Diss. ETH No. 18835
TIK-Schriftenreihe Nr. 113

Network Traffic Anomaly Detection and Evaluation

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by

Daniela Brauckhoff

Dipl. El.-Ing. TU Berlin
born December 23, 1978
citizen of Germany

accepted on the recommendation of
Prof. Dr. Bernhard Plattner, examiner
Prof. Dr. Kave Salamatian, co-examiner
Dr. Xenofontas Dimitropoulos, co-examiner
Dr. Andreas Kind, co-examiner

2010

Berichte aus der Kommunikationstechnik

Daniela Brauckhoff

Network Traffic Anomaly Detection and Evaluation

Shaker Verlag
Aachen 2010

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Zugl.: Zürich, ETH, Diss., 2010

Printed in Germany.

ISBN 978-3-8322-8977-5
ISSN 0945-0823

Shaker Verlag GmbH • P.O. BOX 101818 • D-52018 Aachen
Phone: 0049/2407/9596-0 • Telefax: 0049/2407/9596-9
Internet: www.shaker.de • e-mail: info@shaker.de

Abstract

A worldwide Internet usage growth rate of 380% over the period from 2000, the year of the dot-com bubble burst, until present indicates that Internet technology has become a cornerstone of our daily life. In the same period, cyber-crime has seen an incredible professionalization that makes sophisticated protection mechanisms for computers and networks an absolute necessity. Firewalls as the major defense of the last decade do not provide sufficient protection anymore. This fact has given rise to the development of intrusion detection and prevention systems. Traditional intrusion detection systems are reactive in the sense that they use a set of signatures, which grows at the same rate as new vulnerabilities are discovered, to identify malicious traffic patterns. Anomaly detection systems are another branch of intrusion detection systems that act more proactively. They derive a model of the normal system behavior and issue alerts whenever the behavior changes; making a subtle assumption that such changes are frequently caused by malicious or disruptive events. Anomaly detection has been a field of intensive research over the last years as it poses several challenging problems.

In this thesis, we address three of these challenges. When working with large-scale network data from possibly multiple routers, the curse of dimensionality considerably complicates the problem of anomaly detection. Principal component analysis has been proposed to deal with it. However, as subsequent

work has discovered several deficiencies in the proposed PCA-method, there is room for improvement. A second challenge stems from the underlying assumption of anomaly detection mentioned above, which, unfortunately, does not always hold in practice. As a direct consequence of this circumstance, users are often overwhelmed with false alarms. To cope with high false alarm rates, one could either try to reduce the number of false alarms, or one could try to minimize the time that is required for resolving an alarm. This is where we see the largest discrepancy between research and practice, as the false alarm problem is broadly ignored by the scientific community. Finally, when a research field such as anomaly detection has reached a certain degree of maturity a sound evaluation of the proposed methods should be done. The major challenge with regard to evaluation is due to fact that there are practically no labeled real-world datasets available.

Our contributions are the following. In the first part of this thesis, we revisit the PCA-method and its underlying assumptions. We find that the assumption of independence between measurement points is not given as network traffic statistics typically exhibit strong temporal correlation. Therefore, we extend the PCA-method to stochastic processes and include the temporal as well as spatial correlation in the model. With our extended method we achieve an improvement in accuracy of up to 20 percent. In the second part of this thesis, we address the false alarm problem. We introduce a method that uses histogram-based anomaly detectors and association rules to help administrators with the identification of anomalous flows and event root causes. With our approach we are able to reduce the time for alarm resolution from typically one hour to a few minutes. The third part of this thesis describes several realistic anomaly models for simulation that we have derived directly from flow traces. Moreover, we introduce FLAME, a tool for anomaly injection into real-world traces, which has been used by several researchers for assessing the false negative rates of their algorithms.

