Big Data Analytics for Network Security

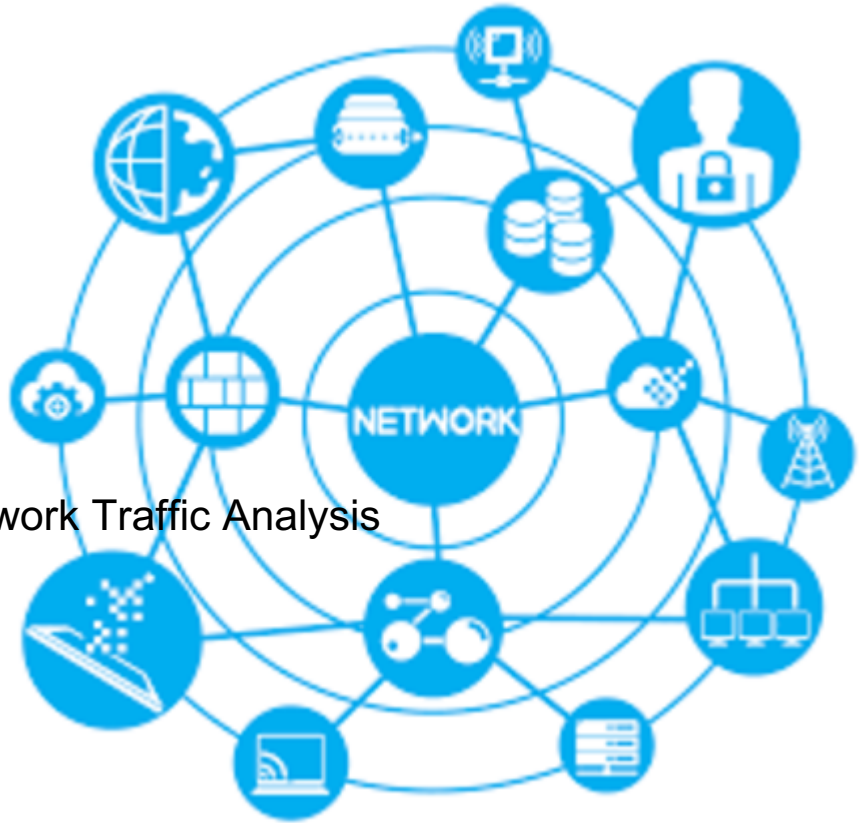
AIM-3 Presentation SoSe 17

Tam Tran & Seema Narasimha Swamy

Img src: https://insiderfinancial.com/airborne-wireless-network-inc-otcmkts-abwn-continues-its-trajectory

Agenda

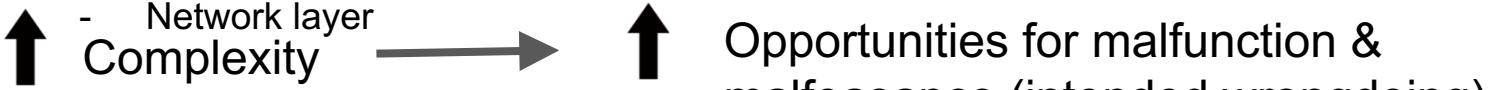
1. Intro to the Internet
2. Network Vulnerabilities and Risks
3. Forms of Defense
4. Anomalies
5. Data Preprocessing and Aspects of Network Traffic Analysis
6. Anomaly Detection Methods
7. Scalable Systems
8. Current Research / Challenges



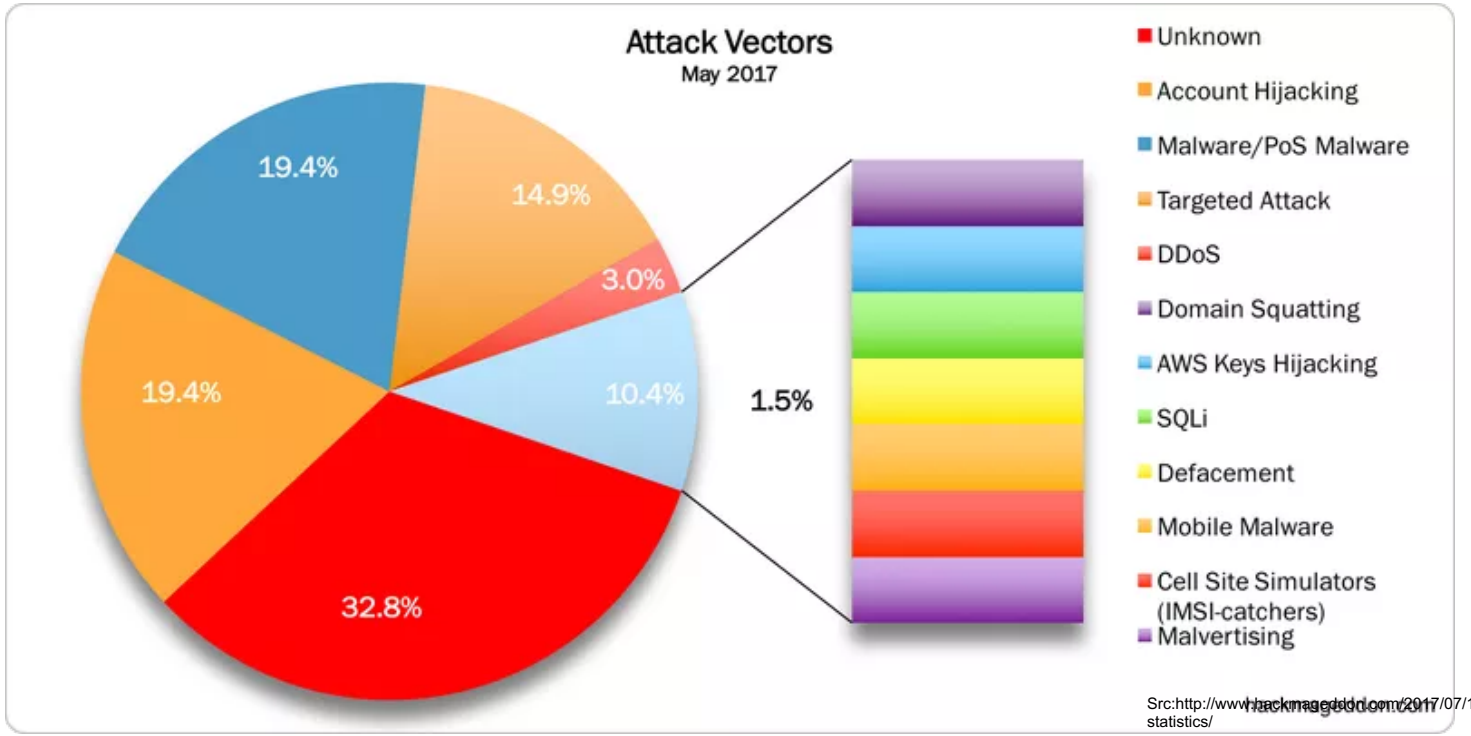
Img src: http://www.goboomtown.com/media/network-web-generic-02.png

