Unconstrained Endpoint Profiling (Googling the Internet)

Florian Haemmerling
(effex@cs.tu-berlin.de)

Seminar "Network Architectures: Internet Measurement" ,
Technische Universität Berlin

WS 2008/2009 (Version vom 23. Januar 2009)

Zusammenfassung

Das Paper „Unconstrained Endpoint Profiling (Googling the Internet)“ [1] stellt eine Methode zur Analyse von Internet Verkehr vor. Im Gegensatz zu gängigen Ansätzen, welche häufig Paketmitschnitte von Knotenpunkten benötigen um Aussagen über Endpunkte zu treffen, ist „Unconstrained Endpoint Profiling“ (UEP) auch ohne solche Daten möglich. UEP geht statt dessen davon aus, dass Informationen zu Endpunkten im Internet frei verfügbar und abrufbar sind: über die Google Suchmaschine [2]. Diese Informationen werden gefiltert und geordnet wodruch ein Bild der untersuchten Endpunkte entsteht. Dieses Ergebnis kann sich durchaus mit auf Paketanalyse gewonnenen Erkenntnissen messen und lässt sich sogar mit diesen kombinieren. Der beschriebene Ansatz wird dann benutzt um einen Einblick in das unterschiedliche Verhalten von Endpunkten in vier Weltregionen (Asien, Südamerika, Nordamerika und Europa) zu erhalten. Die daraus gewonnenen Erkenntnisse zeigen interessante Unterschiede und Gemeinsamkeiten dieser Regionen auf.

1 Einleitung

Was machen Menschen im Internet? Welche Webseiten werden am häufigsten besucht, welche Applikationen am meisten benutzt? Die vorgestellte Methode soll es ermöglichen Antworten auf diese Fragen auf überregionaler Eben zu finden. Es wird versucht ein globales Bild der Internetnutzung zu erstellen. Hieraus werden wiederum kulturelle Unterschiede und Gemeinsamkeiten ersichtlich und Interessen der untersuchten Regionen ablesbar.

Was das neue an dieser Methode ist und warum bisherige Mittel nur bedingt geeignet sind, wird in Kapitel 2 genauer erläutert. Daran anknüpfend wird in Kapitel 3 das grundlegende Vorgehen von UEP erklärt und darauf aufbauend in Kapitel 4 genauer beschrieben.

Das Verfahren wird dann angewendet um Antworten auf die einleitenden Fragen zu finden und die Ergebnisse dann in Kapitel 5 präsentiert. Um heraus zu finden, wie verläßlich diese Ergebnisse sind wird im folgenden Kapitel 6 unter anderem ein direkter Vergleich zu einer anderen Methode gemacht.

In Kapitel 7 werden die gewonnenen Erkenntnisse kritisch hinterfragt und mögliche Probleme aufgezeigt. Ausserdem werden einige Ergänzungsvorschläge zu UEP gemacht.

2 Hintergrund

In diesem Abschnitt werde ich kurz erklären, in welchem Bereich und mit welchem Zweck die Analyse von Netzverkehr am häufigsten angewandt wird.

Grundsätzlich lassen sich zwei große Bereiche unterscheiden, die ein Interesse an der Analyse von Netz verkehr haben. Zum einen ist dies die statistische Auswertung. Hierbei geht es hauptsächlich um das Ablesen von Trends und Zugriffsstatistiken um daraus wiederum z. B. auf andere Eigenschaften und Gewohnheiten der Nutzer Rückschlüsse zu ziehen. Ein weiterer wichtiger Bereich ist die Administration von Netzwerken. Dies beginnt schon bei der Planung und dem Aufbau ist aber mindestens genauso wichtig im Betrieb. Hierbei geht es vor allem um eine sinnvolle Verteilung der Ressourcen, sowie die Einhaltung von Regeln. [3, 4] Hierunter fällt auch die Analyse der Sicherheit eines Netzwerkes. Findet also z. B. nicht erwünschter Nachrichtenverkehr statt, oder werden unerlaubte Dienste benutzt oder angeboten?

Das wichtigste Werkzeug hierfür ist die Paketanalyse.

2.1 Was ist Paketanalyse und wo sind die Grenzen

Wie funktioniert Paketanalyse? Im folgenden soll ein grober und kurzer Überblick gegeben werden, was Pakete sind, welche Informationen sie enthalten und wie man diese auswerten kann.

Der Verkehr in großen Netzwerken, wie dem Internet, wird in Pakete unterteilt. Diese Pakete enthalten, neben den eigentlichen zu übermittelnden Daten, Informationen über den Anfrager (Quell IP), den Empfänger (Ziel IP) und den benutzen Service (Port).