Kurzfassung

Die Internet-Nutzung ist seit dem Jahr 2000, in dem die Dotcom-Blase platzte, weltweit um 380% gestiegen. Dies ist ein klarer Indikator dafür, dass das Internet ein fester Bestandteil unseres täglichen Lebens geworden ist. Gleichzeitig hat sich im Bereich Cyber-Kriminalität eine erstaunliche Professionalisierung gezeigt, welche ausgefeilte Schutzmechanismen für Computer und Netzwerke unerlässlich macht. In den Neunzigern boten Firewalls noch ausreichenden Schutz vor Angreifern. Da dies heutzutage nicht mehr der Fall ist, wurden Intrusion-Detektions- und Präventionssysteme entwickelt. Traditionelle Intrusion-Detektionssysteme bieten reaktiven Schutz durch Signaturen, welche manuell für jede neu entdeckte Schwachstelle und Attacke entwickelt werden müssen. Eine andere Art von Intrusion-Detektionssystemen sind Anomalie-Detektionssysteme. Diese erstellen proaktiv ein Profil für das normale Verhalten eines Systems und melden einen Alarm, sobald das Verhalten vom gelernten Profil abweicht. Dieser Vorgehensweise liegt die subtile Annahme zugrunde, dass Änderungen im Verhalten eines Systems oft von Attacken oder Ausfällen verursacht werden. Anomaliedetektion war ein zentrales Forschungsthema der letzten Jahre und ist es bis heute, da es uns vor einige sehr herausforderne Probleme stellt.

In dieser Arbeit addressieren wir drei spezifische Herausforderungen. Da Modellierungsprobleme oft eine Vielzahl von Dimen-

sionen involvieren, wurden effiziente Methoden entwickelt, um die resultierende Komplexität zu handhaben. Eines dieser entwickelten Verfahren, die Hauptkomponentenanalyse, wurde erfolgreich auf Anomaliedetektionsprobleme in Netzwerken angewendet. Wie weitere Arbeiten zu diesem Thema jedoch gezeigt haben, sind Verbesserungen der Methode und ihrer Anwendung möglich und nötig. Eine zweite Herausforderung resultiert aus der oben genannten Annahme, welche der Anomaliedetektion zugrunde liegt. Da diese Annhame in der Praxis nicht immer erfüllt ist, werden Benutzer von Anomalie-Detektionssystemem oft von Fehlalarmen überschwemmt. Um dieses Problem zu lösen, könnte man versuchen entweder die Anzahl der Fehlalarme zu reduzieren oder die Zeit, welche benötigt wird, um einen Alarm zu beheben. Leider wird das Problem der Fehlalarme in der Forschungsgemeinschaft weitgehend ignoriert. Wenn ein Forschungsgebiet, so wie das der Anomaliedetektion, einen gewissen Reifegrad erreicht hat, sollte eine rigorose Evaluierung der entwickelten Methoden durchgeführt werden. Die Herausforderung hier ist, dass praktisch keine gekennzeichneten reellen Datensätze zur Evaluierung vorhanden sind.

In unserer Arbeit erarbeiten wir die folgenden Lösungen. Im ersten Teil dieser These, betrachten wir die Haptkomponentenanalyse und die ihr zugrunde liegenden Annahmen erneut. Dabei stellen wir fest, dass die Annahme der Unabhängigkeit zwischen einzelnen Messpunkten nicht erfüllt ist, da Netzwerkstatistiken oft stark zeitlich korreliert sind. Wir erweitern daher die Hauptkomponentenanalyse, um sie auf stochastische Prozesse anwenden zu können, indem wir sowohl die zeitliche als auch die räumliche Korrelation ins Model einfliessen lassen. Mit unserer erweiterten Methode errichen wir eine Verbesserung der Genauigkeit um bis zu 20 Prozent. Im zweiten Teil dieser Arbeit, addressieren wir das Fehlalarmproblem. Wir entwickeln eine Methode, welche Histogramm-basierte Detektoren und Assoziierungsregeln nutzt, um Administratoren bei der Aufklärung eines Alarms zu

unterstützen. Mit unserem Ansatz kann die benötigte Zeit zur Aufklärung eines Alarms von typischerweise bis zu einer Stunde auf wenige Minuten reduziert werden. Der dritte Teil unserer Arbeit beschreibt mehrere Anomaliemodelle, welche wir aus vorhandenen Netzwerkdaten extrahiert haben. Weiterhin stellen wir das von uns entwickelte FLAME-Tool vor, welches es erlaubt Anomalien in vorhandene Netzwerkdaten zu injizieren und bereits von mehreren Forschungsinstitutionen eingesetzt wird.

Contents

List of Figures

List of Tables