Internet & Network Architecture

- Internet ~= big communication network
- Networks and networking functions ~= layered architectures
    - ISO/OSI model, TCP/IP model
- Many layers, very complex
    - Physical layer
    - Data-linking layer
    - Network layer
    - Transport layer
    - Session layer



Network Vulnerabilities



Src:http://www.hackmageddon.com/2017/07/13/may-2017-cyber-attacks-statistics/

Globally Cybercrime will costs businesses over $6 trillion by 2021

You are a security risk

- Internet users are at risk
- Computers and laptops
  - Open spam mail, potentially became a botnet
  - Your grandpa who connects to the Internet, but doesn't know a thing about Internet security
- Internet of Things devices
  - Webcams, Digital camera, DVD player, smart home device, Amazon Echo, etc etc

Img src: https://www.irishtimes.com/business/technology/tape-over-your-webcam-to-stop-people-watching-you-1.2742702

Network Vulnerabilities

+

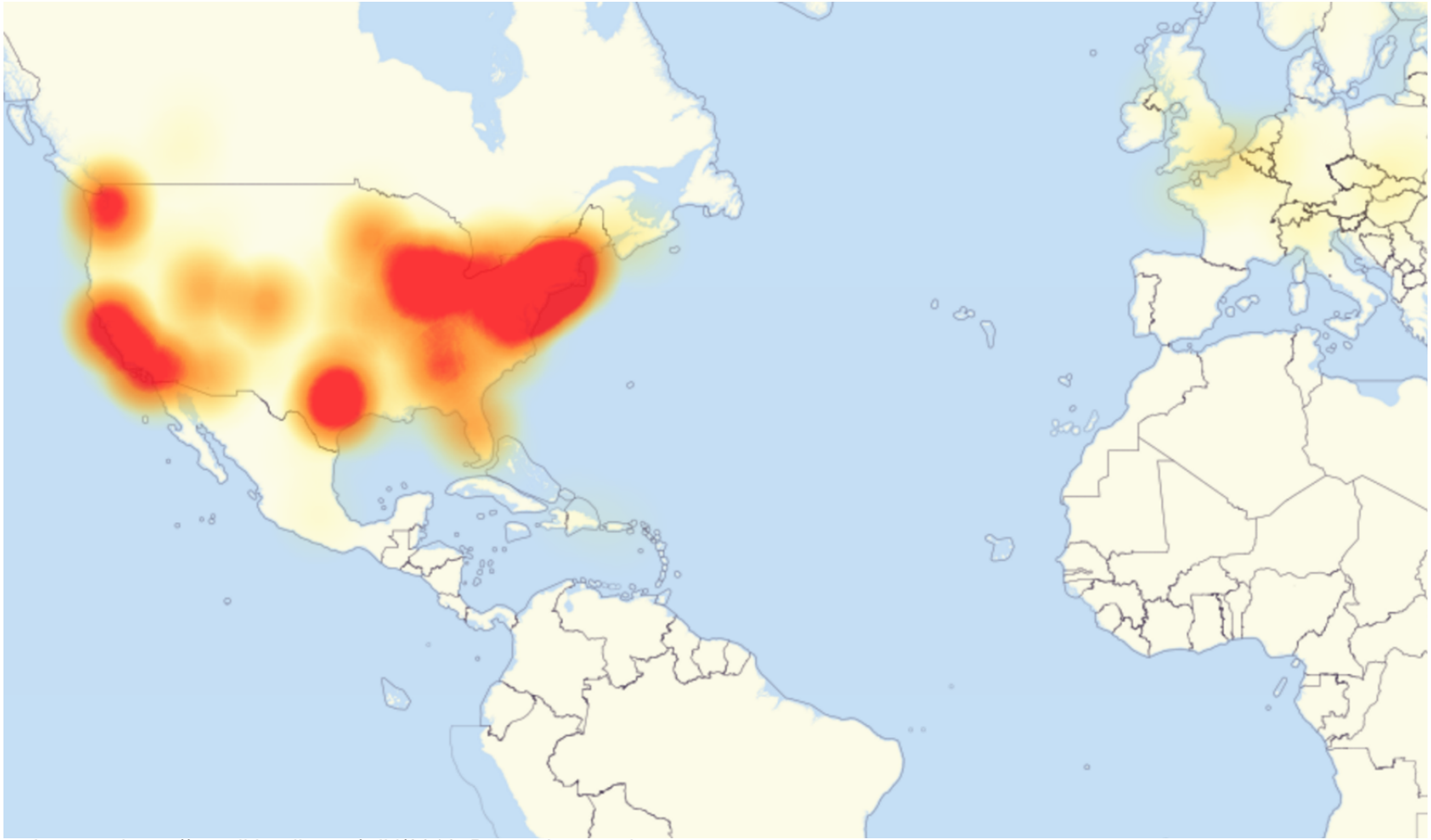
You

+

Limited server resources

=

DDoS Cyberattack of 2016



Img src: https://en.wikipedia.org/wiki/2016_Dyn_cyberattack

DDoS Cyberattack of 2016

- Largest in history
- Botnets included computers **and** IoT devices
    - Bigger army, bigger attack
    - Attack strength as much as 1.2 TB per second
- Took down big websites
    - Netflix, Twitter, Spotify, Reddit, CNN, PayPal, Pinterest and Fox News
- Center of the attack was Dyn
    - Dyn controls much of the internet's domain name system (DNS) infrastructure
    - DNS translates what you type into your browser (ex. amazon.com) into IP addresses that computers can understand

Forms of Defense

Firewalls

Virtual Private Networks (VPN)