Somit ist es z. B. recht einfach zu erkennen, wenn ein Benutzer die Website (Port 80) von Google.de (IP 66.249.93.104) aufruft, oder mittels eines Emailprograms Emails über web.de verschickt (IP 217.72.192.157, Port 25) Für viele Applikationen und Dienste gibt es einen Standardport über den dieser einfach zu identifizieren und zu finden ist. [5]

Obwohl man auf den ersten Blick damit schon eine Menge Informationen in der Hand hat, um ein aussagekräftiges Bild über den Verkehr zu bekommen, wird dies bei näherer Betrachtung sehr weit eingeschränkt. So sind einige Ports mehrfach belegt, z. B. der Port 4000 für das Chatprotokoll „ICQ“ und die Fernadministration „Remote Anything“. Noch weitreichender ist die Tatsache, dass viele Applikationen, wie z. B. Skype, Standard Ports anderer Anwendungen mit benutzen, um somit an Firewalls und den damit verbundenen Regeln und Einschränkungen, vorbei zu kommen [6]. Werden Ports sogar dynamisch ausgehandelt, die Daten verschlüsselt übertragen und die Dienste nicht direkt in Anspruch genommen, sondern z. B. Proxys benutzt, sinkt die Aussagekraft des durch klassische Paketanalyse gewonnenen Bildes weiter.

„Unconstrained Endpoint Profiling“ (UEP) verfolgt daher einen anderen Ansatz.

3 Der Ansatz von UEP

In diesem Absatz wird das grundsätzliche Vorgehen von UEP kurz erläutert und die daraus resultierenden Voraussetzungen genannt.

Wie im vorigen Abschnitt gezeigt wurde, sind die Informationen, die aus Paketen gewonnen werden können nur begrenzt aussagefähig. Außerdem sind Paket-traces zu vielen Endpunkten nur schwer oder gar nicht zu bekommen, da man diese über die Internetprovider direkt beziehen müsste. Und auch wenn man Zugriff auf Traces eines Punktes hat, so kann man nur Informationen über Pakete gewinnen, die auch diesen Knotenpunkt passiert haben. Daher versucht UEP ohne Zugriff auf Paket-traces zu agieren. Diese können zwar zur Verbesserung der Analyse herangezogen werden, sind jedoch nicht notwendig um aussagekräftige Ergebnisse zu bekommen. Anders als eine

pure Auswertung von Paketen, holt sich UEP die Informationen nämlich aus dem gesamten Internet. Anstatt sich also auf die Daten eines Paketes zu verlassen, welche sich letztlich auf Start-Host, Ziel-Host und Port beschränken, wird die Google Suchmaschine [2] zur Identifizierung der Hosts und den benutzen Services verwendet.

Es wird also Vorausgesetzt, dass eine ausreichende Menge von Hosts Ihre laufenden Services im Internet publizieren und der Zugriff und die Benutzung dieser Dienste ebenfalls einsehbar ist. Als Möglichkeiten zur Informationsgewinnung zu Services zieht UEP folgende Quellen in Betracht:

- *Web Logs:* Fast jeder Webserver erstellt Zugriffsdateien und bereitet diese mit Programmen wie z.B. AWStats, Webalizer auf. Aus den so generierten Berichten sind Informationen über den Client wie IP-Adresse, Betriebssystem, Browser und Zugriffszeiten abrufbar.
- *Proxy Logs:* Öffentliche Proxy Dienste, wie z.B. Squid, erstellen ähnliche Berichte aus Ihren Zugrifssdaten und können diese auf einer Website zur Verfügung stellen.
- *Foren:* In Foren werden Beiträge oft mit Benutzernamen, Datum und IP-Adresse versehen. Außerdem werden auf diesem Weg oft Links zu verschiedenen Inhalten auf ftp, http oder streaming shares verteilt.
- *Block Listen:* Im Internet existieren zu fast jedem Service auch öffentlich zugängliche Blocklisten. Beispiellisten sind „spamlists, gaming abuse lists, forum spammers, etc. Diese beinhalten IPs von Denial of service Angreifern oder Spammern usw.
- *Server Listen:* Um die Dienste eines Servers nutzen zu können, muss die IP zum assoziierten Service öffentlich zugänglich sein. Beispiele für solche Angebote sind Domain Name Servers, Domain databases, gaming Server, Mail Servers, etc.
- *P2P Eintrittspunkte:* In P2P-Netzen, wie z. B. eMule, gnutella, edonkey, kazaa, etc., müssen Knoten bekannt sein, damit neue Teilnehmer beitreten können. Diese Eintrittspunkte werden oft auf frei zugänglichen Webseiten publiziert.

Alle diese Quellen werden von UEP über die Google Suchmaschine abgerufen. Wie genau diese ausgewertet werden wird im nächsten Abschnitt erläutert. Auf mögliche Einschränkungen bei der Verfügbarkeit der Daten werde ich in Kapitel 7 noch eingehen.

