Doctoral Thesis

# Detection, classification and visualization of anomalies using generalized entropy metrics

**Author(s):**
Tellenbach, Bernhard Martin

**Publication Date:**
2012

**Permanent Link:**

https://doi.org/10.3929/ethz-a-009795096 →

Diss. ETH No. 20929
TIK-Schriftenreihe Nr. 137

# Detection, Classification and Visualization of Anomalies using Generalized Entropy Metrics

A dissertation submitted to
ETH ZURICH

for the degree of
Doctor of Sciences

presented by

BERNHARD MARTIN TELLENBACH

Master of Science ETH
in Electrical Engineering and Information Technology
born June 30, 1979
citizen of Oberthal, BE, Switzerland

accepted on the recommendation of
Prof. Dr. Bernhard Plattner, examiner
Prof. Dr. Didier Sornette, co-examiner
Dr. Andreas Kind, co-examiner

2012

# Abstract

Today, the Internet allows virtually anytime, anywhere access to a seemingly unlimited supply of information and services. Statistics such as the six-fold increase of U.S. online retail sales since 2000 illustrate its growing importance to the global economy, and fuel our demand for rapid, round-the-clock Internet provision. This growth has created a need for systems of control and management to regulate an increasingly complex infrastructure. Unfortunately, the prospect of making fast money from this burgeoning industry has also started to attract criminals. This has driven an increase in, and professionalization of, cyber-crime. As a result, a variety of methods have been designed with the intention of better protecting the Internet, its users and its underlying infrastructure from both accidental and malicious threats. Firewalls, which restrict network access, intrusion detection systems, which locate and prevent unauthorized access, and network monitors, which oversee the correct functioning of network infrastructures, have all been developed in order to detect and avert potential problems. These systems can be broadly defined as either reactive or proactive. The reactive approach seeks to identify specific problem patterns. It uses models learnt from theory or practice to locate common dangers as they develop. The number of patterns applied grows as each new problem is encountered. Proactive methods work differently. They start defining an idealized model of the normal behavior of a given system. Any significant deviation from this model is assumed to be an aberrance caused by an external danger. However, this assumption may turn out to be incorrect, having actually not arisen from a disruption or a malicious act. Despite considerable improvements, the development of accurate proactive detection and classification methods is still an area of intense research. This is particularly true of methods fit for high speed networks. To cope with the huge amounts of data at hand, these methods utilize highly aggregated forms of data. Volume

measurements and traffic feature distributions such as the number of connections per time unit or the distribution of connection sources form their primary sources of information. Various methods have been developed to detect anomalous changes in these distributions. Among them, entropy based methods have become widely used, and demonstrate considerable success in both research and production systems. Nonetheless, there remain many challenges regarding the use of entropy.

In this thesis, we address three of these challenges. In high speed networks, packet sampling methods are widely employed to reduce the amount of traffic data measured. However, we possess no empirical data about how this affects the visibility of anomalies when using entropy or volume metrics. Another area where additional analysis is required is the value of entropy with regard to anomaly detection. A study published by Nychis *et al.* found that entropies of common traffic feature distributions correlate strongly with simple volume measurements. The authors use this to suggest that they therefore do not contribute much. However, their claims do not match the practical evidence furnished by the many successful applications of this method. The second issue is the characterization and visualization of changes in distributions. In high-speed networks, the sheer quantity of information involved makes the concise representation of changes in distributions essential. However, many of the most commonly used methods, such as the Shannon entropy, are hampered by their limited descriptive power. This stems from the fact that they capture change using a single number. Other methods, including histograms, suffer by the fact that their optimal use depends on parameters which differ across various types of change.

The third problem to consider is the way in which the detection and classification capabilities of entropy-based anomaly detectors can be improved. Existing systems do show good detection rates. They can even, to an extent, successfully classify the largest anomalies. However, there remains scope to refine their performance, specifically when dealing with small to medium sized anomalies. Furthermore, studies on distributed denial of service and port scan anomalies from malware point out that parameterized entropies such as the Tsallis entropy might be superior to non-parameterized entropies. However, how these preliminary results can be linked to arbitrary types of anomalies, as well as appropriate detection and classification systems, remains underexplored.