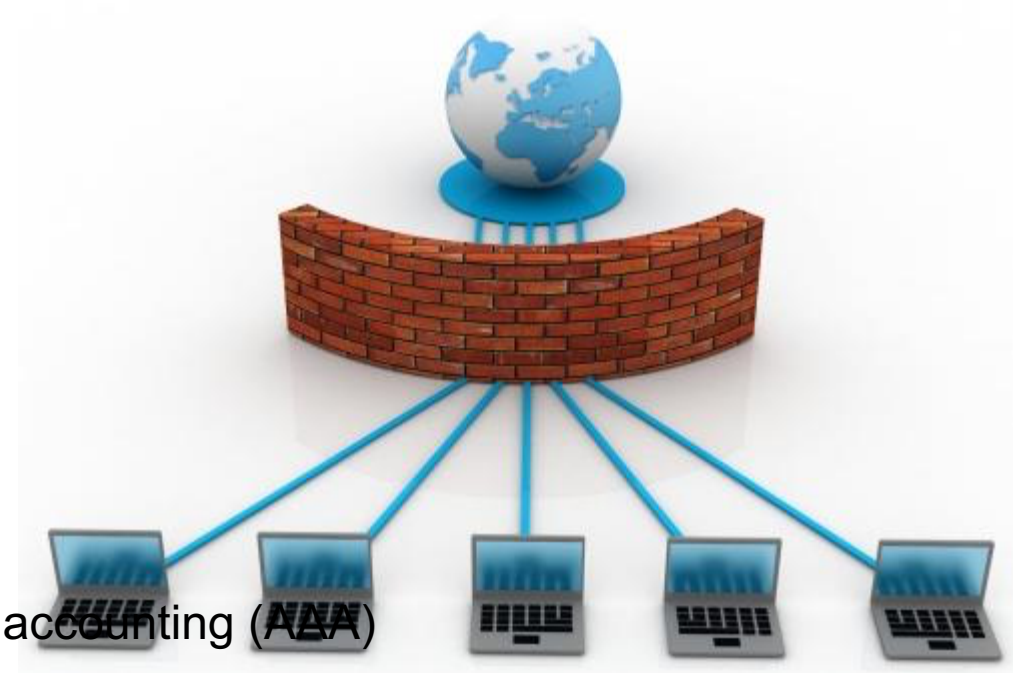
Tunneling

Network Access Control (NAC)

Authorization, authentication and accounting (AAA)

Security Scanners

Protocol Analyzers



Img src: http://www.pcquest.com/8-open-source-firewalls-to-secure-your-business/

Intrusion Detection and Prevention System (IDS)

Monitoring of network traffic, with analysis of historical and current cyber activity to detect and respond to network security breaches

Signature-Based Detection

- Each packet sent around a network contains a signature
- Signatures are used for identification of attacks
- Can only detect already known attacks / invalid signatures
- Ineffective against attack variants and polymorphism

Anomaly-Based Detection

- Find unknown attacks
- Packets converted to byte streams in vector space
- Hence can apply machine learning
- Must have concise description of normal / valid data (ex. One class SVM)
- Familiarity of unknown attacks not necessary