4 Methodisches Vorgehen

Wie genau UEP die verschiedenen Quellen verwendet um darüber wiederum das Verhalten von Endpunkten (IP-Adressen) zu identifizieren wird nun im folgenden beschrieben.

Auf der untersten Ebene von UEP wird eine IP als Suchanfrage an Google geschickt. Aus den verschiedenen URLs des Suchergebnisses werden nun Kennzeichnungen (tags) für diese IP extrahiert. Zu einer Suchanfrage können am Ende verschiedenste tags stehen. Möglich sind z. B. *mail server, forum user, etc.*

Auf höherer Ebene werden Eigenschaften einer IP anhand von aus dem Internet gewonnenen Informationen festgelegt. Diese Eigenschaften sind um so zuverlässiger, je mehr Suchergebnisse es zu einer Anfrage gibt und je größer die Anzahl der bestätigten Dienste einer IP sind (z. B. ein *mail server*).

In Abbildung 1 wird das Vorgehen anschaulich zusammengefasst. Es besteht im einzelnen aus den drei Schritten: *(1)* Regelgenerierung, *(2)* Klassifizierung und *(3)* Kennzeichnung. Diese werden nun im einzelnen erläutert.

4.1 Regelgenerierung

Am Ende der Regelgenerierung ist es möglich, anhand von Schlüsselworten die Funktion (Klasse) einer Website zu bestimmen.

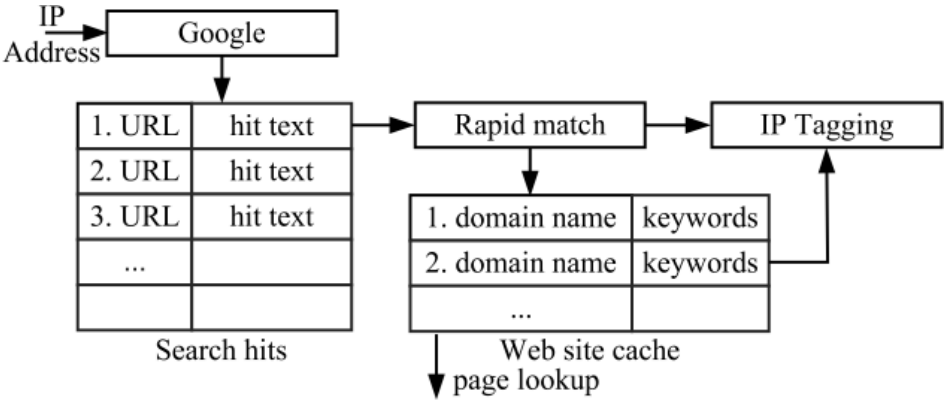

Abbildung 1: Web-basierte Endpunkt Profilbildung

Hierzu werden zunächst Schlüsselworte aus Ergebnisse von Google-Anfragen gefiltert und sortiert. Diese Keywords werden dann einmalig manuell interpretiert und einer Klasse zugeordnet. Die dadurch entstehende Abbildung wird anschließend genutzt, um die Funktion einer Website zu bestimmen. Einen Auszug dieser Relationen sieht man in Tabelle 1.

Schlüsselwörter	Klasse	Kennzeichnung
{ 'ftp' | 'webmail' | 'dns' | 'email' | 'proxy' | 'smtp' | 'mysql' | 'pop3' | 'mms' | 'netbios' }	Protocols and Services	<protocol name> server
{ 'yahoo' | 'gtalk' | 'msn' | 'qq' | 'icq' | 'server' | 'block' }	Chat servers	<protocol name> server
{ 'torrent' | 'emule' | 'kazaa' | 'edonkey' | 'announce' | 'tracker' | 'xunlei' | 'limewire' | 'bitcomet' | 'uusee' | 'qqlive' | 'pplive' }	p2p node list	<protocol name> p2p node
'mail server'	Mail server list	mail server
'mail server' & { 'spam' | 'dictionary attacker' }	Malicious mail servers list	mail server [spammer] [dictionary attacker]

Tabelle 1: Schlüsselwörter - Klassen - Kennzeichnungen

4.2 Klassifizierung

Die zuvor generierte Tabelle wird in diesem Schritt genutzt, um automatisch die Klasse einer Seite zu bestimmen.

Dazu werden die Schlüsselworte einer Website aus dem Namen der Domain, unter der sie zu erreichen ist, und aus dem Text der Seite selbst extrahiert. Durch die vorher geleistete Zuordnung von Keywords und Klasse ist somit die Website einer Klasse zugeordnet.

Diese Zuordnung von Schlüsselworten und Klasse wird für jede Website gespeichert (*Web site cache*) um sie im letzten Schritt zur Kennzeichnung von IP-Adressen zu verwenden.

