Unwanted traffic identification problems

Martins Ekmanis
Department of Telecommunications, Riga Technical University, Azenes iela 12, LV-1048, Riga, LATVIA,
Martins.Ekmanis@mil.lv

Abstract - Nowadays networks transport a huge amount of information. New applications like VTC, VoIP, streaming video, network games, P2P come in market every day. Different traffic need different service level, some traffic we can classify even as unwanted. Existing network devices like IDS, firewalls and routers don't have mechanisms for automatic traffic classification, so a lot of handwork must by done to set up network and adapt it day by day. In the same time a lot of job is done in field of artificial intelligence, data mining, machine learning etc. In this paper I will summarize existing solutions and create simplified model for network traffic identification and classification purposes.

Keywords: network misuse, network abuse, unwanted traffic.

I. INTRODUCTION

If we look at known threats in data networks from point of unwanted traffic, we can separate the following groups:

1. Denial of service attacks.
2. Port scans and remote vulnerability searching and virus spread.
3. P2P files exchange networks.
4. Email spam and web popup.
5. Open resources misuse (open DNS, open mail relay, open proxy, Trojan horse etc).

In each of these fields have done a lot of experiments and achieved several useful conclusions. But still does not exists universal well working model for network misuse and abuse identification and prevention.

II. TRAFFIC CLASSIFICATION

Generally classification requires training set T which consists of n samples. Each sample $N = \{x, y\}$ has class id y and vector of attributes $x = \{x_1, x_2, .., x_d\}$

Result model will be $y = f(x)$. In other words we need model returning class id to with belong traffic source by given attributes. Traffic sources usually apply to several classes in same time, so we need degree of belonging $d_n = f(x, y_n)$.

In case of IP traffic as a sample, we can use known to be bad traffic source. For example, found workstation with virus, capture its network traffic and ask model to identify all other sources with the same virus.

IV. EXISTING METHODS AND APPROACHES

A. Entropy model

Model is built on hypothesis that network anomalies cause significant changes in traffic distribution [4]. Four attributes: source address, target address, source port and target port are analyzed. Entropy is used as a metric to describe the attribution dispersion. At first the empiric histogram is made $X = \{n_i, i = 1,..., N\}$, where n_i displays attribute's i occurrence in defined time frame.

Entropy is defined as (1), where S is total observation

$$H(X) = -\sum_{i=1}^{N}\left(\frac{n_i}{S}\right)\log_2\left(\frac{n_i}{S}\right), \text{ where } S = \sum_{i=1}^{N} n_i \quad (1)$$

count in histogram. Value of entropy is in the limits $H(X) \in (0, \log_2 N)$, reaching 0, if attributes are identical and $\log_2 N$, if all observations occur equally often $n_1 = n_2 = ...n_N$. Acquired metrics are grouped with the clustering technologies, trying not only to identify the anomaly, but also to determine the anomaly type.

This model [4] does not answer, what really happens; does not identify sources and targets, but rather gives overview that something is going wrong. Author hints at this model to be used to determine DoS attack, port scanning and network functionality interruption detection.

B. Normalized Compression Distance

Original method is presented for unknown virus identification [18]. It is offered to use normalized compression distance (NDC), which supervenes from the Kalmogorov complexity. The Kalmagorov complexity $C(x)$ for definitive binary string x equals with shortest length of the program, which is able to return x. There is no definition about how the program should look like, which language it should be written in, hence authors chooses bzip2 (zlib) standard archiving algorithm, obtaining $x_{compressed}$. NDC between x and y is (2), where (3).

$$NCD(x, y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}} \quad (2)$$

$$C(x) \approx length(x_{compressed}) \quad (3)$$

If strings x and y are alike, their combination can be compressed effectively, because the information redundancy is high. The more distinctive strings are the less compression is possible.

Results show [18] that alike or derivative virus codes present small distance and are effectively grouped with the clustering methods. Also harmless program codes, as, for example, Windows command line utilities with analogical functionality present small mutual distance. Author also points that this method cannot always determine the analogy particularly, when a virus code is encrypted or masked with random value insertion.

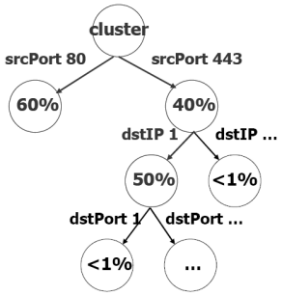
C. Behavior Models

Author examines communication patterns proposing a traffic profiling model [19]. 5-tuple flows are used as the source data. Flows are automatically clustered. These

clusters are classified by behavior. Behavior clusters are represented as structural models.