Hybrid Intrusion Detection

- A combination of both methods

Source:
TUB Machine Learning 2 lecture slides

Types of Anomalies

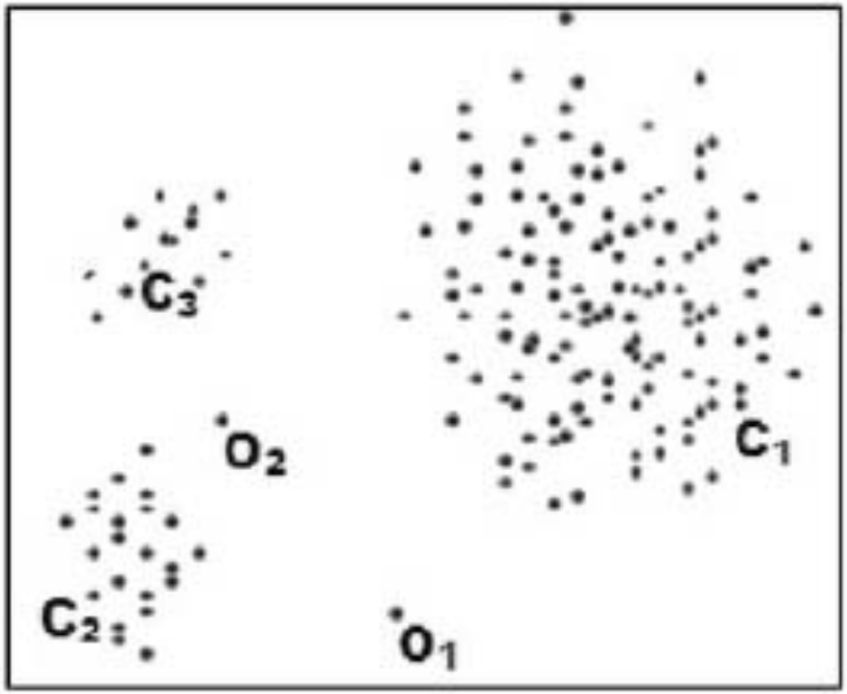


FIGURE 2.18: Point, collective and contextual anomalies

anomaly = deviation from the normal model of the data

Point Anomaly = O1
Ex. very high credit card charge

Collective Anomaly = C3
Group of instances as an anomaly

Contextual Anomaly = O2
In a given context, it is anomalous

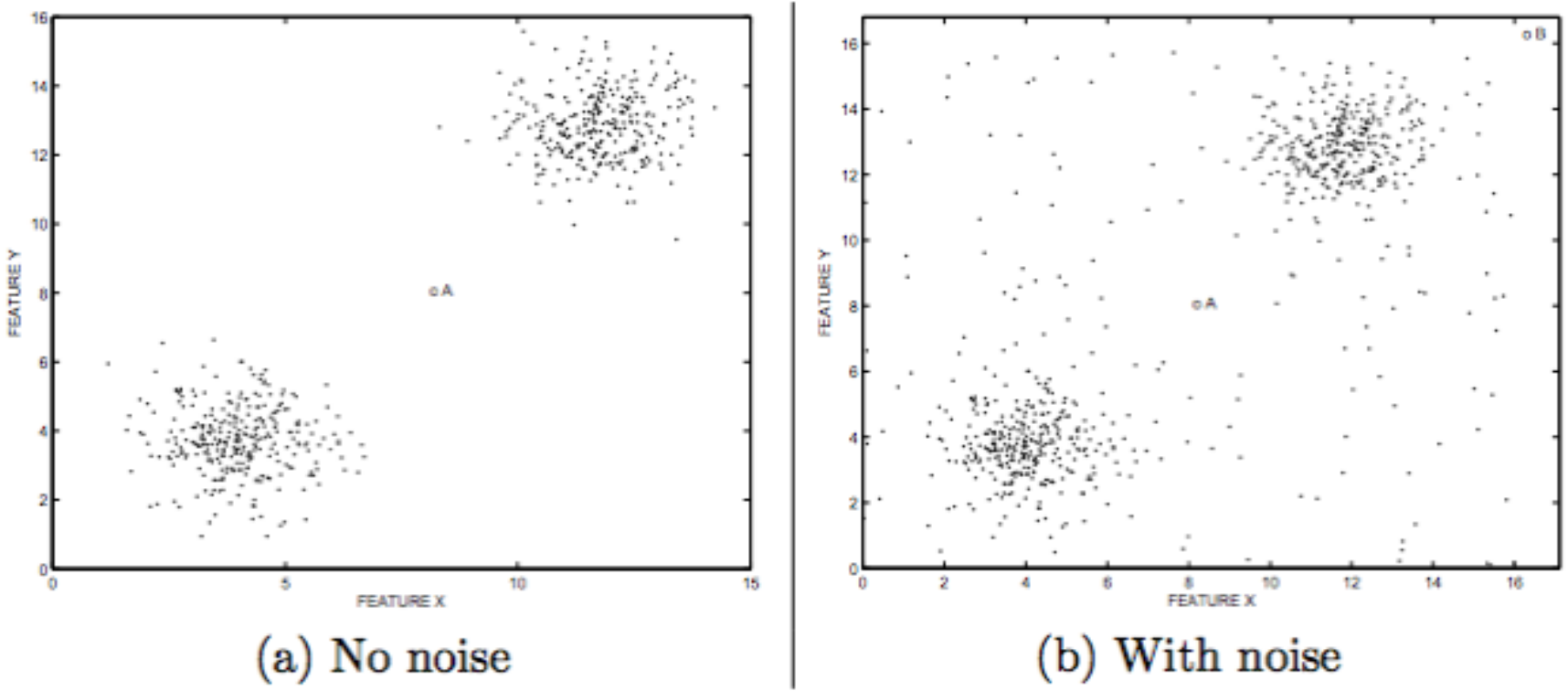
Img src: Outlier Analysis pg 23

Applications of Anomaly Detection

- fraud detection for credit cards
- insurance or health care
- intrusion detection for cyber security
- military surveillance for enemy activities
- medical diagnosis

Img src: https://www.callcentrehelper.com/nice-robotic-automation-assists-in-credit-card-fraud-prevention-86808.htm

Anomaly-based detection



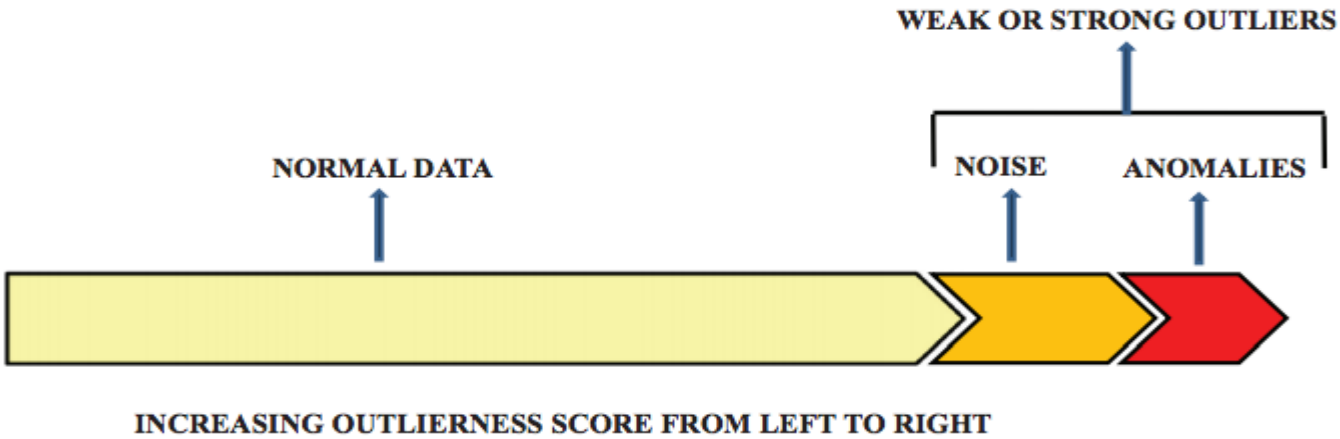
Src: http://charuaggarwal.net/outlierbook.pdf

Outlier Score

Every data point lies on a continuous spectrum from normal data to noise, and finally to anomalies

Anomalies will typically have a much higher outlier score than noise

Dep



The spectrum from normal data to outliers

Src: http://charuaggarwal.net/outlierbook.pdf

Data Preprocessing

Network data is huge, noisy, high dimensional, continuous and dynamic

Knowledge extraction in high-dimensions is difficult due to the curse of dimensionality

Aim of the preprocessing is to remove irrelevant and redundant information present in the data in order to extract knowledge from the data more accurately and reduce processing time during later phases of the data mining process.

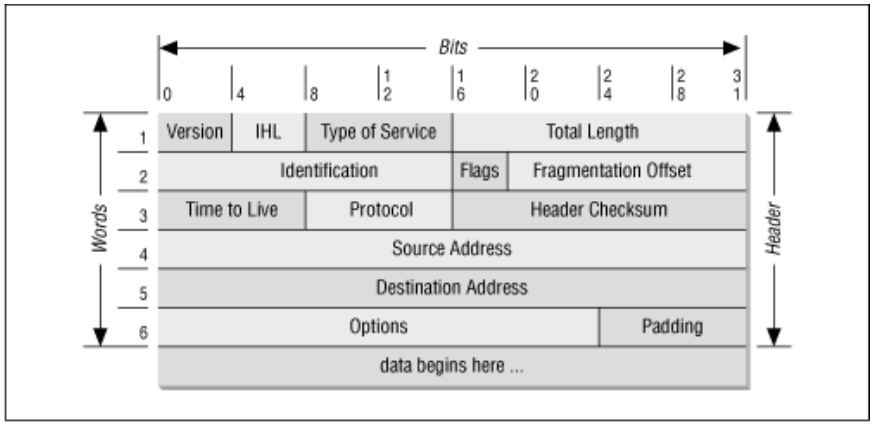
- Data cleaning
- Feature extraction and selection (Dimensionality reduction)
- Standardization and normalization

Pre-processing step takes 50 % of the effort and directly impacts on the accuracy and capability of the downstream algorithm

What aspects of the network traffic are analyzed?