4.3 Kennzeichnung

Der zuvor aufgebaute *Web site cache* wird nun verwendet, um IP-Adressen zu kennzeichnen.

Ähnlich wie beim klassifizieren der Webseiten, kann die zugehörige URL der IP Adresse schon ausreichend Informationen liefern, um daraus einen tag zu erstellen. In den meisten Fällen reicht dies jedoch nicht aus. Dann wird die IP im zuvor aufgebauten cache gesucht. Die zu der IP Adresse gefundenen Schlüsselwörter und Klassen bestimmen dann die Kennzeichnung der IP. Zur Verdeutlichung möchte ich ein einfaches Beispiel heranziehen.

Die Website projecthoneypot.org ist im cache mit der Klasse 'malicious mal servers' eingetragen. Eine IP die nun den cache durchläuft und mit den Schlüsselworten 'spam' und 'mail server' auf projecthoneypot.org gefunden wird, bekommt die Kennzeichnung 'mail server' und 'spammer'. Wird lediglich 'mail server' gefunden, ist die Kennzeichnung ebenfalls nur 'mail server'.

Da der cache sehr viel mehr Websiten, Schlüsselworte und daraus resultierende Informationen bzw. Kennzeichnungen enthält, kann eine IP Adresse am Ende eines Durchlaufs auch mit sehr vielen Kennzeichnungen versehen sein. Diese geben dann ein gutes Bild der benutzten und angebotenen Dienste dieser IP wieder.

5 Ergebnisse der Profilbildung von Endpunkt

In diesem Kapitel wird die zuvor vorgestellte Methode nun angewendet um das Benutzungsverhalten verschiedener Regionen zu untersuchen.

Zuerst muss die Frage beantwortet werden, wie Internetaktivitäten einer bestimmten Region zugeordnet werden können. Der einfachste Weg hierbei ist die IP-Adress-Räume verschiedener großer Internet-Service-Provider (tier-1) zu nehmen. Genau dieser Ansatz wurde in „Unconstrained Endpoint Profiling“ auch gewählt. Ein Auszug der untersuchten IP-Adress-Räume ist in Tabelle 2 aufgelistet.

Asien (China)	S. Amerika (Brasilien)	N. Amerika (US)	Europa (Frankreich)
XXX.39.0.0/17	XXX.96.128.0/17	XXX.0.0.0/11	62.147.0.0/16
XXX.83.128.0/17	XXX.101.0.0/17	XXX.160.0.0/12	81.56.0.0/15
XXX.69.128.0/17	XXX.103.0.0/17	XXX.70.0.0/16	82.64.0.0/14
XXX.72.0.0/17	XXX.163.0.0/17	XXX.168.0.0/14	

Tabelle 2: Untersuchte Netzwerke

Die Anonymisierung wurde vollzogen, da diese Anbieter für die in Kapitel 6 abgehandelte weitere Auswertung Daten zur Verfügung gestellt haben.

Aus jeder dieser so definierten Regionen wurden dann ca. 200.000 zufällige IP-Adressen ausgewählt und mittels UEP analysiert. Die Auswertung hat zum Teil sehr interessante Einblicke in regionale Internetnutzung offenbart.

5.1 Trends in der Benutzung von Systeme, Programme und Dienste

Mit dem Verfahren konnten eine Menge Statistiken erhoben werden, aus denen ich nun einige, meiner Meinung nach interessante, Aspekte hervorheben möchte

.

Aus der linken Diagramm von Abbildung 2 ist ersichtlich, dass in Frankreich das Betriebssystem debian öfter im Einsatz ist, als Windows. Dieser Trend lässt sich auch für