Mathematic model is based on relative uncertainty (4) $Ru(X)=\frac{H(X)}{H_{\max}(x)}$, where $H(X)=-\sum p(x_i)\cdot\log p(x_i)$ (4)

Maximal entropy is $H_{\max}(X)=\log[\min(m,N)]$. Relative uncertainty $Ru(X)\in[0,1]$ adopts value 0 if X is determined distribution, but 1 if X is random distribution. From relative uncertainty of three attributes (scrPort, dstPort, dstIP) the cube is created, where each edge is split into three parts: low, average and high. As a result each flow can be related as one of 27 behavior classes. The traffic clusters, which belong to same behavior class $BC_0, BC_1,..., BC_{26}$, represent analogical structural model, but they may represent different applications and usages in some case. Structural model is described as tree or rule set (Fig.1.).



$$BC_2 \Rightarrow \begin{cases} srcPort(1025)->dstPort(137)->dstIP(*) \\ srcPort(1081)->dstPort(137)->dstIP(*) \\ srcPort(1153)->dstPort(1434)->dstIP(*) \end{cases}$$

Fig.1. Structural traffic model representation as a tree and rule set [19]

Author shows canonical behavior profiles as a tool to describe service, server, host, virus etc. For example, server communicates with multitude clients WEB, DNS, SMTP, FTP. $srcBC\{6,7,8\}, dstBC\{18,19\}$ shows that source belongs to one of three behavior classes, but target – one of two other behavior classes.

This model does not include time information of the flow as well as transmitted information size and packet counts in flows. The model offered cannot identify concrete application, but points to the application and event groups with similar behavior.

D. Manual traffic profiling

Authors are reviewing p2p file sharing application identification and are developing traffic profile by their own [2]. Several p2p applications are chosen, which not only choose ports dynamically, but also encrypt application level data. These applications use dynamic protocols without defined ports, protocols and dedicated servers.

Well known protocols such as DNS, NTP, IRC, net games and other applications which are characterized with strongly determined ports, are filtered out off the flow in the very beginning. From remaining data set protocol identifier, target and source {ip, port} are selected, building five attribute set. Common attribute time sequences create connections (flows), assuming that if no packet arrives in a specific flow for 64 seconds, the flow expires. As a result of traffic analysis, several conclusions are made:

P2P often uses TCP and UDP connections in parallel between host computers participating in network, which rarely happens between client and server in other applications (if server runs one applications) [2].

New client when connecting to the p2p file sharing network usually knows one or several other network participant addresses, where the primary connection is made to. Within this framework the open port address for other participants is sent as well as list of locally available resources to be distributed further. N-count incoming and outgoing connections between network participants are then following. Important informative attribute from here is connection sequence in time frame.

If one host computer is marked as p2p participant, then automatically can be assumed, that other computers which this participant has continuous connection, also falls into this network – community.

In order to exclude incorrect classification, connections with length of one or several packets are filtered off, because they most probably are virus or port scanners created noise, also the connections, which are known to belong to other application are excluded from searching.

Acquired conditions are gathered in tree-type structure with IF and THEN conditions, creating algorithm to identify p2p traffic, where each data source is allocated with classification: belongs to group, doesn't belong or is not determinable.

Experiments accomplished by authors prove that this algorithm can identify p2p connections more effectively then classic packet analyzers.

E. Machine Learning Traffic Identification

Authors compare different possible attribute groups which might be used to identify the traffic [20]. It's been indicated that ports distributed by IANA is no more a useful parameter to recognize application since many applications use random ports or allows a free configuration of the ports. Also the payload analysis is not returning desired results since applications often encrypt transmittable data. Also the huge data size is required for analysis. These methods return good results only when identifying previously known applications with available signatures.

Author analyzes machine learning models which initially are adapted with training data in order to identify same application later on. Controlled and un-controlled methods are used.

$$P(A\mid B)=\frac{P(B\mid A)P(A)}{P(B)} \quad (5)$$

From controlled methods the Native Bayesa (5) is being used. From un-controlled method perspective data clustering is used. Session attributes are chosen: packet count, average packet size (separately for each direction and total), average data packet length, connection length and average packet arrival time. Three parameters are used

for method comparison (6), where TP represents correctly classified connections as appropriate, FP – incorrectly classified connections as appropriate, FN represents incorrectly classified appropriate connections as inappropriate and n represents possible number of classes to be classified.

$$precision = \frac{TP}{TP + FP} \qquad accuracy = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} (TP_i + FP_i)} \qquad (6)$$
$$recall = \frac{TP}{TP + FN}$$

The clustering produced 90% precision when Bayes method 9% less. Only the valuable traffic such as DNS, FTP, HTTP and POP3 were classified in tests.

