil an leicht angreifbaren Geräten[13]. Hierbei ist

[13]Wie bereits erwähnt verwendete der Autor nicht alle angreifbaren Geräte, die wahren Zahlen dürften also noch größer ausfallen.

doi: 10.2313/NET-2014-03-1_03

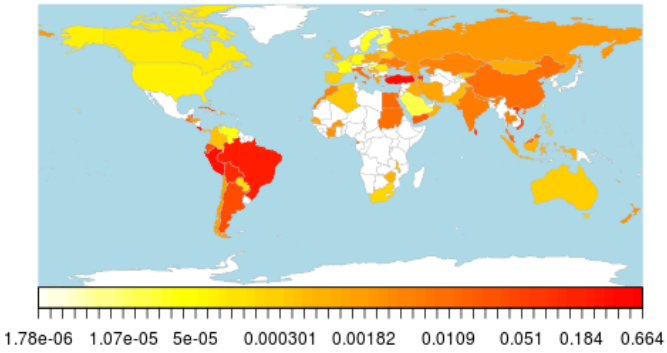

Abbildung 5: Anteil der pro Land betroffenen verschiedenen IP-Adressen als Quelle gegenüber den dem Land zugeordneten IP-Adressen in Prozent

zu beachten, dass weiße Länder auf der Karte nicht zwangsläufig als „sicher“ zu betrachten sind sondern als unbekannt, da aus diesen keine Geräte übernommen wurden. So scheint vor allem Südamerika im Vergleich zum Rest der Welt relativ viele unsichere Geräte zu verwenden, gefolgt von Asien. Europa und Nordamerika hingegen scheinen in dieser Hinsicht vergleichsweise sicher zu sein.

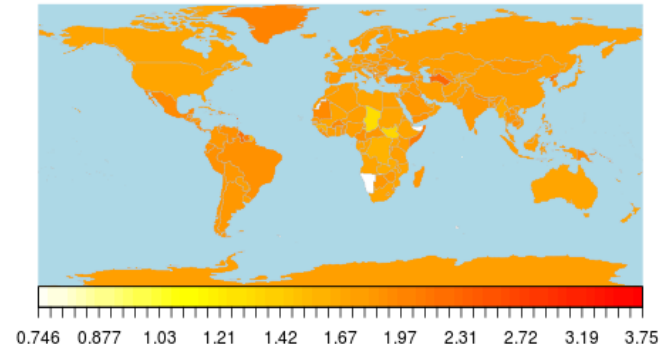

Abbildung 6: Anteil der pro Land betroffenen verschiedenen IP-Adressen als Ziel gegenüber den dem Land zugeordneten IP-Adressen in Prozent

Abbildung 6 zeigt schließlich den Anteil der IP-Adressen eines Landes, welcher als Ziel von mindestens einer Traceroute diente. Hier zeigt sich, dass der im letzten Abschnitt bemerkte große Anteil an Zielen in den USA (40,1%), bezogen auf die Anzahl der den USA zugeordneten IP-Adressen(knapp 1,6 Mrd.[14]), gar nicht so groß ist. Vom Großteil der Länder wurden durchschnittlich zwischen 1,7% und 2% der zugeordneten IP-Adressen adressiert.

4.6.3 Hops

Analog zu den Hops auf IP-Adressebene (siehe Abschnitt 4.4.3) betrachteten wir auch die Hoplänge auf Länderebene. Das bedeutet in diesem Fall die Anzahl der Länder die an einer Traceroute beteiligt waren. Im Unterschied zu den IP-Adressen zählten wir hier also auch das Land aus dem die Traceroute gestartet wurde dazu. Eine Traceroute von Brasilien über Spanien in die USA hätte somit einen Hopwert von 3. Wir betrachteten hierzu nur die Traceroutes, bei denen Start- und Zielland bekannt waren und bei denen nicht alle IP-Adressen auf der IP,Trace unbekannt waren (nicht alle IP-Adressen konnten Länder zugeordnet werden, siehe Abschnitt 3.2). Auch hier berechneten wir, wieviele der Tracroutes, die über eine bestimmte Anzahl von Ländern gehen, eine Routingschleife auf ihrer zugehörigen IP-Trace aufwiesen. Auf Länderebene können Wiederholungen des selben Landes durchaus vorkommen, weswegen uns eine Schleifenerkennung auf dieser Ebene nicht sinnvoll vorkam. Ein Beispiel für eine vermeintliche Schleife auf Länderebene wäre eine Traceroute von Brasilien nach Alaska, auf der die USA doppelt vorkommen (BR->US->CA->US(Alaska)).

[14]Das entspricht ca. 45% der gesamten zugeteilten IP-Adressen.

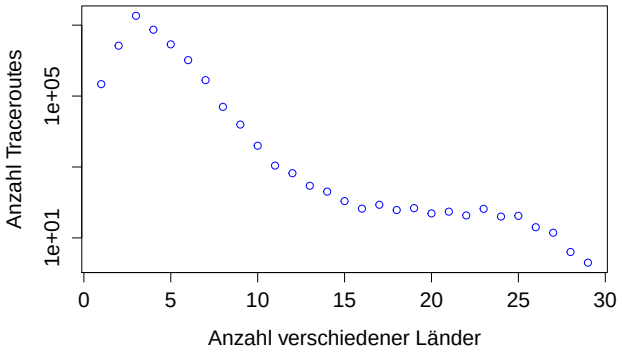

Abbildung 7: Anzahl der Traceroutes nach der Anzahl der Länder von Start- bis zum Zielland

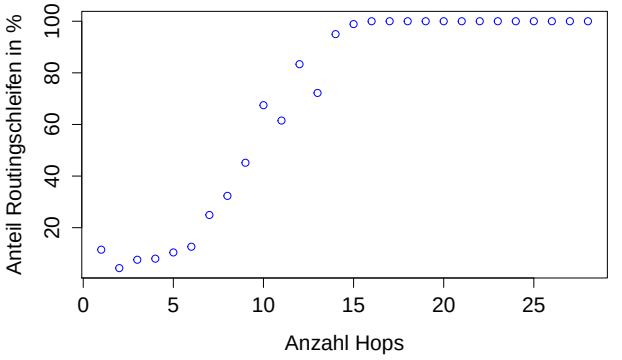

Abbildung 8: Anteil der Traceroutes mit Routingschleifen an gesamten Traceroutes nach Anzahl der Länderhops von Start- bis zum Zielland

Abbildung 7 zeigt wie viele Traceroutes jeweils eine bestimmte Anzahl an Ländern beinhalten. Im Gegensatz zu der Hopzahl bei den IP-Traceroutes, ist hier schon relativ schnell ein Abfall in der Anzahl Traceroutes mit höherem Länder Hopcount zu beobachten. Noch klarer ersichtlich wird dies, wenn man den Anteil an Routingschleifen in Abhängigkeit von der Länderzahl mit einbeieht (Abb. 8). Betrachtet man beide Abbildungen zusammen, sieht man dass der Großteil

der Traceroutes, die sehr wahrscheinlich ihr Ziel erreicht haben, weniger als ungefähr 7 mal das Land wechselten. Der Mittelwert der Anzahl beteiligter Länder an den Traceroutes ohne Schleifen betrug 3,44.