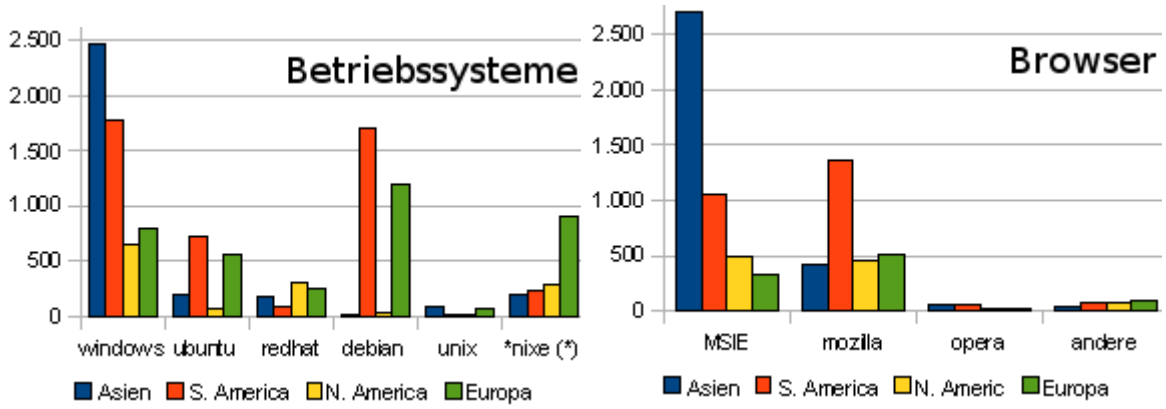

Abbildung 2: Betriebssystem / Browser Nutzung

Brasilien erkennen. Ähnliches lässt sich über die Browsernutzung sagen (rechtes Diagramm von Abbildung 2). Diese korreliert mit dem eingesetzten Betriebssystem, da bei Linux-Systemen i.d.R. ein Mozilla basierter Browser vorinstalliert installiert ist, während dies unter Windows Systemen der Internet Explorer ist.

Sehr deutlich ist auch das Ergebnis der besuchten Webseiten. Google ist hier in allen Regionen die meist besuchte Webseite. Ebenso ist die weltweite Popularität von Wikipedia[1] zu erkennen. Es liegt überall auf Platz zwei, ausser in China, wo es sehr weit hinten liegt (Platz 9).

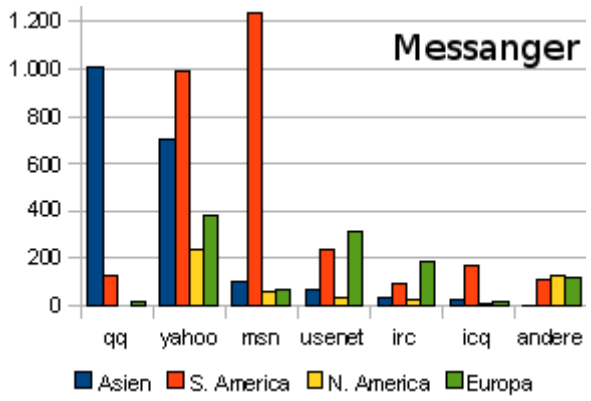

Abbildung 3: Messanger Nutzung

Die regional sehr unterschiedliche Beliebtheit von Messangern lässt sich aus dem Diagramm in Abbildung 3 sehr gut ablesen. So führt das hierzulande völlig unbekannte QQ[2] die Liste im asiatischen Raum an, während z. B. im brasilianischen Raum msn sehr oft benutzt wird, was im europäischen Raum wiederum nur durchschnittlich oft eingesetzt wird.

Die in Kapitel 6 vorgenomme Auswertung findet für einige der hier genannten Trends eine plausible Erklärung und geht auf weitere Vergleichskriterien ein.

[1] http://www.wikipedia.org/
[2] http://www.qq.com/ (deutsche Erläuterung: http://de.wikipedia.org/wiki/QQ)

5.2 Lokalität des Internetverkehrs

Neben regional unterschiedlichem Verhalten in der Benutzung von Systemen, Programmen und Diensten wird noch die Lokalität des Internetverkehrs untersucht. Hierbei geht es darum, zu untersuchen, von wo Anfragen einer Region überwiegend beantwortet werden.

Hierzu werden zunächst Paare von IP-Adressen gebildet, welche miteinander kommunizieren. Diesen IP Adressen wird dann mit Routeviews[3] das jeweils zugehörige Autonome System (AS)[4] zugeordnet und die AS-Entfernung zwischen Sendern und Empfängern gemessen, das heißt wie viele verschiedene AS eine Antwort auf Ihrem Weg durch das Netz durchlaufen hat. Dann wurden die Ländercodes der verschiedenen AS bei den entsprechenden Routing-Register Datenbanken (ARIN, RIPE, APNIC, LACINIC) herausgefunden. Die Ergebnisse sind in Abbildung 4 zu sehen.