Packet header anomaly detection(Traditional Method)

- packet header attributes from each data instance are compared to the trained model.and then given an anomaly score . The total anomaly score for the packet is the sum of the anomaly score for each of its attributes.



Statistical packet anomaly detection engine SPADE (Staniford et al., 2002)

Src: https://webee.technion.ac.il/labs/comnet/netcourse/CIE/Course/Section3/7.htm

- developed to detect stealthy scans,
- requires basic features extracted from protocol headers such as the source and destination IP addresses and ports.
- A traffic distribution model is built in real time by tracking joint probability measurements
- During detection, packets are compared to the probability distribution to calculate an anomaly score.

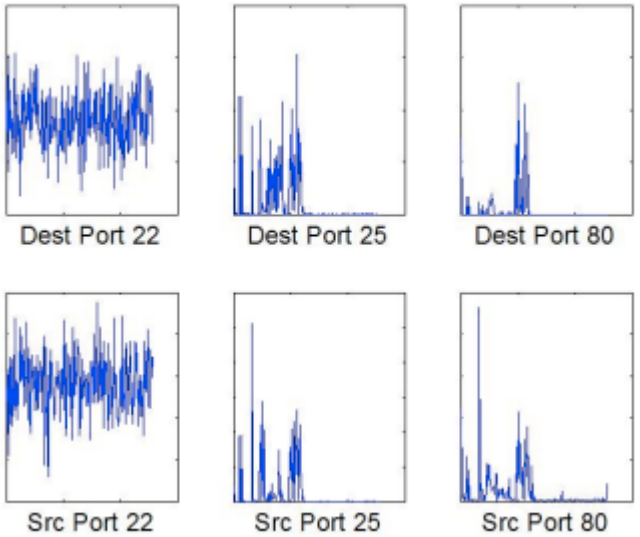
Content based anomaly detection

Payload-based anomaly detector(PAYL)

Training phase: a profile byte frequency distribution built using 1-gram and their standard deviation of the application payload flowing to a single host and port.

Testing Phase: Captures incoming payloads and tests the payload for its consistency (or distance using Mahalanobis distance) from the centroid model.

Test payload is too distant from the normal expected payload -> anomalous ->alert is generated.



https://pdfs.semanticscholar.org/29e2/97cf086d822c4d4ca6f968ecd2b798b16caa.pdf

Nimda worm and how to combat it?

- September 2001, the Nimda worm spread throughout the Internet.
- Took advantage of Escaped Character Decoding Vulnerability
- Used 16 different URLs to probe Microsoft IIS servers for known vulnerabilities, including the double hex encoding one.

  scripts/..%255c../winnt -> scripts/..%5c../winnt -> scripts/..\../winnt

  %25 is the hex encoding equivalent of the % character

  %5c is the hex encoding equivalent of the backslash.

Protocol based Anomaly Detection!!!!

Why is ML important for Anomaly Detection?

10 years ago, one of the oft-repeated mantras was "the best practice is to review all your logs every day."

Humans - good at finding patterns and noticing odd things

+

Computers - good at repetitive work and working on large scale

Machine Learning can complement human analyst and work on big scale

Adversaries will change their techniques -> increases false alarms

The model needs to be dynamically retrained over time

Algorithms

Supervised methods:

In this case, a training dataset containing normal and outlying instances, which is used to learn a model. The learned model is then applied on the test dataset in order to classify unlabeled records into normal and anomalous records

Normally used in signature based Intrusion Detection and analysed offline.

Examples: Decision Trees, k-Nearest Neighbor (kNN)

Unsupervised methods:

the data does not contain any labeling information and no separation into a training and testing phase is given.

Unsupervised learning algorithms assume that only a small fraction of the data is outlying and that the outliers exhibit a significantly different behavior than the normal records.

Examples: k-means clustering, and single linkage clustering.

Soft Computing Approaches:

There is no known algorithm that can compute an exact solution in polynomial time

Examples: Fuzzy logic, Neural Networks, Bayesian Networks

One-Class SVM for Anomaly detection

It is a semi- supervised method

The proposal to use traditional SVM method was hindered by inability to learn non-linear decision boundaries as well as the inability to account for outliers

A one-class SVM uses an implicit transformation function $\varphi(\cdot)$ defined by the kernel to project the data into a higher dimensional space. The algorithm then learns the decision boundary (a hyperplane) that separates the majority of the data from the origin

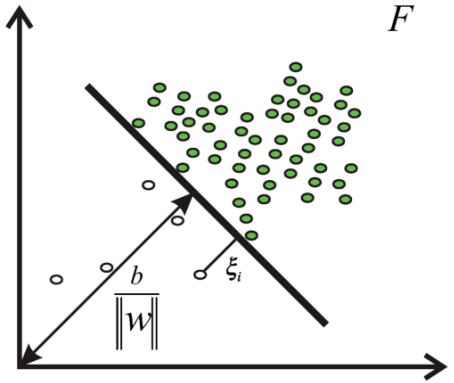
Data points lying on the other side of boundary is considered as outliers

One-Class SVM for Anomaly detection

The Gaussian kernel in particular guarantees the existence of such a decision boundary

In order to separate the data from the origin, the following quadratic program must be solved

$$\min_{w \in F, b \in \mathbb{R}, \xi \in \mathbb{R}^N} \frac{1}{2}\|w\|^2 + \frac{1}{\nu N}\sum_i^N \xi_i - b$$