4.6.4 *Interessante Traceroute Verläufe*

Abschließend betrachteten wir auch noch die Strecken selbst, welche Pakete von einem Land in ein anderes nahmen. Einleitend zeigt Abbildung 9 wie oft die verschiedenen Länder auf den Routen (also nicht als Quelle oder Ziel) vorkommen.

Hier ist zu sehen, dass die USA eine wichtige Rolle im Internetverkehr spielt, da sie in vergleichsweise vielen Traceroutes vorkommt (≈ 23%). Auf Platz zwei folgen einige europäische Länder sowie Europa[15] selbst mit 3%-12.5%.

TeleGeography [22] hat eine Karte erstellt, welche alle Unterseekabel enthält die momentan[16] aktiv sind[23]. Auf dieser ist zu erkennen, dass vor allem die Länder häufig in den Traceroutes vorkommen, die eine gute Anbindung an Unterseekabel vorweisen können. Gut zu sehen ist dies auch an den afrikanischen Ländern, von denen fast nur Küstenländer vertreten sind. Auch zu erkennen ist, dass vor allem Afrika für die Differenz der Länderzahl zwischen Ländern auf den Routen (199) und Ländern als Ziel (243) verantwortlich ist, die in Tabelle 4 gezeigt wird. Tabelle 7 zeigt einen Überblick über die am meisten vertretenen Strecken im Traceroute Datensatz.

Ursprung	Ziel	Anzahl
Brasilien	USA	8.300.000
Costa Rica	USA	5.700.000
China	USA	4.400.000
Türkei	USA	3.500.000
Peru	USA	2.500.000

Tabelle 7: Top 5 der Strecken für Traceroutes

Da die GeoIP Daten von Maxmind einen nicht genauer bekannten Fehler aufweisen, entschieden wir uns, nur Länder zu betrachten, die an einer Route mit mindestens 1% an der Summe der Vorkommen aller Länder beteiligt sind.

Als erstes Beispiel suchten wir uns den größten Datensatz mit den Traceroutes von Brasilien in die USA aus und sind auf eine auf den ersten Blick überraschende Tatsache gestoßen. Abbildung 10 zeigt die Länder die Pakete auf dieser Route durchquert haben.

Wir hätten erwartet, dass Pakete entweder von Brasilien aus den direkten Weg in die USA über ein Unterseekabel nehmen oder eventuell noch über die schmale Landbrücke zwischen Nord- und Südamerika gehen. Interessanterweise geht aber ein relativ großer Anteil(≈ 26%) erst nach Spanien, bevor die Pakete über das nächste Unterseekabel zurück nach Amerika gesendet werden. Die Landbrücke wird hingegen gar nicht genutzt. Pakete von USA nach Brasilien nehmen den umgekehrten Weg, allerdings verlaufen nur noch ≈ 14% über Spanien. Kanadas Beteiligung ist unserer

[15] Wie erwähnt hat Europa einen eigenen Coutry Code (EU), den wir nicht auf der Karte abbilden konnten.
[16] Letzter Stand 17.September 2013.

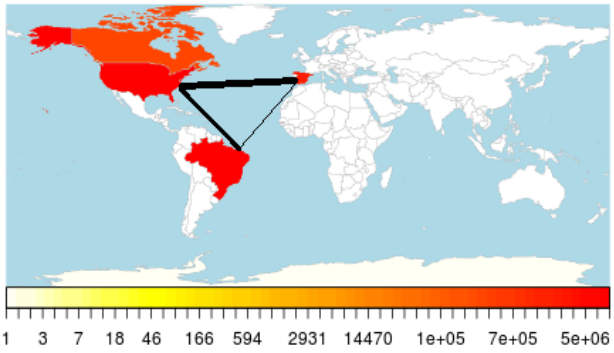

Abbildung 10: Anzahl Beteiligungen von Ländern an Routen von Brasilien in die USA

Meinung nach den Paketen geschuldet die an IP-Adressen in Alaska gesendet wurden. Aus Ressourcengründen war es uns leider nicht möglich eine genauere Auflösung als auf Länderebene vorzunehmen. Auf der Unterseekabelkarte [23] von Telegeography [22] ist zu sehen, dass Brasilien eigentlich relativ gut an die USA angebunden ist, wohingegen Spanien nur über ein einziges Unterseekabel erreichbar ist.

Eine mögliche Erklärung für den Umweg über Spanien könnte der Fakt sein, dass Routingentscheidungen unter anderem auch auf finanzielle Aspekten beruhen und nicht unbedingt nur auf geographischen. Internetanbieter die Pakete über Verbindungen anderer Anbieter routen, müssen diesen häufig Gebühren zahlen. Von daher wird versucht Pakete möglichst über eigene Routen zu versenden, so dass diese Zusatzgebühren möglichst gering ausfallen. Für die Verbindung Brasilien über Spanien in die USA vermuten wir, dass die starke Vertretung von Telefónica in Brasilien und Spanien ein Grund dafür ist, dass die Pakete nicht alle direkt in die USA gesendet werden. Telefónica ist teilweise Miteigentümer eines Unterseekabel jeweils zwischen Brasilien und Spanien, sowie eines weiteren zwischen Spanien und den USA. Daher die Vermutung, dass Telefónica[17] diese Unterseekabel nutzt, um im Vergleich zur direkten Verbindung Kosten einzusparen. In Abbildung 10 wurden die Unterseekabelverbindungen die hier relevant sind schematisch eingetragen, die Dicke der Striche steht hier für die Anzahl der Kabel, nicht zwangsläufig deren Kapazität.

Ebenfalls interessant ist der Verkehr von China nach Deutschland und umgekehrt. So verlaufen rund 17% der Traceroutes von China nach Deutschland über die USA, in umgedrehter Richtung sind es sogar rund 23%. Die beteiligten Länder der Traceroutes von China nach Deutschland sind in Abbildung 11 zu sehen.

Auffällig ist hierbei auch der Fakt, dass ein relativ großer Teil der Daten den direkten Weg ohne Zwischenstationen nimmt. Auch hier scheint der Datenverkehr über ein Unterseekabel abgewickelt zu werden, welches, unter anderem, Deutschland und China direkt verbindet (SeaMeWe-3 [23]).