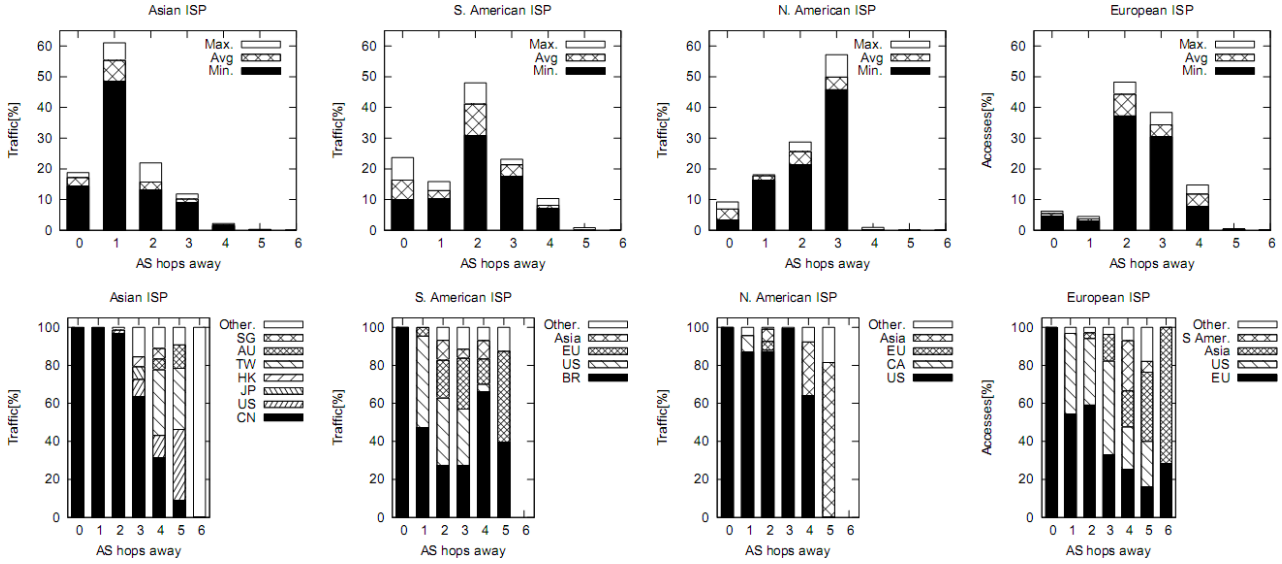

Abbildung 4: Lokalität von Internetverkehr

Hieraus ist z. B. zu erkennen, dass die meisten Anfragen im asiatischen Raum einen „AS-Sprung“ gemacht haben und dass nahezu alle Ländercodes dieser Entfernung ebenfalls in Asien liegen. Für Nordamerika ergibt sich ein ähnliches Bild.

Das heißt, sowohl in Asien, als auch in Nord Amerika ist der Internetverkehr sehr lokal, während in Europa und Südamerika eine Menge Verkehr nach aussen geht. Wie aussagekräftig diese gemessene Lokalität ist, werde ich in Kapitel 7 kurz erläutern.

6 Vergleich und Auswertung der Ergebnisse

Es soll nun versucht werden, einen Vergleich für die erzielten Ergebnisse zu bekommen.

Hierzu wurden zum einen gemessene Trends anhand von weiteren Quellen überprüft. So konnte zum Beispiel die beobachtete Benutzung von Linux in Frankreich und Brazilien anhand von Nachrichtenquellen bestätigt werden. So setzen in diesen Ländern

[3] http://www.routeviews.org/

[4] Ein Autonomes System (AS) ist eine Ansammlung von IP-Netzen, welche als Einheit unter einer gemeinsamen Verwaltung stehen (z. B. durch Internet Service Provider, internationale Firmen, Universitäten). Diese AS sind untereinander verbunden und bilden so das Internet.

Behörden, Ämter und Schulen vermehrt auf Linux als Betriebssystem [7, 8, 9] Eine Erklärung für die wenigen Aufrufe von Wikipedia aus dem Asiatischen Raum lässt sich sicherlich auf die strikte Kontrolle des chinesischen Internetverkehrs und daraus resultierende Blockierungen zurückführen [10, 11]. Ausserdem wurden Packet-Traces analysiert und neben die Ergebnisse von UEP gelegt. Die Packet-Traces wurden mit der in [3] vorgestellten Methode (BLINC)[5] ausgewertet und dann mit den Ergebnissen von UEP verglichen. Ein Auszug der Ergebnisse dieses direkten Vergleiches ist in Tabelle 3 zu sehen.

Klassen	**BLINC**	**UEP**
Chat	0,398 %	3,38 %
Browsing	23,16 %	44,70 %
P2P	4,72 %	11,31 %
Gaming	0,14 %	0,15 %
Malware	2,93 %	2,3 %
Streaming	0 %	0,18 %
Mail	0 %	1,58 %
FTP	0 %	0,1 %
zugeordnet	**29,60** %	**62,14** %

Tabelle 3: Netzverkehr Klassifizierung

Es wurde untersucht, wie gut sich Netzverkehr mit den beiden Methoden in Klassen einteilen lässt. Es ist deutlich zu erkennen, dass mittels UEP sehr viel mehr Verkehr zugeordnet werden kann, als mit einer reinen Pakatanalyse. Letztlich ergibt sich daraus der Schluß, dass die mittels UEP gewonnenen Erkenntnisse durchaus fundiert sind und ein gutes Bild des untersuchten Netzverkehrs wiedergeben.