subject to $w \cdot \varphi(xi)) \geq b - \xi i \;; \xi i \geq 0,\ v \in (0,1],$

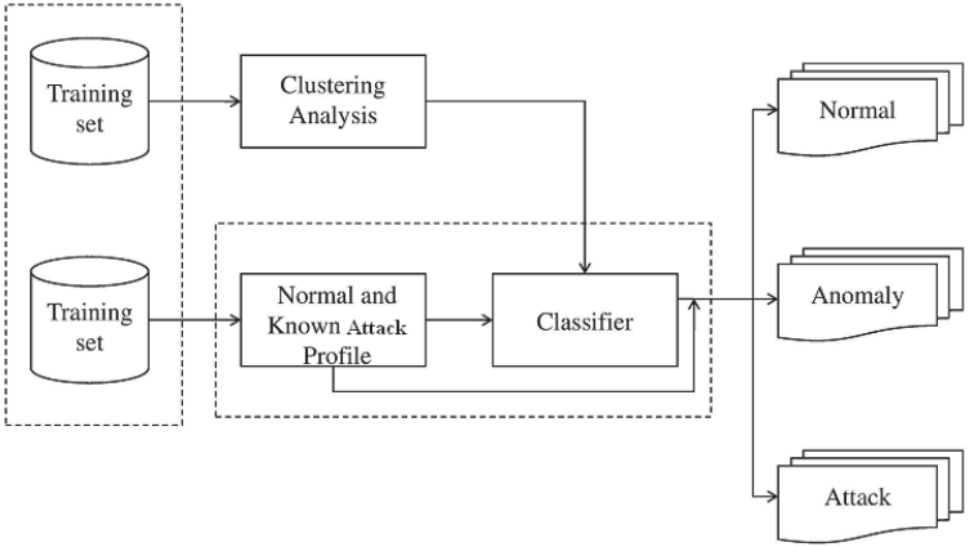
(where w is the normal vector, φ is a map function A→F , b is the bias, ξi are nonzero slack variables, v is the outlier parameter control, and k(x,y)=(φ(x),φ(y)).

Moreover, the decision function is given by **f(x)=sgn((w · φ(xi))−b).**

src:http://www.mdpi.com/1099-4300/17/9/6239/htm

Clustering-Based Anomaly Detection Methods

- Clustering methods group data into clusters-based on a similarity measure or distance computation
- Clustering is performed on the training data points so that one can select some clusters as well-known attacks and others as normal on the basis of some predefined criteria
- During the testing stage, the method uses instances as normal or anomalous.



Case Study -IDS in Wireless Sensor Networks

Sensor nodes - used to collect information from its surroundings where physical networks can't reach

- low battery power supply, limited bandwidth support, distributed operations using open wireless medium, multihop traffic forwarding, and dependency on other nodes makes it more vulnerable to security threats at all layers
- Ex: In physical layer, Radio jamming attack, Battery exhaustion attack
- Unknown attacks can be used detected using neural networks and learn time related changes using Markov Model
- Different mechanisms are used for different attacks

Anomaly based IDSs

Mechanism	Attacks
Artificial neural network	Time related changes
Set of techniques at OSI layers	Masquerade, routing attacks
Cluster based	Periodic route error attack, sink hole attack
Support vector	Black-hole attacks
Cross feature	Packet dropping attacks
Sliding window	Route depletion attack

src:http://journals.sagepub.com/doi/pdf/10.1155/2013/167575

Case Study -IDS in Wireless Sensor Networks cont'd

Characteristics	Anomaly based IDS
Detection rate	Medium
False alarm	Medium
Computation	Low
Energy consumption	Low
Attack detection	Few
Multilayer attack detections	No
Strength	Capable of detecting new attacks
Weakness	Misses well known attack
Suitable for WSN	Yes

src:http://journals.sagepub.com/doi/pdf/10.1155/2013/167575

Challenges:

Due to the energy and computational power limitations, designing appropriate IDS for WSN is a challenging task.

Small sized WSNs use Anomaly based IDS but causes false alarms as traffic pattern is almost same

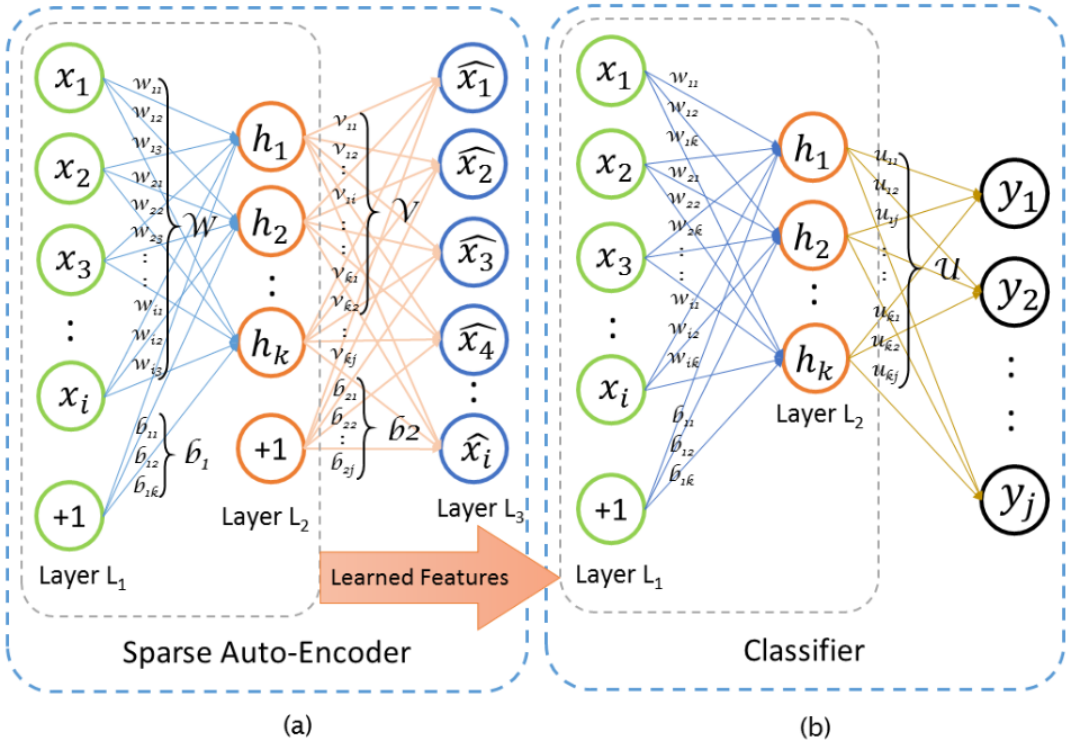
Relatively large WSN use Signature based as more susceptible to security threats but challenge is compilation and insertion of new attack signatures in the database

Deep Learning for Anomaly Detection

- Emphasis on feature selection to enhance classification results
- Voids the need for cumbersome feature engineering
- Need a lot of input data and strong computing power
- Learning parameters for a model that contains several layers of nonlinear transformations

- Self-taught learning
- 2 stages

Source: covert.io



Unsupervised Feature Learning: A good feature representation is learned from a large set of unlabeled data

Application: The learnt representation is applied to a different set of labeled data
Classify this newly mapped data with soft-max regression

Scalability

	Present	Future
# of IoT devices	13 billion in 2017	50 billion in 2020
Network Security Analytics	- examines packets in a stateless manner - performs deep packet analysis	- also compute the correlation of attributes from heterogeneous sources - improve agency's security through the continuous monitoring of stream data