4.6.5 *Fazit*

[17] Telefónica dient hier vor allem als Beispiel, es kann natürlich noch andere Gründe geben.

doi: 10.2313/NET-2014-03-1_03

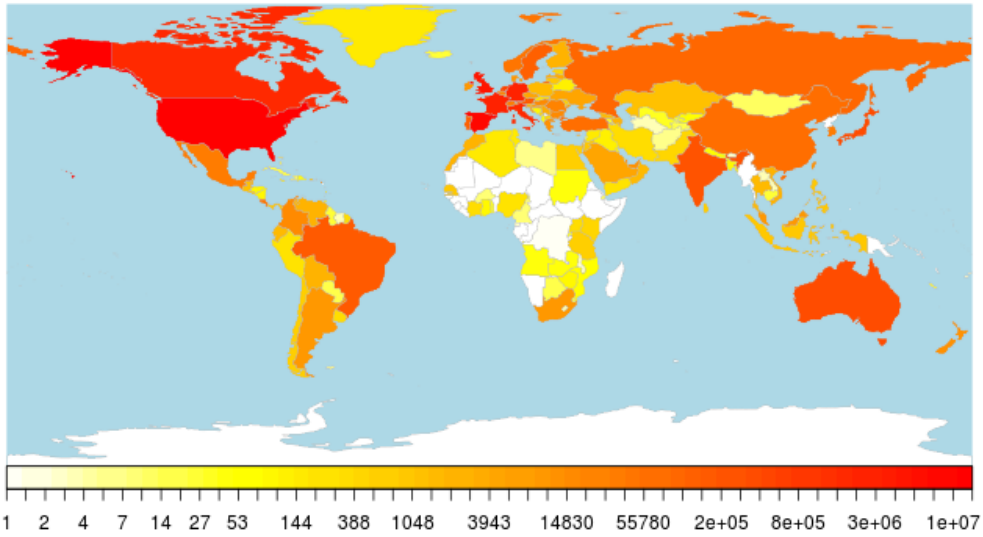

Abbildung 9: Anzahl Beteiligungen von Ländern an allen Traceroutes ohne Start- und Zielland

Zusammenfassend lassen sich aus dem Traceroute Datensatz einige interessante Schlüsse ziehen, beziehungsweise bekannte Vermutungen bestätigen. So sind Nordamerika und Europa Infrastrukturmäßig enorm wichtig für das Internet, viele Pakete werden durch Länder dieser Kontinente gesendet. Ebenfalls interessant sind Südamerika und Asien. Viele Länder in diesen Regionen sind durchaus wichtig für das Internet, allerdings zeigen die Traceroute Daten, dass in diesen Gebieten noch besonders viele Geräte durch einfaches Passwortraten angreifbar sind. Afrika hingegen ist aus Internetsicht noch relativ unerschlossen. Desweiteren sahen wir, dass Pakete um an ihr Ziel zu kommen nicht immer den geographisch gesehen kürzesten Weg nehmen, sondern unter Umständen auch Umwege machen um beispielsweise finanziellen Aspekten gerecht zu werden.

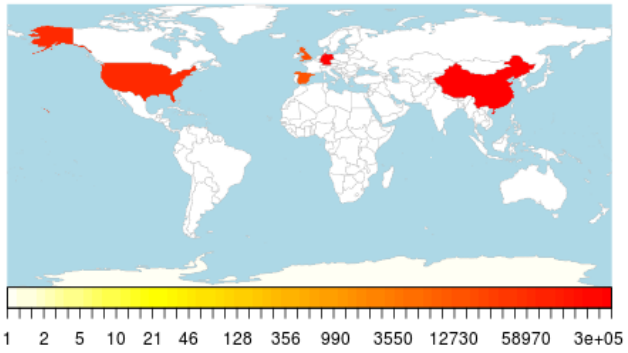

Abbildung 11: Anzahl Beteiligungen von Ländern an Routen von China nach Deutschland

5. RDNS

Das Domain Name System (DNS) ist eine zentrale Datenbank des Internets um eine Abbildung von Namen auf IP-Adressen zu ermöglichen. Dabei kann eine einzelne Domain auf verschiedene IP-Adressen zeigen und eine IP-Adresse kann mehreren Domain Namen zugeordnet sein. Die Zuordnung einzelner IP-Adressen zu deren Domains wird dabei als rDNS (reverse DNS) bezeichnet. Im Rahmen des Internet Census wurde dabei für einen großen Bereich des IPv4 Adressraums der jeweilige reverse DNS Eintrag bestimmt und abgespeichert. Im weiteren Verlauf des Kapitels erfolgt eine Analyse der Daten.

Für das Resultat einer rDNS Abfrage gibt es verschiedene Optionen. Ist ein rDNS Eintrag hinterlegt wird dieser in Form eines Hosts zurückgegeben. Falls kein Eintrag gefunden wurde, wird der Fehlercode (3) zurückgegeben. Eine leere Antwort kann entweder keiner Antwort oder keinem gefundenem Host entsprechen. Darüber hinaus existieren noch eine Reihe weiterer Fehlercodes die in Tabelle 8 näher spezifiziert werden.

5.1 Zahlen und Fakten

Während des Internet Census 2012 wurden im Gesamten 10,5 Milliarden Anfragen bezüglich des rDNS Eintrages gestellt, somit wurde jede verfügbare IP-Adresse im Durchschnitt 2,45 mal abgefragt. Insgesamt wurden dabei 3.7 Milliarden verschiedene IPs abgefragt, was einer Abdeckung von rund 86% entspricht (inkl. privater Blöcke). Von diesen 10,5 Milliarden Anfragen erhielten ca 2.8 Milliarden eine Antwort mit gesetzter rDNS Domain. Dies entspricht einem Anteil von $\approx$ 26%. Folglich lieferten rund 7.8 Milliarden ($\approx$ 76%) angefragte Host Adressen einen Fehlercode, keine oder eine leere Antwort. Über die Antworten mit leerer Nachricht ($\approx 41,7$ Millionen) kann, mangels fehlender Information, keine Aussage getroffen ob deren rDNS Eintrag leer ist, oder der entsprechende Host nicht geantwortet hat. Die Verteilung der Fehlercodes kann in Tabelle 9 betrachtet werden. Weiterhin auffallend ist: Einige der zurückgegeben Fehlercodes sind innerhalb der entsprechenden manpages nicht dokumentiert.

Auffällig ist, dass sämtliche gewonnen Daten deutlich von den „offiziellen“ Daten innerhalb der ursprünglichen Veröffentlichung abweichen. Da jedoch nichts über die genaue Art der Gewinnung der offiziellen Daten bekannt ist, kann nur

doi: 10.2313/NET-2014-03-1_03

Code	Nachricht
0	No error condition.
1	The name server was unable to interpret the query.
2	The name server was unable to process this query due to a problem with the name server.
3	The domain name does not exist
4	The name server does not support the requested kind of query.
5	The name server refuses to reform the specified operation for policy reasons.
65	The reply was truncated or ill-formated.
66	An unknown error occurred.
67	Communication with the server timed out.
68	The request was canceled because the DNS subsystem was shut down.

Tabelle 8: rDNS Fehler Meldungen [6]

gemutmaßt werden wie diese Abweichungen nach Oben (bis maximal Faktor 3) auftreten können. Eine mögliche Ursache ist der untersuchte Zeitraum: Der Autor schreibt in seiner Veröffentlichung, dass nur Messungen zwischen Mai und Oktober berücksichtigt werden. Im Oktober wurden jedoch noch weitere reverse DNS Messungen durchgeführt. Somit ist es wahrscheinlich, dass der hier untersuchte Datensatz signifikant Größer als der (noch nicht vollständige) Datensatz des Autors ist.

Fehlercode	#	%
2	1.690.584.869	21,85
3	5.885.552.163	76,08
5	11.603	0,00015
65	111.054	0,0014
66	1.256.886	0,02
67	40.038.354	0,52
70	118.858.282	1,54

Tabelle 9: Verteilung der Fehlercodes von rDNS Anfragen

5.2 Auswertungen

Aus der Datenbank können eine Reihe verschiedener Daten extrahiert werden, die im folgenden kurz vorgestellt werden. Eine Interpretation und weiterführende Auswertung der Daten wird anschließend im nächsten Abschnitt vorgenommen.