In this work we make the following contributions. We analyze the robustness of entropy in the presence of packet sampling. Based on traffic traces

from the outbreak of the Blaster and Witty worm, we find that entropy is not only robust but, depending on the traffic mix, might even lead to an improvement in the location of anomalies for sampling rates of up to 1:10,000. Next, we analyze whether the entropy of various traffic feature distributions provides valuable information for anomaly detection. We refute the findings of previous work, which reported a supposedly strong correlation between different feature entropies. Our core contribution is the *Traffic Entropy Spectrum (TES)*, a method for the compact characterization and visualization of traffic feature distributions. We also propose a refined version of the TES, which hones its capabilities with regard to anomaly classification. To demonstrate the descriptive power of the TES, we use traffic data containing real anomalies. Finally, we build the Entropy telescope, a detection and classification system based on the TES. We provide a comprehensive evaluation using three different detection methods, and one classification method. Our evaluation, based on a rich set of artificial anomalies combined with real traffic data, shows that the refined TES outperforms the classical Shannon entropy by up to 20% in detection accuracy and by up to 27% in classification accuracy.

# Kurzfassung

Das heutige Internet ermöglicht jederzeit und praktisch überall Zugriff auf eine schier endlos erscheinende Menge an Informationen und Dienstleistungen. Zahlen wie die Versechsfachung des via Internet erzielten Umsatzes des US Einzelhandels seit 2000 weisen deutlich auf dessen zunehmende Bedeutung für die Weltwirtschaft aber auch auf die damit verbundene wachsende Abhängigkeit hin. Neben erhöhte Anforderungen an das Management und Überwachung aufgrund der zunehmenden Komplexität der Infrastruktur, führte dies insbesondere auch zu einer Zunahme und Professionalisierung der Cyber-Kriminalität. In den letzten Jahren wurden deshalb verschiedenste Methoden entwickelt, um das Internet, seine Teilnehmer und die zugrunde liegende Infrastruktur besser vor mutwilligen aber auch unbeabsichtigten Störungen und Bedrohungen zu schützen. Dazu gehören Systeme wie Firewalls zur Beschränkung des Netzwerkzugriffs, Systeme zur Erkennung und Verhinderung eines unerlaubten Eindringens oder auch Systeme zur reinen Überwachung des korrekten Funktionierens einer Netzwerkinfrastruktur. Zur Erkennung und Vermeidung von Störungen und Bedrohungen gibt es grundsätzlich zwei Ansätze: Erstens, der auf Mustererkennung basierte reaktive Ansatz, der die Erkennung von in der Theorie oder Praxis bekannten Bedrohungen ermöglicht. Und zweitens, der proaktive Ansatz, der auf der Annahme basiert, dass jegliche Abweichung von einem spezifizierten normalen Verhalten eines Systems auf eine Bedrohung oder Störung hindeutet. In einer Analyse der Abweichung kann sich dann aber durchaus herausstellen, dass es sich weder um eine Bedrohung noch um eine Störung gehandelt hat.

Trotz einiger viel versprechender Ansätze ist eine präzise Erkennung und Klassifizierung mit proaktiven Methoden noch immer ein Gebiet intensiver Forschung. Dies gilt insbesondere auch für Methoden, die für den Einsatz in Hochgeschwindigkeitsnetzen geeignet sind. Um den riesigen Datenmengen

Herr zu werden, basieren die meisten dieser Methoden auf hochaggregierten Informationen. Dazu gehören primär Volumen- oder Verteilungsinformationen wie z.B. die Anzahl Verbindungen pro Zeit oder die Verteilung der Quell- und Zieladressen oder auch der Verbindungsdauer von Verbindungen. Eine Klasse von Methoden, die sowohl in der Forschung als auch in der Industrie mit Erfolg eingesetzt wird, identifiziert ungewöhnliche Veränderungen mit Hilfe der aus den Verteilungen der Quell- und Zieladressen und der Quell- und Zielports der beobachteten Verbindungen berechneten Entropiewerte. Trotz dieses Erfolgs gibt es aber noch viele offene Fragen und Herausforderungen.