7 Diskussion

In diesem Abschnitt möchte ich auf Einschränkungen von UEP hinweisen und, falls vorhanden, Möglichkeiten diese zu minimieren, bzw. das Verfahren gezielt zu erweitern. Daraus resultierend werde ich noch etwas zur Anwendbarkeit des Verfahrens sagen.

7.1 Einschränkungen bei der Erkennung von Trends

Wie in Kapitel 3 und 4 beschrieben, bildet die Google Suchmaschine das Fundament von UEP zur Erkennung von Trends in der Benutzung von Systemen, Programmen und Diensten. Hieraus ergeben sich gleich mehrere mögliche Probleme. So werden nicht nur Werbeeinblendungen (Google Ads) bei Google regional abhängig eingeblendet, sondern auch die Suchergebnisse selbst sind nicht immer gleich. Diese sind unter anderem davon Abhängig, an wen die Anfrage abgesetzt wurde (an google.com oder an google.de etc.), aber auch die IP des Anfragers (und dadurch sein Standort) wird hierbei mit ausgewertet. Abhängig davon werden bestimmte Ergebnisse gefiltert und mit anderer Priorität dargestellt. Dies dürfte sehr große Auswirkungen auf die Ergebnisse von UEP haben.

Um dieses Problem zu minimieren, sollten gleiche Anfragen von verschiedenen Orten durchgeführt werden und Ergebnisse anderer Suchmaschinenanbieter mit einbezo-

[5]BLINC ist eine Methode zur Analyse von Packet-Traces. Hierbei werden nicht alle mitgeschnittenen Pakete analysiert, sondern anhand von Eckdaten einzelner Pakete Aussagen über eine Menge von Paketen getroffen.

gen werden. Allerdings wird man das Problem hierdurch nur minimieren können, denn auch andere Suchmaschinen gehen ähnlich vor.

Ein von den Autoren aufgewiesenes Problem ist die Aktualität der Suchergebnisse. Darauf hat UEP keinen direkten Einfluss. Hier bestimmt wiederum Google, wie oft Webseiten nach Aktualisierungen durchsucht werden. Dies zieht vor allem Verzögerungen beim Einbeziehen neuer Informationen nach sich. Veraltete Ergebnisse könnten wiederum (in einem gewissen Rahmen) durch eine Erweiterung von UEP herausgefiltert werden. Als Beispiel wird die Einbeziehung des Datums bei der Auswertung von Forenbeiträgen genannt. Dies wird oft mit den Beiträgen veröffentlicht.. Hieran könnten veraltete Informationen erkannt und herausgefiltert werden.

Als ebenfalls sehr wichtig empfinde ich die Bewertung der Quellen, auf die UEP aufbaut. Diese sind in Kapitel 3 genauer erläutert. So möchte ich die öffentliche Zugänglichkeit zu vielen Proxy- bzw. Web-logs anzweifeln. Proxys werden oft genutzt um dadurch eine Anonymisierung der eigenen Zugriffe zu erreichen. Als viel genutze Proxy-Netzwerke seien hier „TOR“[6] und „Jap“[7] genannt. Daten über deren Benutzung sind z. B. nicht öffentlich zugänglich. Bei vielen Fällen, wo ein Zugriff auf solche Logs doch möglich ist, würde ich dies auf ein geringes Sicherheitsbewusstsein der Administratoren zurück führen. Ähnliches gilt z. B. für Foren, in denen man Zugriff auf die IPs der Mitglieder hat.

Von den Autoren ebenfalls angesprochen wurden dynamische IP-Adressen. So werden in der momentanen Umsetzung von UEP hauptsächlich Informationen von Servern, die auf statische IPs gehostet werden ausgewertet. Aktivitäten dynamischer IPs führen mitunter zu sehr vielen unterschiedlichen Kennzeichnungen ein und derselben IP. Meines Erachtens ist dies aber eher zu vernachlässigen, wenn man Profile von Endpunkten großzügig zusammenfasst, wie es auch in 5.1 gemacht wurde.

Eine interessante Erweiterung dürfte auch die von den Autoren angestrebte Auswertung anderer Alphabete sein. So werden die Schlüsselworte zur Regelgenerierung in 4.1 lediglich aus dem lateinischen Alphabet gebildet. gerade im asiatischen Raum dürften durch eine Ausweitung auf nicht-lateinische Alphabete die Ergebnisse noch verfeinert werden.

Zusammenfassend möchte ich die von den Autoren genannte Ausweitung des Systems auf andere Informationsquellen als die sinnvollste hervorheben. Als Beispiel sei der P2P-Verkehr genannt. So sind zwar die Eintrittspunke bekannt und durch Google auffindbar. Alles was dann passiert ist aber durch Google nicht mehr nachvollziehbar. Durch eine Einbeziehung weiterer Quellen könnten sich die Ergebnisse durchaus noch differenzierter zeigen.