Big data Analytics tools - A survey by BIGDAMA

3 objectives

- *Analytics - scalable online/offline ML and DM based techniques for large networks*
- *Big NTMA(Network Traffic Monitoring and Analysis) Frameworks - tailored to anomaly detection/network security*
- *Benchmarking - focus on stream analysis algorithms and online processing tasks*

Big Data Analysis Frameworks

Data Stream Management Systems

- Gigascope, Borealis support continuous online processing

NoSQL Systems:

- MapReduce systems such as Apache Hadoop, Spark supports unstructured data based on offline processing
- Spark Streaming, Indoop, Muppet, SCALLA - promising recent work on enabling real time analytics

Big data Analytics tools - A survey by BIGDAMA cont'd

Other Big Data Systems available in Market

NoSQL

Storm

Samza

Flink

SQL Oriented

- Hawk
- Hive
- Greenplum

Graph Oriented

- Hawk
- Hive
- Greenplum

Challenges:

❑ Memory limitations

❑ To process data fast and in a single run

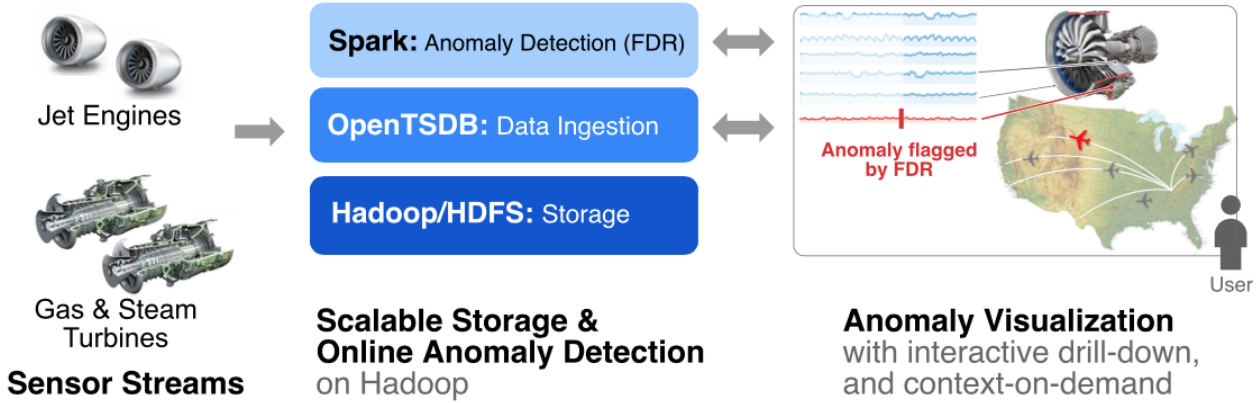
❑ Breadth of functions

❑ Reporting and Visualization

Scalability: Use Case

Power Generating Assets

- Jet engines, gas turbine
- 1000s of sensors to monitor physical state and performance
- Anomaly detection to detect faults and then set off alarms
- Falsely detected anomalies -> costly maintenance
- Similar to how network security is maintained

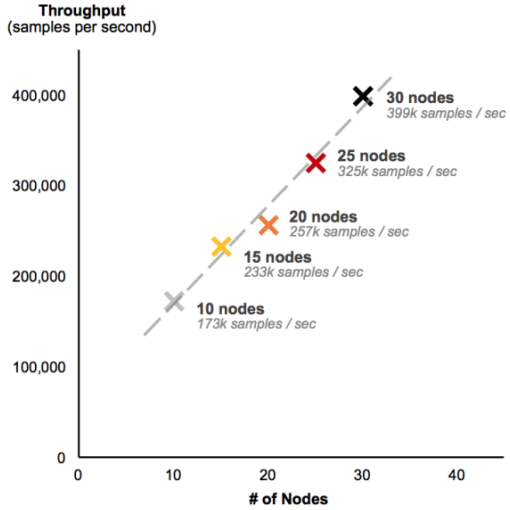


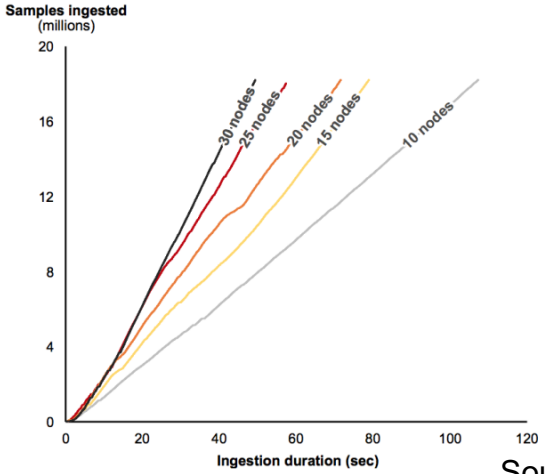
Source: arxiv.org

Scalability: Use Case cont'd

Scalable Ingestion & Storage Architecture

- Store streaming sensor data
- Open Time Series Database (OpenTSDB)
  - Supported by HBase
    - underlying distributed file storage system
    - Apache
  - Manage data in distributed manner and provide horizontal scalability
  - Easily horizontally scale out system to more storage nodes
  - While maintaining a stable, linear scaleup in streaming ingestion