In dieser Doktorarbeit adressieren wir drei dieser offenen Fragen und Herausforderungen. Die erste Herausforderung betrifft die Analyse der Auswirkungen von unvollständigen Messdaten. In Hochgeschwindigkeitsnetzen wird zur Reduktion der Systemlast oft nur ein Teil der effektiv über das Netzwerk fliessenden Datenpackete für eine Messung berücksichtigt. Bei zufälliger Wahl der gemessenen Datenpakete ist somit die Chance gross, dass die Zahl der nicht erfassten Verbindungen für Verbindungen, die nur aus wenigen Datenpaketen bestehen, grösser ist als für Verbindungen mit vielen Datenpaketen. Bis anhin ist unklar, wie sich dies bei der Verwendung von Entropie als Metrik auf die Sichtbarkeit von Anomalien auswirkt. Unklarheit besteht auch beim Nutzen von Entropie-basierten Metriken im Hinblick auf die Erkennung von Anomalien. Eine von Nychis *et al.* publizierte Studie stellte hierzu fest, dass Entropie kaum mehr Informationen liefert, als bereits in einfachen Volumenmessungen enthalten ist. Die bisherigen Erfolge mit Entropiemetriken stehen allerdings im Widerspruch dazu. Eine zweite Herausforderung stellt die Erfassung und Visualisierung von Veränderungen in Verteilungen dar. In Hochgeschwindigkeitsnetzen ist eine kompakte und mit Fokus auf Veränderung informative Erfassung und Darstellung von Verteilungen aufgrund der schieren Menge von Informationen von grosser Relevanz. Bisher verwendete Verfahren haben entweder nur eine beschränkte Beschreibungskraft, weil sie, wie die Shannon-Entropie, die Veränderung mittels einer einzigen Zahl beschreiben. Oder deren optimalen Erfassung hängt wie beim Histogramm primär von Parametern ab, die von der Veränderung selbst abhängig sind. Die dritte Herausforderung betrifft die Verbesserung der Erkennungs- und Klassifizierungsleistung von entropiebasierten Anomalie Detektoren. Existierende Systeme zeigen bei massiven Anomalien gute Detektions- und teilweise auch Klassifikationsleistungen. Für kleinere Anomalien ist ihre Leistung hingegen wenig erforscht. Studien zu Distributed Denial of Service Anomalien und Portscans von Malware weisen zudem auf die Überlegenheit von parametrisierten Entropien wie der Tsallis Entropie hin. Eine Ausweitung auf beliebige

Anomalien sowie die Frage nach passenden Detektions- und Klassifikationssystemen bleibt aber unbeantwortet.

In dieser Arbeit machen wir die folgenden Beiträge: Wir analysieren die Robustheit der Entropie beim Einsatz von Messstrategien, die für die Generierung der Verbindungsinformationen im Durchschnitt nur jedes n-te Paket berücksichtigen. Basierend auf dem Ausbruch des Blaster und Witty Wurms zeigen wir, dass Entropiemetriken robust sind und je nach Verkehrsmix und Anomalie sich deren Sichtbarkeit bis zu Abtastraten von 1:10,000 sogar verbessern kann. Ein weiterer Beitrag ist eine Analyse der Relevanz der Entropie von verschiedenen Verbindungsmerkmalen in Bezug auf die Anomaliedetektion. Wir widerlegen dabei eine Studie, die eine starke Korrelation zwischen verschiedenen Entropie- und Volumenmerkmalen fand. Unser wichtigster Beitrag jedoch ist die Entwicklung des *Traffic Entropy Spectrum (TES)*, eine auf der Tsallis Entropy basierende Methode zur kompakten Charakterisierung und Visualisierung von Verteilungen von Verbindungsmerkmalen. Wir ergänzen diesen Beitrag durch eine Verfeinerung des TES im Hinblick auf die Klassifizierung von Anomalien. Zur Demonstration der Beschreibungskraft des TES verwenden wir Verbindungsdaten mit echten Anomalien. Schliesslich bauen wir das Entropie-Teleskop, ein auf dem TES basierendes System zur Erkennung und Klassifizierung von Anomalien und liefern eine umfangreiche Evaluation basierend auf drei verschiedenen Detektionsmethoden und einer Klassifikationsmethode. Die Auswertung mit einer grossen Zahl an künstlichen Anomalien kombiniert mit realen Verkehrsdaten zeigt, dass der verfeinerte TES Ansatz der klassischen Shannon-Entropie bei der Detektion um bis zu 20% und bei der Klassifikation um bis zu 27% überlegen ist.