7.2 Aussagekraft der Lokalitätsmessung

Die Autoren nennen selbst eine sehr wichtige Einschränkung, wenn es um die Korrektheit der Lokalität des Internetverkehrs geht, wie er in Kapitel 5.2 dargestellt wurde. So kann mit der Methode zwar festgestellt werden, wie weit Anfragen durch das Internet laufen, dies sagt jedoch nicht zwingend etwas darüber aus, wie regional die Anfrage nun wirklich ist. Als Beispiel sei ein in Amerika gehosteter Server genannt, auf dem ein deutsches Forum betrieben wird. So wird jede Aktivität aus Europa in diesem Forum nicht als lokale Aktivität gewertet, obwohl es sich eigentlich um ein lokales Forum handelt. Auf diesen Umstand wird von den Autoren auch hingewiesen. So kann Europa und Südamerika einen sehr überregionalen Internetverkehr haben, wenn in China und Amerika mehr Server betrieben werden, als in Europa oder Südamerika. Hier sollten

[6] http://www.torproject.org/
[7] http://anon.inf.tu-dresden.de/

meines erachtens noch genauer untersucht werden, wie die Zusammenhänge zwischen inhaltlicher Lokalität und verkehrs Lokalität sind.

7.3 Anwendbarkeit von UEP

Die genannten Einschränkungen wirken sich direkt auf die Anwendbarkeit von UEP aus. Meiner Meinung nach stellt UEP einen durchaus interessanten Ansatz dar um sich eine Übersicht über Netzwerkverkehr auf einem sehr großem Level zu verschaffen. Das heißt als Werkzeug für einem Administrator eines kleinen Netzwerkes ist UEP nur bedingt einsetzbar. Um sich ein grobes Bild über großflächige Verhaltensweisen zu verschaffen ist es aber durchaus eine sinnvolle Ergänzung. So bieten die in 5.1 dargestellten Daten einen guten Ausgangspunkt um vertiefende Analysen über Benutzerverhalten, Trends in Programmen und Internet-Traffic-Lokalität zu machen.

8 Zusammenfassung

Zu Beginn wurde die Frage aufgeworfen, was Menschen im Internet machen. Nach einer genaueren Differenzierung der Fragestellung wurde gezeigt, warum sich vorhandene Methoden nur bedingt zur Beantwortung der Frage eignen. UEP wurde als neuer Ansatz vorgestellt und das genaue Vorgehen erläutert. Anschließend wurde das Verfahren angewandt um Trends in der Benutzung von Programmen, Diensten und besuchten Webseiten auf globaler Ebene aufzuzeigen. Die gewonnenen Erkenntnisse wurden anhand weiterer Quellen verifiziert und mit bereits vorhandenen Methoden veglichen. Hierbei zeigte sich die Stärke von UEP. So konnte mittels der vorgestellten Methode ein genaueres Bild erstellt werden, als dies bisher möglich war und das mit geringerem Aufwand. Zum Abschluß wurden die Ergebnisse kritisch hinterfragt, sowie mögliche Schwachpunkte des Verfahrens aufgezeigt. Es hat sich gezeigt, dass es eine Menge Ausbaumöglichkeiten gibt, wodurch sich die Ergebnisse noch verfeinern lassen. Dennoch hat man durch die Methode schon jetzt ein gutes Werkzeug in der Hand um die Frage nach Trends und lokalen Unterschieden in der Internetbenutzung zu beantworten.

Literatur

[1] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Unconstrained Endpoint Profiling (Googling the Internet). In Proceedings of ACM SIGCOMM 2008, Seattle, WA, August 2008.

[2] Google.com Suchmaschine: http://www.google.com

[3] T. Karagiannis, K. Papagiannaki, and M. Faloutsos. BLINC: Multilevel Traffic Classification in the Dark. In ACM SIGCOMM, Philadelphia, PA, August, 2005.

[4] L. Bernaille, R. Texeira, and K. Salamatian. Early Application Identification. In CONEXT, Lisboa, Portugal, December 2006.

[5] Standard Ports: http://www.iana.org/assignments/port-numbers (2008-12-15)

[6] Skype Port-Konflikt: http://support.skype.com/en_US/faq/FA528/Conflicts-with-applications-such-as-Apache-or-IIS-working-on-port-80-443 (2008-12-15)

[7] http://www.brazzil.com/2004/html/articles/mar04/p107mar04.htm (2008-12-22)

[8] http://www.linuxinsider.com/story/35108.html (2008-12-22)

[9] http://www.redhat.com/about/news/prarchive/2007/frenchministry.html (2008-12-22)

[10] http://www.heise.de/newsticker/meldung/print/48343 (2008-12-15)

[11] http://www.heise.de/newsticker/meldung/print/113650 (2008-12-15)