Novel techniques for monitoring network network traffic at the flow level

Author(s):
Glatz, Eduard

Publication Date:
2013

Permanent Link:

https://doi.org/10.3929/ethz-a-010025541 →

Diss. ETH No. 21466
TIK-Schriftenreihe Nr. 140

Novel Techniques for Monitoring Network Traffic at the Flow Level

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by

Eduard Glatz

Dipl. El.-Ing. ETH
born July 15, 1955
citizen of Zurich and Basel

accepted on the recommendation of
Prof. Dr. Bernhard Plattner, examiner
Dr. Xenofontas Dimitropoulos, co-examiner
Prof. Dr. Björn Scheuermann, co-examiner
Dr. Walter Willinger, co-examiner

2013

Abstract

Research in Internet measurement provides us with new ways to understand, operate and improve the Internet. Learning from network traffic data requires a well-chosen set of analysis techniques. We envision a rich toolbox available for this task, and delve into novel techniques and their application on large data sets to extend the choice of analysis schemes. In particular, we focus on traffic data at the network level that is readily available from commercial routers in the form of flow metadata (e.g. NetFlow) to enable analyzes of ever growing traffic volumes with low demands on the measurement infrastructure.

This thesis consists of two major parts.

In a *first part*, we explore a promising approach to study unsolicited traffic without the need to reserve unpopulated IP address ranges to this task, as has been done in the past. Our approach is to study one-way traffic, i.e., packets that never receive a reply in live networks. We introduce a novel scheme to classify one-way traffic at the flow level into interpretable classes. We validate this scheme based on a data set that we prepare using all informative details available from packet data (e.g. header and payload contents). We use our classifier to shed light on the composition of one-way traffic, and illustrate how the particular class of "Unreachable Services" can be used to passively detect network service outages by processing flow-level traffic data only. Moreover, to obtain a comprehensive view on one-way traffic, we

conduct a large-scale study covering eight years of traffic data leading to new insights about the evolution of this exotic piece of traffic over time and space.

In *part two*, we present novel visualization methods following the well-known adage "A picture is worth a thousand words". In particular, we tackle the problems of how to summarize data to extract the most relevant information from big data sets, and how to visualize this information in an easy interpretable way. We envision a top-down workflow that in a first step identifies probably hidden patterns in a data set captured from a potentially large network, followed by a second step that involves a closer inspection of the traffic of individual end systems or subnets. Specifically, we use frequent itemset mining to obtain a list of most relevant patterns from the traffic data of a network that we then visualize through hypergraphs. Then we make use of a graph representation and a domain specific summarization scheme, which is based on the characteristics of typical host roles (e.g. client, server, P2P) to provide a quick overview of what roles a host assumes and what applications it runs. We demonstrate the usefulness of our approach by using proof-of-concept implementations in a number of illustrative case studies.

Kurzfassung

Forschung im Gebiet der Internet-Verkehrsdatenanalyse zeigt uns neue Ansätze, um das Internet zu verstehen, zu betreiben und zu verbessern. Der Gewinn neuer Erkenntnisse aus Verkehrsdaten bedingt jedoch den Einsatz gut ausgewählter Analysetechniken. Unser Ziel ist die Bereitstellung eines reichhaltigen Instrumentariums zu diesem Zweck, weswegen wir neue Analysetechniken und ihre Anwendung auf grossen Datenbeständen zur Entwicklung dieses Instrumentariums erforschen. Im Speziellen fokussieren wir uns auf Flowdaten auf der Netzwerkschicht (z.B. NetFlow), die von kommerziellen Routern einfach zur Verfügung gestellt werden, um die stets wachsenden Verkehrsvolumina mit geringem Infrastrukturaufwand zu analysieren.

Diese Dissertationsschrift ist in zwei Hauptteile gegliedert.

Im *ersten Teil* erforschen wir einen vielversprechenden Ansatz zur Analyse von unangefordertem Verkehr, ohne dass wir dazu einen ungenutzten IP-Adressbereich reservieren müssen, wie das bisher gemacht wurde. Wir studieren das Phänomen des Einwegverkehrs, d.h., von Netzwerkpaketen, die in operativen Netzen keine Antwort erhalten. Wir führen ein neuartiges Klassifizierungsssschema ein, um Einwegverkehr auf der Flow-Ebene in interpretierbare Klassen einzuteilen. Wir validieren dieses Schema mittels zusätzlicher Detailinformationen (z.B. Rahmendaten aller Pakete, Nutzlastinhalte) die nur Paket-Verkehrsdaten liefern können. Wir benutzen unseren

Klassifizierer um die Zusammensetzung von Einwegverkehr sichtbar zu machen, und illustrieren die Nützlichkeit der speziellen Klasse “Unerreichbare Dienste” um Dienstausfälle ausschliesslich aufgrund von Flow-Verkehrsdaten passiv zu detektieren. Darüber hinaus führen wir eine umfangreiche Studie durch, in der wir Einwegverkehr aus einem Zeitraum von acht Jahren analysieren und neue Einsichten in die Eigenschaften dieses exotischen Verkehrsanteils und seiner Entwicklung über Zeit und Raum hinweg gewinnen.

Im *zweiten Teil* der Arbeit stellen wir neue Visualisierungsmethoden vor, dem Sprichwort “Ein Bild sagt mehr als tausend Worte” folgend. Insbesondere befassen wir uns mit Methoden für die Verdichtung sehr umfangreicher Daten, die Extraktion relevanter Informationen und entwickeln Verfahren für die Interpretation und Visualisierung solcher Informationen. Unsere Methodik ist ein Top-Down Vorgehen, bei dem in einem ersten Schritt die potenziell vorhandenen, versteckten Muster einer Verkehrs-Datensammlung, die in einem möglicherweise sehr grossen Computernetz erfasst wurde, identifiziert werden, gefolgt von einem zweiten Schritt, bei dem Verkehrsdaten einzelner Endsysteme oder Subnetze im Detail inspiziert werden. Im Speziellen benutzen wir Frequent-Itemset Mining um eine Liste der relevantesten Muster zu extrahieren, die wir in Form von Hypergraphen visualisieren. Anschliessend benutzen wir einen domänenspezifischen Ansatz der Datenverdichtung, der auf den Eigenschaften typischer Endsystemrollen (Client, Server P2P) basiert, um selektiv den Netzwerkverkehr eines einzelnen Rechners überblicksartig in einem Graphen darzustellen, so dass seine Rollen und die von ihm ausgeführten Applikationen unmittelbar erkennbar sind. Mit Proof-of-Concept Implementierungen zeigen wir anhand von illustrativen Fallstudien die Nützlichkeit unseres Ansatzes auf.