III. ATTRIBUTES AND SIGNALS

A. DoS and DDoS

This traffic is targeted to interrupt or fully stop victim's system [1]. In these attacks the botnet (remotely controlled systems) is used. This is practically impossible to determine attack source, because unwanted traffic is going to come from hundreds or even tens of thousands different sources. As a victim for these attacks can be any on-line business, web resource, bank or other company or governmental institution.

This traffic has huge size (volume) which might be identified with statistical methods [6, 8]. Despite the ease of identification it is very difficult to protect systems from such attacks, due the fact that attack sources are diffused and very many in counts.

Protections should be made as close as possible to attack source, which, in fact, is not determinable. In practice such security actions are implemented by ISP before unwanted traffic reaches client's (victim's) system [7].

B. Port scanning

Depending on attack objective the traffic may be and also may not be huge in volume. If the objective is to identify weak points of system, long performance process can be used, which is not creating traffic volume to be identified. During few days or even weeks this is possible to poll all interested system ports and get the whole picture about available system services and their weakest points. However, in case of virus distribution the scanning process is automated and fast as possible. This can be identified with volume based signals and statistics [3,4]. In case of virus attack, usually there is high number of connections open (one service – many recipients). Network viruses can be identified with help of signatures, which actually is most commonly used practice. In order to identify slow scanning it is necessary to obtain data volume in long term, which is not currently possible to exercise.

C. Pear to Pear files exchange

Fail exchange networks relieve with high volume traffic and decentralized management. To identify such it is difficult to use volume based methods, because traffic is masked and statistically is not much different from other, preferable usages. Signature based methods also is not returning desired result, because connection ports very often are random, transferred data is encrypted or compressed [12]. Also exchange protocols are regularly updated and changed [9]. Good results are presented by behavior (performance) based identification solutions [2]. P2P traffic is characterized by traffic flow directions, host group and their behavior. P2P participants most often is closed group, who communicate between themselves.

D. Application level data

Higher level unwanted traffic has to be analyzed in high level OSI context [14]. To differentiate between wanted and unwanted e-mails huge attribute set are used [15]. In order to obtain attributes, many different data bases, signatures, white, black and gray lists, key-words, attached file analysis and many other different methods are used. All this complicated process introduces additional traffic, load on system resources and additional labor configuring the system, meanwhile providing only partly decoding of messages [16]. Likewise the web page unwanted content identification is done with help of signatures and unwanted URL data bases. Unfortunately the content of such databases is imperfect and cannot fully protect against rapidly changing threats in the Internet environment.

E. Open resources

Great harm to system performance is caused by configuration inaccuracy and backdoors caused by viruses. Such systems generate unwanted traffic by executing tasks for third – unauthorized – parties [17]. Today's solutions to identify such threads include network security scanners, which tries to provoke each host to perform unwanted actions by using pre-made check procedures, such as e-mail relay, open DNS, open share etc. Unfortunately these methods cannot identify presence of unknown services, brought by viruses and letting remotely control hosts

Summarizing methods, used to identify different unwanted traffic groups, common attributes can be noticed:

Signatures in different ways of execution are used to identify viruses, spam and web threats. Also unwanted traffic on routers and firewalls are described with manufacturer specific signatures, specifying ports, protocols, addresses, flags etc. This method is easy to implement, but very time-consuming in maintenance, because each signature system has different syntaxes and all they are generally hand wrote.

Volume based signals also requires people effort when describing threshold values and choosing appropriate counters.

Behavior based identification looks promising, which, rarely used in practice, are very well appreciate in different publications by many authors. Within this method, the functional role of the source (server, client, P2P system participant) is being reviewed. Also this system reviews time diagrams in making connections, host groups and their interaction within group and to outside etc. Each of behavior analyzes tool use their hard-coded identification

system which limits summary of found algorithms, grouping and exchange [13].

IV. NETWORK MODEL

I will use Flow information as base element. Major part of routers already has implemented NetFlow functionality. It defines the flow as unidirectional sequence of packets [5]. There are already methods and products using NetFlow data for early worm and other abnormal network activity detection.

As more information will be available to classification model as best results are possible. Topological information as proven before [2] also is valuable in traffic source classification. Thus network will be graph (Fig.2.) $G=(S,F)$, where S is non empty set of traffic sources (hosts) and F is set of flows between hosts. Each flow $F=\{x_1,x_2,...,x_n\}$ is a vector consisting of several attributes. NetFlow v5 define following attributes [5]: the source and destination IP address, next hop address, input and output interface number, number of packet in the flow, total bytes in the flow, the source and destination port, the protocol, ToS, and TCP flags (cumulative OR of TCP flags).

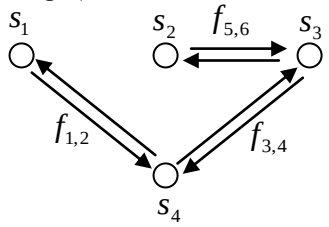


Fig.2. Network traffic flow graph

More information is available if look at flows as time diagram (Fig.3.). If connections is following one after other it is possible, then first connection give information for second. If one flow causes another sub flow, it is possible then host ask other service like database for additional information.