Im Verlauf der Arbeit muss beachtet werden, dass sämtliche Daten über einzelne IP-Adressen genommen wurde. Der Datensatz der reverse DNS Daten umfasst mehr als 10 Milliarden Einträge, eine Speicherung der Daten in eine Tabelle war nicht möglich, da der verfügbare Arbeitsspeicher nicht ausreichte um die gesamte Datenmenge laden zu können. Bedingt dadurch konnten mehrfach vorhanden Domains nicht gefiltert werden und somit entsteht eine gewisse Unschärfe in den Daten. Bei mehr als 10 Milliarden Datensätzen kann jedoch davon ausgegangen werden, dass bedingt durch das Gesetz der großen Zahlen [21] eine starke Korrelation zwischen den absoluten und relativen Werten vorhanden ist.

Weiterhin ist es, aufgrund der Verschlossenheit seitens MaxMind, nicht möglich, detaillierte Aussagen über die Genauigkeit der geographischen Positionen der Domains zu treffen. Es ist weder bekannt wie die Positionen bestimmt werden, noch wie die zugehörigen Fehlerraten ermittelt werden. Um qualitative Aussagen über die nachfolgenden Daten treffen zu können, wäre dies aber unabdingbar.

Begonnen wird zunächst mit der einfachsten Form der Auswertung, einer Abbildung der lokalisierten Domains in die zugehörigen Länder. Siehe hierzu die Abbildung 12 und die zugehörige Tabelle 10. In dieser Ansicht werden dabei nur distinkte Domains betrachtet, somit werden mehrfach genutzte Domains (zum Beispiel dynamische Hosts von Internetzugängen) ignoriert.

Land	# IPs
USA	975.064.787
Japan	282.994.662
Deutschland	138.511.437
Großbritannien	127.073.049
Frankreich	110.419.032
Italien	87.925.156
Niederlande	70.746.202
China	70.081.097
Brasilien	68.774.971
Spanien	62.006.960

Tabelle 10: Top 10 der Länder mit den meisten rDNS Einträgen

Aus der Anzahl der rDNS Einträge pro Land lassen sich Rückschlüsse über die Verbreitung des Internets in den einzelnen Ländern getroffen werden. Ausführungen hierzu werden in Abbildung 13 und der Top 10 Tabelle 11 dargestellt.

Land	Domains/Bewohner
Schweden	4,36
Norwegen	4,30
Niederlande	4,20
Finnland	3,36
USA	3,11
Dänemark	2,72
Taiwan	2,56
Japan	2,24
Schweiz	2,20
Australien	2,17

Tabelle 11: Top 10 Anzahl der Länder mit den meisten rDNS Einträgen pro Person, es werden nur Länder mit mehr als 1 Mio Einwohner angetragen

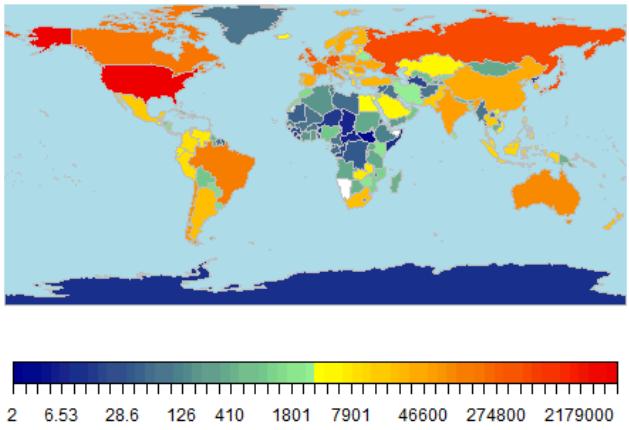

Abbildung 12: Absolute Anzahl der distinkten Domains pro Land

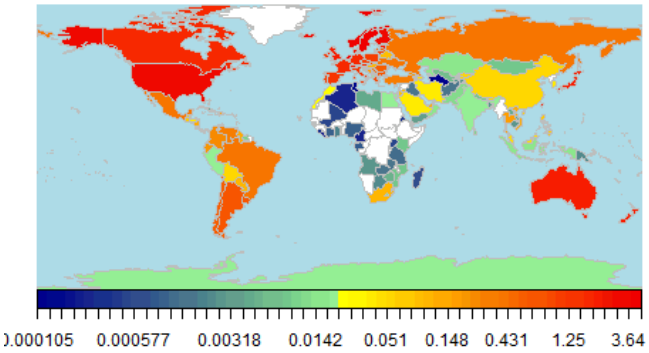

Abbildung 13: Verhältnis Domains pro Einwohner, Länder mit einem Verhältnis von weniger als 10^{-5} Domains pro Person sind weiß dargestellt.

Weiterhin gibt es durch den Datensatz auch die Möglichkeit, die Abhängigkeit einzelner Staaten voneinander darzustellen. Hierfür gibt es zwei verschiedene Varianten. Eine Variante wird in Tabelle 12 und Abbildung 14 gezeigt. Dabei wird anhand des Beispiels der USA gezeigt wie stark einzelne Staaten vom jeweiligen Land abhängig sind. Bei den Auswertungen dieser Abhängigkeiten werden im folgenden nur distinkte Domainnamen herangezogen.

Auch eine Invertierung dieser Ansicht bietet interessante Einblicke in die Infrastruktur verschiedener Länder. Daraus können dann Rückschlüsse über den Einsatzzweck, aber auch diplomatische Beziehungen beziehungsweise Vertrauensverhältnisse gezogen werden. Dies wird anhand der Domains von Deutschland, Togo und Nordkorea gezeigt (siehe hierzu die Abbildungen 15, 16 und 17).

Im Zuge des PRISM [24] Skandals wird abschließend noch eine Analyse über die „5 Eyes“ [5] durchgeführt, deren Resultate aus der Tabelle 14 gezogen werden. Da in dieser Auswertung erneut eine Betrachtung einzelner Hosts von Interesse ist, wird die Auswertung unter Berücksichtigung dynamischer IP Adressen durchgeführt.

5.3 Interpretation

Die einfachste Möglichkeit der Auswertung auf Basis der absoluten Werte der Domains in den verschiedenen Ländern bringt kaum Überraschungen mit sich. Es zeigt sich eine klare Dominanz der westlichen Industrienationen. 9 der 10 Länder sind diesen zugehörigen und stellen die Heimat für 1,7 Milliarden Domains, was einem Anteil von $\approx$ 62% an allen aufgelösten IP-Adressen darstellt.

Wenn die so gewonnen Daten jedoch einer Normalisierung über die aktuelle Einwohnerzahl [1] unterzogen werden, ergeben sich interessante neue Aspekte. So heben sich dann besonders hochtechnisierte Länder wie Taiwan hervor. Aber auch die skandinavischen Länder, die sich zum Zeitpunkt des Census durch eine sehr liberale Haltung bezüglich Urheberrechts und Meinungsfreiheitssachen ausgezeichnet haben, sind in dieser Statistik in den führenden Positionen vertreten. Weiterhin ist auffallend, dass weder Deutschland, Frankreich, noch England in dieser Darstellung in einer der führenden Rollen vertreten sind, obwohl diese Länder innerhalb Europas mithin die wirtschaftlich leistungsfähigsten Staaten darstellen.