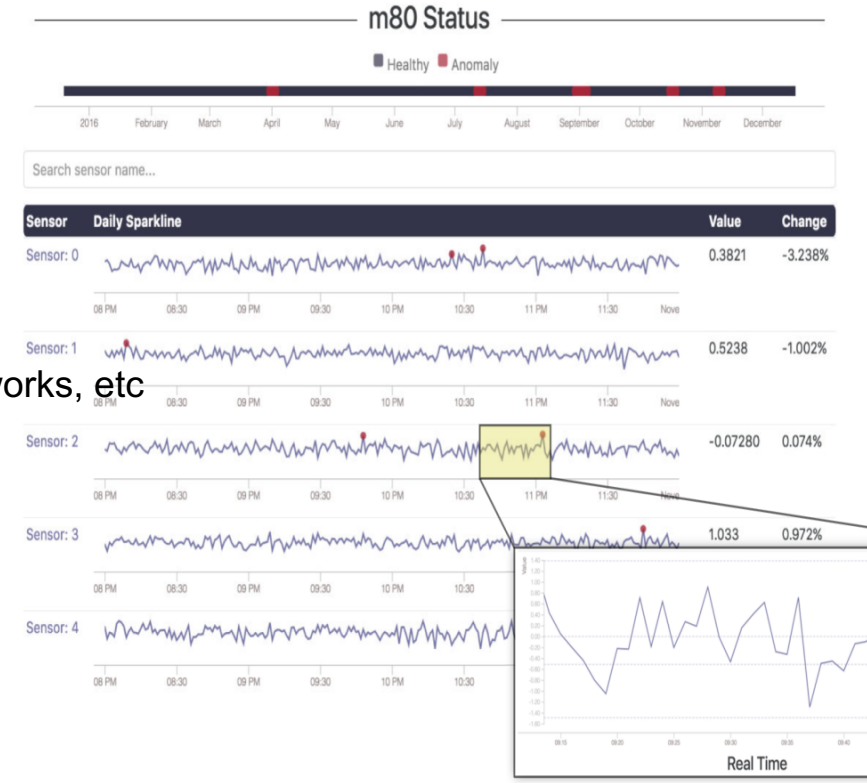


Source: arxiv.org

Scalability: Use Case cont'd

Anomaly Detection

- to control the rate of false positives
- False Discovery Rate algorithm
  - Other options: rule-based systems, neural networks, etc
- Spark to evaluate FDR algorithm
  - Rich distributed matrix computation libraries
- Offline training component
  - running in batch mode
  - MLlib to implement distributed matrix factorization



Source: arxiv.org

Strengths/weaknesses of the state-of-the-art anomaly detection algorithms

Methods	Pros	Cons
Support Vector Machine	Insensitivity to input data dimension High training rate and decision rate Better learning ability for small samples	Limited to binary classifiers which cannot give additional information about detected attack type Training time is lengthy
Decision Tree	High accuracy in detection Works well with huge data sets	It is computationally intensive to build
Bayesian Network	Can incorporate both prior knowledge and data Can encode probabilistic relationships among the variables of interest	If prior knowledge is incorrect, it is possible not to contain any good classifiers Hard to handle continuous features
Genetic Algorithm	Biologically inspired and uses evolutionary algorithms Can derive best classification rules and select optimal parameters	Can be over-fitted Constant optimization response times are not assured
Neural Networks	Do not need expert knowledge and can find novel or unknown intrusions Generalization capable from limited, incomplete, and noisy data	Possible to over-fit during training Not suitable for real-time detection because of a slow training process
Fuzzy Logic	Effective, especially for port scans and probes	Difficult to identify reduced, relevant rule subset and to update dynamic rules at runtime High resource consumption

Src: http://article.sapub.org/10.5923.j.ijnc.20170701.03.html

Challenges of using big data analytics

- The critical challenge is using this data when it is still in motion and extracting valuable information from it.
- Ingestion and storage of big data volumes
- Has to study sequential interrelation between transactions
- False positives are overwhelming as data set is huge
- Scalability is expensive
- Need for a security domain expert to be trained in Machine Learning and Big data analytics field
- characterized by high dimensionality and large sample size
    - (i) High dimensionality brings noise accumulation, spurious correlations, and incidental homogeneity
    - (ii) High dimensionality combined with large sample size creates issues such as heavy computational cost and

Summary

- Cyber Crime is escalating
- Anomaly detection methods are good in detecting network-level attacks
- Accuracy, False Negative Rate(FNR), and False Positive Rate(FPR) are three practical evaluation criteria for the IDS
- Removing redundant and irrelevant features helps feature selection and PCA is an approach to extracting features from high dimension data
- Storm, Flink, and Spark Streaming are primary open source platforms for distributed stream-processing
- Characteristics of ML techniques makes it possible to design IDS that have high detection rates and low false positive rates while the system quickly adapts itself to

Current Research / Future Challenges

Need a security solution with the capability to predict, detect and prevent long term attacks as well as solutions capable of correlating unstructured data from different data sources.

Methods for monitoring Social Engineering Threats

Integration – Manage all data on one platform

Workload Optimization – Improve upon efficient processing and storage

Sources

Aggarwal, C. Outlier Analysis. Second Edition. Chapter 1. “http://charuaggarwal.net/outlierbook.pdf”

Bhattacharyya, D. and Kalita, J. Network Anomaly Detection: A Machine Learning Perspective. CRC Press

Casas, P., D’Alconzo, A., Zseby, T., and Mellia, M.. 2016. Big-DAMA: Big Data Analytics for Network Traffic Monitoring and Analysis. Big-DAMA. “https://bigdama.ait.ac.at/wp-content/uploads/2017/03/bigdama_lancomm16.pdf”

Patcha, A. and Park, J.. 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends. Science Direct, Computer Networks 51. “https://pdfs.semanticscholar.org/7c38/70e4c6f75aefeb7801c75026cefb512304f6.pdf”

Amer, M., Goldstein, M., and Abdennadher, S. Enhancing One-class Support Vector Machines for Unsupervised Anomaly Detection. “http://www.outlier-analytics.org/odd13kdd/papers/amer,goldstein,abdennadher.pdf”

Alrajeh, N., Khan, S., and Shams, B.. Intrusion Detection Systems in Wireless Sensor Networks: A Review. International Journal of Distributed Sensor Networks. “http://journals.sagepub.com/doi/pdf/10.1155/2013/167575”

Fan, J., Han, F., and Liu, H. 2014. Challenges of Big Data Analysis. National Science Review. “https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4236847/”

Jain, P., Tailor, C., and more. 2017. Scalable Architecture for Anomaly Detection and Visualization in Power Generating Assets. “https://arxiv.org/pdf/1701.07500.pdf”

Niyaz, Q., Sun, W., and more. A Deep Learning Approach for Network Intrusion Detection System. The University of Toledo. “http://www.covert.io/research-papers/deep-learning-security/A%20Deep%20Learning%20Approach%20for%20Network%20Intrusion%20Detection%20System.pdf”

Davis, J. and Clark, A. Data preprocessing for anomaly based network intrusion detection: A review. Science Direct. “http://docshare01.docshare.tips/files/31397/313977977.pdf”

Additional References

Books

Information Security Analytics: Finding Security Insights, Patterns, and Anomalies in Big Data
Machine Learning and Data Mining for Computer Security
Network Anomaly Detection: A Machine Learning Perspective
Network Security Through Data Analysis: Building Situational Awareness

Blogs

covert.io
Data Driven Security Blog
mlsecproject
Automating OSINT
BigSnarf Blog

Security Data

Collection of Security and Network Data Resources.

See Covert.io Data Page
See Covert.io Threat Intelligence Page
See secrepo.com is more comprehensive