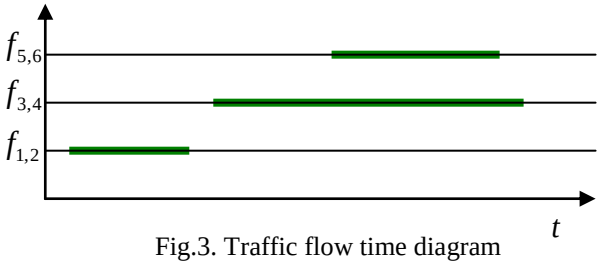


Fig.3. Traffic flow time diagram

In real network it will not be possible to get all information. We can read flow information only on finite number of cross-cut where flow detectors are set. We must take it in account then some data will be hidden to system, and model must assume probability than other hidden flows exist.

CONCLUSIONS

In this work I have summarized existing sources of unwanted traffic. Majority existing methods of traffic classification, identification and profiling are analyzed and conclusions about most informative attributes are made. New network model for unwanted traffic identification is purposed.

In order to achieve effective network traffic control and defense, the mechanism model should include as much non overlapping information as possible. Purposed model includes flow information, topological data and flow time diagrams with cover main attributes used in existing models. When creating more sophisticated models it becomes more and more difficult to compare knowledge base objects; to create metrics and distances between them. If it is not possible to compare objects, methods like case based reasoning, clustering and others will be difficult to use.

My ongoing work is centered on deeper analysis of inductive and deductive reasoning methods. Also new knowledge base architecture must be created or existing one adapted.

ACKNOWLEDGMENT

This work has been partly supported by the European Social Fund within the National Programme “Support for the carrying out doctoral study programm`s and post-doctoral researches” project “Support for the development of doctoral studies at Riga Technical University”.

REFERENCES

[1] Cisco systems, “White paper: Leading DDoS protection for service providers and their customers”, 2005.
[2] T. Karagiannis, A. Broido, M. Faloutsos, "Transport Layer Identification of P2P Traffic", 2004,
[3] D. Marchette, “A Statistical Methods for Profiling Network Traffic”, Proceedings of the Workshop on Intrusion Detection and Network Monitoring, ASV, 1999.
[4] A. Lakhina, M. Crovella, C.Diot, “Mining Anomalies Using Traffic Feature Distributions”, 2005.
[5] Cisco Systems, “Introduction to Cisco IOS NetFlow”, 2007.
[6] P. Barford, D. Plonka, “Characteristics of Network Traffic Flow Anomalies”, University of Wisconsin, Madison, 2001.
[7] A. Akella, A. Bharambe, M. Reiter, S. Seshan, “Detecting DDoS Attacks on ISP Networks”, Carnegie Mellon University, 2003.
[8] P. Barford, J. Kline, D. Plonka, A. Ron, “A Signal Analysis of Network Traffic Anomalies”, Proceedings of ACM SIGCOMM Internet Measurement Workshop, 2002.
[9] K.Tutschku, “A Measurement-based Traffic Profle of the eDonkey Filesharing Service”, University of Wurzburg, Germany.
[10] Kuai Xu, Zhi-Li Zhang, Supratik Bhattacharyya, “Reducing Unwanted Traffic in a Backbone Network”, 2005.
[11] T. Abbes, A. Bouhoula, M. Rusinowitch, "Protocol Analysis in Intrusion Detection Using Decision Tree", France, 2004.
[12] S. Sen, O. Spatscheck, D. Wang, “Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures”, AT&T Labs-Research, 2004.
[13] A. Tauriainen, "A Self-learning System for P2P Traffic Classification", Helsinki University of Technology, 2005.
[14] M. Ekmanis, “Analysis of Spam Filtering Optimization Methods”, Scientific Proceeding of RTU, Series 7, Telecommunications and electronics, vol.5., pp 44-47, 2005.
[15] „The Apache SpamAssassin Project”, http://spamassassin.apache.org
[16] “Snort”, http://www.snort.org
[17] “Spam and Open-Relay Blocking System”, http://www.sorbs.net
[18] S. Wehner, “Analyzing Worms and Network Traffic using Compression”, Amsterdam, Netherlands, 2005.
[19] Supratik Bhattacharyya, Kuai Xu, Zhi-Li Zhang, “Profiling Internet Backbone Traffic: Behavior Models, Applications and Implementation”, University of Minnesota, 2005.
[20] J. Erman, A. Mahanti, M. Arlitt, “Internet Traffic Identification using Machine Learning”, University of Calgary, Canada, 2006.