Aus der Abhängigkeitsdarstellung einzelner Staaten anhand der USA ergeben sich sodann auch einige interessante Faktoren. So ist zum Beispiel eine protektive Nutzung dieser Abhängigkeit durchaus möglich, aber auch offensiv-aggressive Handlungen sind aus einer zu großen Abhängigkeit möglich. Protektion könnte beispielsweise durch die Auslagerung bzw. redundante Haltung kritischer Systeme durch schützende Nationen erfolgen. So könnte eine tiefgreifende technische Kontrolle, bedingt durch die Abhängigkeit in der Internet Infrastruktur, genutzt werden um unmittelbaren diplomatischen Zwang anzuwenden. Die Resultate am Beispiel der USA werden im Folgenden kurz dargestellt:

Besonders auffällig ist hierbei die sehr starke Abhängigkeit einiger Staaten Afrikas. Aber auch die sehr schwache Abhängigkeit einiger Staaten, beispielsweise sei hier Israel genannt ist auffällig. Obwohl politisch eine tiefe Verbundenheit zwischen den USA und Israel zu beobachten ist, ist nur ein Bruchteil ($< 1\%$) israelischer Domains in den USA gehostet. Die Erklärung dieses Phänomens dürfte aber in der sehr starken technischen Infrastruktur Israels liegen, so dass grundsätzlich kein Bedarf für eine Abhängigkeit Israels von anderen Ländern besteht.

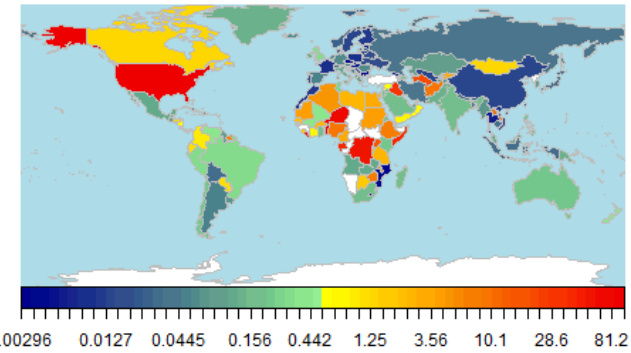

Abbildung 14: Landkarte zur Abhängigkeit einzelner Staaten von den USA in %

doi: 10.2313/NET-2014-03-1_03

Land USA	Domains Total	Domains	Abhängigkeit [%]
USA	6.465.676	6.491.618	99,60
IO	275.933	284.817	96,88
Niederlande	891	1.111	80,20
Rrasilien	9.465	12.284	77,05
Kongo	2.686	4.998	53,74
Jungferninseln	708	1.450	48,83
West-Samoa	20.379	48.317	42,18
Südgeorgien	569	1.463	38,89
Kokosinseln	50.862	130.823	38,88
VC	374	1.057	35,38
Montenegro	10.992	33.947	32,38

Tabelle 12: Top 10 der Länder (zu lange Namen sind mit dem Country Code abgekürzt) mit den Prozentual am meisten in den USA gehosteten Domains, US selbst als Vergleich.

Aber auch die inverse Betrachtungsweise kann interessante Fakten offenbaren. Mit dieser Ansicht können dann Rückschlüsse darüber gezogen werden, welchen Ländern ein Land *A* Teile seiner, mehr oder weniger, kritischen Infrastruktur anvertraut. Auch ein Outsourcing zum Zwecke von Zensurmaßnahmen ist dabei denkbar. Im folgenden Beispiel wird dies anhand Deutschlands veranschaulicht. Naheliegend ist, dass der größte Anteil der länderspezifischen Domains im Heimatland der jeweiligen Domain gehostet wird. Im Falle von Deutschland sind dies mehr als 90%. Somit wäre es beispielsweise für die deutsche Regierung ein leichtes weitreichende Zensurmaßnahmen über den Zugriff auf deutsche Rechenzentrumsbetreiber zu realisieren, da nur innerhalb der eigenen Staatsgrenzen Zwang ausgeübt werden müsste.

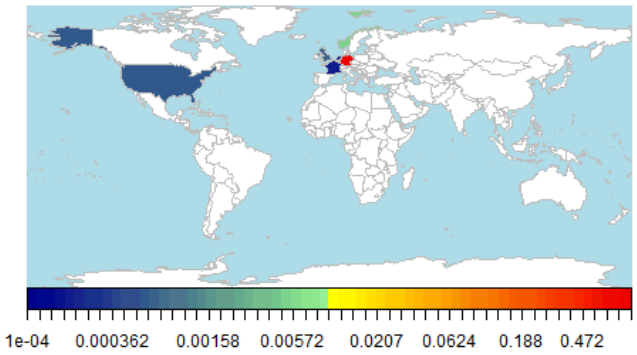

Abbildung 15: Landkarte zur Verteilung von .de Domains über verschiedene Länder in %

Ein gegenteiliges Beispiel, mit einer sehr homogen Verteilen über die einzelnen Ländern stellt die Domain Endung „.to“ [25] von Togo in Abbildung 16 dar. “.to“-Domains sind relativ gleichmäßig über viele verschiedene Länder verteilt. Ursächlich für dies sind aber eher nicht diplomatische Gründe, sondern liegen im Registrierungsprozess der „.to“-Domains begründet. Bei „.to“-Domains ist keine whois Abfrage möglich, sodass der Eigentümer der jeweiligen Domain einen gewissen Grad an Anonymität genießt. Aus diesem Grund werden to Domains oftmals für Webseiten am Rande der Legalität genutzt.

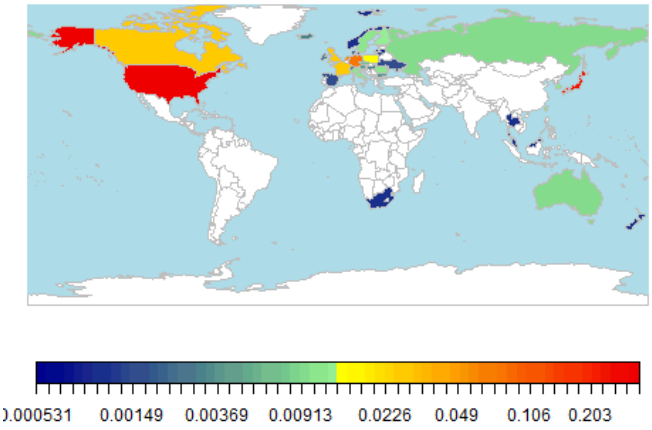

Abbildung 16: Landkarte zur Verteilung von .to Domains über verschiedene Länder in %

Eine sehr interessante Fragestellung fokussiert die geographische Verteilung von Domains diplomatisch isolierter Staaten. Im folgenden Beispiel wird dabei Nordkorea mit der Landesdomain „.kp“ untersucht. Nordkorea ist international weitestgehend diplomatisch isoliert. Nordkorea hat dabei nur einen sehr kleinen Adressblock mit gesamt 1024 IP-Adressen [29]. Bedingt dadurch sind die meisten der Verfügbaren Domains staatlicher Natur - die interessantere Fragestellung ist dabei aber: Welche Länder stellen Nord Korea Kapazitäten zur Verfügung und brechen somit die fast weltweiten (China hat keine Embargos gegen Nordkorea) Embargos. Die Beurteilung, ob faktisch auch ein Bruch der Embargos vorliegt kann dabei von uns nicht abschließend beurteilt werden, da diese Fragestellung komplexe völkerrechtliche Aspekte mit sich bringt [3]. Wie in der Abbildung 17 und der Tabelle 13 zu sehen ist, gibt es bei der Anzahl der nordkoreanischen IPs keine Überraschungen. Hinter den im aufgeführten IP-Adressen steckt dabei hauptsächlich technische Infrastruktur oder regierungsnahe Adressen. Ebenfalls wenig überraschend ist der Fakt, dass keine der Domains oder IP-Adressen von einem deutschen Internetzugang erreichbar ist. Interessanter ist die Domain, die einen großen, internationalen Lieferdienst vermuten lässt, welcher aber seinen Dienst in Nord Korea bereits seit mindestens 2011 wieder eingestellt hat [27], jedoch der zugehörige DNS Eintrag weiterhin vorhanden ist. Weiterhin ist bemerkenswert, dass keinerlei „.nk“ Domains auf chinesischen IP Adressen erreichbar sind, aufgrund der Verbundenheit Nordkoreas mit China [9] wäre dies tendenziell zu erwarten gewesen.

Im Hinblick auf den kürzlichen Spionageskandal der NSA [24] ergibt sich aus dem Census die Möglichkeit, den direkten Zugriffe der großen Geheimdienste, den „Five Eyes“ [5] abzuschätzen. Diese fünf Geheimdienste umfassen dabei die Länder: USA (NSA), England (GCHQ), Neuseeland (GCSB), Australien (DSD) und Kanada (CSEC). Über die genauen Details dieses Zusammenschlusses ist nur wenig bekannt, da die zu Grunde liegenden Verträge geheim sind. Es wird jedoch davon ausgegangen, dass die 5 Augen Einsicht in die gesamten Geheimdienstdaten aller beteiligter Staaten

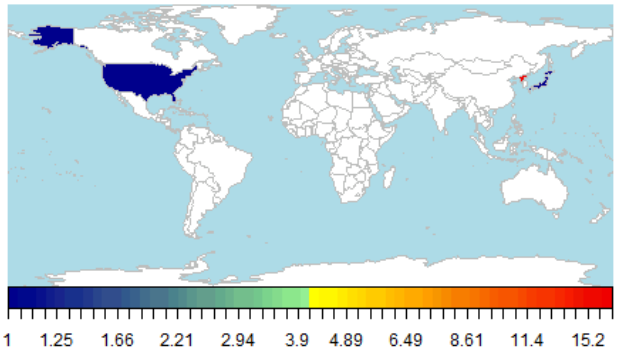

Abbildung 17: Landkarte über die Verteilung nordkoreanischer Domains

#	Domain	Land
1	kita.no.kuni.kara.-83-winter.kp	JP
1	www.globalbusiness.ups.kp	US
1	spinef1.star.net.kp	KP
2	ns1.kptc.kp	KP
2	ns2.kptc.kp	KP
1	mail.silibank.net.kp	KP
2	naenara.com.kp	KP
4	smtp.star-co.net.kp	KP
1	ns1.star.edu.kp	KP
4	mail.star.edu.kp	KP

Tabelle 13: Übersicht über die gesammelten nordkoreanischen Domains. Bedingt durch mehrfach gefundene Einträge können Domains mehrfach auftreten, zb. durch redundante Systeme

haben.

Grundsätzlich muss davon ausgegangen werden, dass diese Geheimdienste in der Lage sind innerländisch den gesamten Datenverkehr zu überwachen und zu manipulieren, da diese mittels der jeweiligen Rechtsprechung (z.B. Patriot Act [11]) die jeweiligen Betreiber zur Kooperation zwingen können. In Tabelle 14 können dabei die Ergebnisse der reinen Domainzahlen betrachtet werden. Gesamt kommen diese fünf Länder auf mehr als 1 Milliarde reverse DNS Einträge, was mehr als 44% aller rDNS Einträge des gesamten Census beträgt. Auch wenn dabei dynamische IP-Adressen etc. inkludiert sind, haben diese 5Geheimdienste, beauftragt mit dem Schutz von weniger als 10% der Weltbevölkerung, indirekten Zugriff auf mehr als 40% der weltweiten Internet Infrastruktur.

Ein weiterer Aspekt ist dabei auch die Abhängigkeit anderer Länder von diesen fünf Augen, diese wird in Abbildung 18 sichtbar. Darin ist klar zu sehen, die Geheimdienste in der Lage sind mehr als 50% der IPs eines fremden Landes zu kontrollieren, analysieren und manipulieren. Für die Top 10 dieser so theoretisch kontrollierbaren Länder kann die Tabelle 15 betrachtet werden.

5.4 Fazit

Wie im vorherigen Abschnitt gesehen, bietet auch der rDNS Datensatz eine Reihe interessanter Ansätze um Aussagen

Land	”5 Eyes”	Gesamtanzahl	Abhängigkeit in [%]
IO	281.251	284.817	98,75
Niger	933	1.111	83,98
Belize	9.671	12.284	78,73
Kongo	3.618	4.998	72,39
CX	5.952	10.768	55,27
VG	781	1.450	53,86
Samoa	21.288	48.317	44,06
GS	643	1.463	43,95
Tuvalu	168.818	391.714	43,10
CC	53.913	130.823	41,21

Tabelle 15: Abhängigkeit einzelner Länder (zu lange Namen sind mit dem Country Code abgekürzt) von den 5 Eyes.

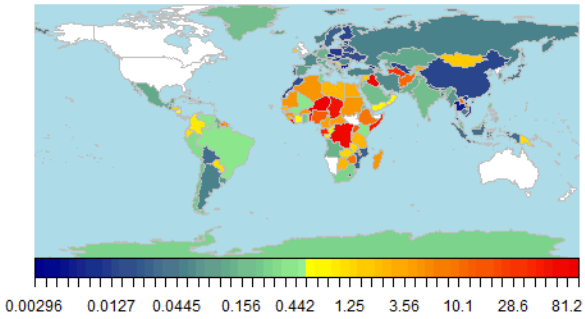

Abbildung 18: Abhängigkeit verschiedener Länder von den „5 eyes“ in %

über den Zustand des Internets durchführen zu können. Die Qualität dieser Aussagen ist dabei aber von vielen externen Faktoren abhängig, die nur schwer zu kontrollieren und evaluieren sind. Auf der einen Seite steht dabei die Qualität des Census Datensatzes. Aufgrund Zeit, aber auch Kapazitätsfaktoren ist eine Evaluierung der Daten im Rahmen dieser Arbeit nicht möglich.

Für die Zukunft können mit diesem Datensatz noch eine Reihe weiterer Experimente durchgeführt werden, hierfür sollte jedoch ein System mit mehr Arbeitsspeicher oder einem anderen, nicht RAM zentrischen, DBMS verwendet werden. Weiterhin wären auch Auswertungen interessant, bei denen auf eine deutlich geringere Abstraktionsebene zurückgegriffen wird. Als Datengrundlage könnten hierfür die, von Maxmind, angebotenen Geolokationsdatensätze auf Stadtebene benutzt werden, die im Verlauf dieser Auswertung nicht berücksichtigt wurden.

6. VERWANDTE ARBEITEN

Im Bereich der Messung und Analyse von Internetverkehr gibt es einige Arbeiten, welche sich direkt oder indirekt mit dem Internet Census 2012 in Verbindung bringen lassen.

So schlugen Mark Allman und Vern Paxson in [2] 2007 einige Richtlinien vor, welche Forschern bei solchen großflächigen Messungen dabei helfen sollen die Daten auszutauschen, ohne dabei in ethische oder rechtliche Probleme zu geraten. Ein Beispiel dafür ist das Aufstellen klarer Regeln was mit den Daten getan werden darf und was nicht (z.B.: ”The user

Geheimdienst	Domains	Anteil [%]	Bevölkerung	Domains pro Einwohner	Gesamtkontrolle [%]
USA	975.064.149	80,05	313.847.465	3,11	35,45
Großbritannien	127.073.049	10,43	63.047.162	2,02	4,62
Kanada	60.007.066	4,93	34.030.586	1,76	2,18
Australien	47.267.158	3,88	21.766.711	2,17	1,72
Neuseeland	8.622.234	0,71	4.290.347	2,01	0,31
Total	1.218.033.656		436.982.271		44,29

Tabelle 14: Übersicht über die Anzahl der von den Five Eyes kontrollierten Domains

may use the data to develop new techniques for finding subverted hosts that are part of botnets."[2]). Anonymisierung und Aggregation von Daten sind weitere Vorschläge für die sichere Weitergabe von Daten. Grundsätzlich wird im von uns betrachteten Teil des Internet Census 2012 keine dieser Regeln angewandt, allerdings bertrat der anonyme Urheber durch Verwendung seines Botnetzes bereits eine Grauzone.

Einen effizienten Scanner für internetweite Scans namens ZMap [7] stellten Durumeric et al. im August 2013 vor. Dieser ist in der Lage, von einem einzelnen Rechner aus den kompletten IPv4 Adressraum in weniger als 45 Minuten auf einen bestimmten offenen Port hin zu untersuchen. Der Schwerpunkt bei der Entwicklung von ZMap lag dabei darauf, den Scan so schnell wie möglich durchzuführen. So geht ZMap beispielsweise davon aus, dass der Scanner weder Quell- noch Zielnetzwerk überlastet. Aufgrund dieser Annahme umgeht NMap [16] den TCP/IP Stack, generiert Ethernet Frames selbst und sendet diese so schnell es dem Hostrechner möglich ist aus. Im Vergleich dazu nutzte der Autor des Census für seine Scans das CARNA Botnetz, welches dezentral auf ungefähr 420.000 Geräten lief. Dabei benutzte der Carna nur so viel Ressourcen, dass der normale Betrieb des genutzten Gerätes nicht eingeschränkt wurde.

Direkt dem Internet Census 2012 gewidmet hat sich Parth Shukla[20]. Sein Schwerpunkt lag dabei aber weniger auf den erhobenen Daten, als vielmehr auf dem CARNA Botnetz und den angreifbaren Geräten. Für diese Analysen verwendete er Daten des Census Authors, welche nicht öffentlich verfügbar sind. Er betrachtete dabei unter anderem Hersteller, Standort und den verfügbaren Speicher der ungefähr 1,2 Millionen angreifbaren Geräte. Den Hersteller konnte Shukla aus den verfügbaren MAC Adressen auslesen, dabei stellte er fest, dass bestimmte Hersteller relativ häufig vertreten waren. Aufgrund einer Länderzuordnung der angreifbaren Hardware machte Shukla sichtbar, dass allein 56% der angreifbaren Geräte China zuordenbar waren. Rund 69% der angreifbaren Geräte besaßen zwischen 32 MB und 256 MB Arbeitsspeicher, es gab allerdings beispielsweise auch eines mit 4,5 TB RAM (China). Abschließend berechnete Shukla noch, dass durchschnittlich ≈0,088 Geräte pro /24 Subnetz[18] durch den von CARNA genutzten Telnet Zugang angreifbar waren. Speziell auf China bezogen waren durchschnittlich sogar ≈0.56 Geräte pro /24 Subnetz verwundbar. Generell bestätigen unsere Untersuchungen diese Unsicherheit in China, allerdings zeigte auch Südamerika gewisse Verwundbarkeiten.

Anja Feldmann betrachtete 2013 den Census Datensatz mit Schwerpunkt auf ICMP Scans genauer[8]. Dabei stellte sie unter anderem fest, dass die Daten zwar authentisch sind, allerdings mit Vorsicht behandelt werden sollten. Grund dafür ist beispielsweise das Fehlen von Angaben, wie genau bestimmte Daten erhoben wurden. Auch konnte sie Veränderungen der Messmethoden im Laufe der Zeit beobachten, was vom Ersteller des Census nicht dokumentiert wurde. Unsere eigenen Beobachtungen zeigen, dass auch der Traceroute und der rDNS Datensatz gewisse Auffälligkeiten vorweisen, welche sich ohne weitergehende Informationen wie die Daten genau erhoben wurden nicht erklären lassen.

[18]Ein /24 Subnetz beinhaltet 256 Adressen.

7. ZUSAMMENFASSUNG

Der Internet Census 2012 bietet mit seinen 9 Terabyte an Daten Unmengen an Analysemöglichkeiten. Nachdem wir uns kurz mit der Methodik befasst haben, mit der der Autor die Daten gesammelt hat (CARNA Botnetz), nahmen wir uns in dieser Arbeit zwei der insgesamt acht Datensätze heraus und untersuchten diese eingehender. Um mit den bis zu 10 Milliarden Einträgen umfassenden Daten möglichst effizient und effektiv umgehen zu können, nahmen wir uns das DBMS MonetDB zu Hilfe. MonetDB fehlten zwar noch einige Features, welche Datenbanksysteme wie PostgreSQL, die wesentlich ausgereifter sind, besitzen, allerdings zeigte MonetDB im Vergleich zu PostgreSQL einen enormen Geschwindigkeitsvorteil bei der Auswertung von Datenbankabfragen. Dieser Vorteil liegt hauptsächlich in der spaltenorientierten Arbeitsweise von MonetDB. Bei Tabellen mit mehr Spalten als Zeilen ist Postgres idR. die bessere Wahl.

Ein weiteres Tool, welches wir verwendeten, war die Pythonbibliothek `pygeoip` um eine Zuordnung der IP-Adressen zu den einzelnen Ländern durchzuführen. Möglich war dies unter Zuhilfenahme der frei verfügbaren MaxMind [13] GeoLite Datenbank. Allerdings waren wir nicht in der Lage, hierzu genaue Fehlerquoten zu bestimmen, die Daten sind also nicht hundertprozentig korrekt.

Die, zur Extraktion der einzelnen Domainbestandteilen, genutzte Bibliothek `tldextract` hinterließ einen gemischten Eindruck. Einerseits wurde durch die Nutzung der Suffix Listen ein großer Teil der Hostadressen korrekt in die einzelnen Bestandteile aufgeteilt, aber bei mehreren Milliarden Einträgen reichte bereits eine sehr niedrige Fehlerquote aus um umfangreiche manuelle Nacharbeiten erforderlich zu machen. So musste jeder erzeugte Datensatz im weiteren Verlauf händisch nachbearbeitet werden um Daten weiter verarbeiten zu können.

Der erste der betrachteten Datensätze bestand aus 68 Millionen Traceroutes, welche von rund 275.000 verschiedenen

Geräten weltweit aus gestartet wurden. Ein Problem, dass die Traceroutes aufwiesen, war die ungenaue Dokumentation des Autors, welcher den Datensatz ohne weitergehende Informationen veröffentlichte. So fanden wir einige Traceroutes, die ihr Ziel aufgrund von aufgetretenen Routingschleifen vermutlich nie gefunden haben, was aber ohne eigene Analyse der Daten nicht ersichtlich wurde, da die Ziel IP der Traces ganz normal angegeben war. Die IP-Adressen der Traces ordneten wir unter anderem auf Länderebene zu, um Rückschlüsse auf die Standorte der verwendeten Geräte und den Weg der versendeten Pakete ziehen zu können. So zeigte sich, dass Pakete teilweise einen geographisch gesehen längeren Weg nahmen als nötig gewesen wäre. Desweiteren wurde aus den Daten ersichtlich, welche Kontinente eine besonders wichtige Rolle für das Internet spielen (Nordamerika, Europa). Dies ist zwar keine neues Erkenntnis, bestätigt aber die bisherigen Annahmen.

Trotz der Tatsache, dass es systematische Abweichungen von der initialen Auswertung des Census Autors gibt und der Problematik der nicht eindeutigen Domainnamen, zeigt der rDNS Datensatz des Census einige interessante Fakten auf und bestätigt gehegte Vermutungen. So war bereits im Vorfeld erwartbar, dass die westlichen Industrienationen das Backbone des Internets darstellen. Zu mehr Erstaunen führt dabei schon, dass die pro Kopf Anbindung in hochtechnisierten Staaten wie Taiwan den gesamten Westen deutlich überbieten. Auch die untersuchte nicht Abhängigkeit einzelner Staaten stellt interessante Aspekte heraus. Die finale Auswertung im Bezug auf den Zusammenschluss der Geheimdienste - den „5 Eyes“ - bestätigt, dass der aktuelle Geheimdienstskandal keinesfalls unterschätzt werden darf. Eine generische Auswertung des Datensatzes gestaltet sich aber, aufgrund der mannigfaltig enthaltenen Informationen als schwierig, so dass eine spezifische und gezielte Auswertung erforderlich ist.

Abschließend lässt sich sagen, dass es uns gelang, in den Daten des Internet Census 2012 einige interessante Fakten zu finden. Und bei dem Umfang der restlichen, von uns hier nicht genauer betrachteten Daten, ist es sehr wahrscheinlich, dass sich hier noch weitere Analysemöglichkeiten bieten.

8. LITERATUR

[1] Übersicht über die Weltbevölkerung. http://www.internetworldstats.com/list2.htm.

[2] M. Allman and V. Paxson. Issues and etiquette concerning use of shared measurement data. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 135–140. ACM, 2007.

[3] S. Arnold. Das Handelsembargo - völkerrechtliche, europarechtliche, nationale Grundlage. *Bucerius Law School, Hamburg - Seminararbeit*, 2004.

[4] H. Asghari. Python ip to asn lookup module pyasn. https://code.google.com/p/pyasn/, 2009.

[5] G. Braune. Geheimbund ”Five Eyes Der exklusive Club der Geheimdienste. Tagesspiegel http://www.tagesspiegel.de/politik/geheimbund-five-eyes-der-exklusive-club-der-geheimdienste/8450796.html.

[6] Doxygen. evdns.h file reference. http://monkey.org/~provos/libevent/doxygen/evdns_8h.html, 09 2013.

[7] Z. Durumeric, E. Wustrow, and J. A. Halderman. Zmap: Fast internet-wide scanning and its security applications. In *22nd USENIX Security Symposium*, 2013.

[8] A. Feldmann. Internet census taken by an illegal botnet - a qualitative analysis of the measurement data. In *Talk at Dagstuhl*, 2013.

[9] R. Kirchner. Nordkoreas einziger Verbündeter. http://www.dradio.de/dlf/sendungen/einewelt/1969023/, Januar 2013.

[10] J. Kurkowski. tldextract projekt webseite. https://github.com/john-kurkowski/tldextract.

[11] J. Laas. Der Patriot Act und Datenschutz in der EU. http://www.telemedicus.info/article/2477-Der-Patriot-Act-und-Datenschutz-in-der-EU.html, 12 2012.

[12] Maxmind. Übersicht über die Datenqualität. http://www.maxmind.com/de/city_accuracy.

[13] Maxmind. Offizielle Webseite. http://www.maxmind.com/de/home.

[14] Maxmind. Zugeordnete IP Adressen pro Land. http://www.maxmind.com/de/techinfo.

[15] MonetDB DBMS Projekt Webseite. http://www.monetdb.org, 2013.

[16] Nmap. http://nmap.org/.

[17] Pivotel Greenplum DBMS Projekt Webseite. http://www.gopivotal.com/pivotal-products/data/pivotal-analytic-database.

[18] Postgres DBMS Projekt Webseite. http://www.postgresql.org/.

[19] PyGeoIP Projekt Webseite. https://pypi.python.org/pypi/pygeoip/, 2013.

[20] P. Shukla. Compromised devices of the carna botnet. In *ausCERT 2013*, 2013.

[21] Statista. STATISTA Statistik-Lexikon: Definition Gesetz der großen Zahlen. http://de.statista.com/statistik/lexikon/definition/58/gesetz-der-grossen-zahl/.

[22] TeleGeography. http://www.telegeography.com/.

[23] TeleGeography. Submarine cable map. http://www.submarinecablemap.com/.

[24] The Guardian. Prism. http://www.theguardian.com/world/prism.

[25] Tonic Domain Registry. Tonic Domain Registry. http://www.tonic.to/.

[26] Unbekannt. Internet Census 2012. http://internetcensus2012.bitbucket.org/paper.html, 2012.

[27] UPS. UPS - Trade Embargos Import/Export. http://www.ups.com/ga/CountryRegsPrint?loc=en_US&origcountry=US&destcountry=KP&cat=0150160170110180190200210230240090040140\discretionary{-}{}{}006007008002012010005013003&PrintRegulations=PrintRegulations.

[28] Wikipedia. AS Typen. http://en.wikipedia.org/wiki/Autonomous_System_(Internet).

[29] M. Williams. Wagt Nordkorea den Schritt ins Web? *Computerwoche*, 06 2010.