ФИЗИКО-МАТЕМАТИЧЕСКИЕ И ТЕХНИЧЕСКИЕ НАУКИ

УДК 004.023
МРНТИ 49.38.49

METHODS AND TOOLS FOR NETWORK TRAFFIC CLASSIFICATION

ABSATTAR D.

Kazakh-British Technical University

Abstract: *Since the first releases of intercommunication models between end-devices, like ring topology, the whole picture of now days network infrastructure was changed unrecognizably. Modern systems of networking consist of many complex intermediate modules like switches, routers, firewalls, hubs etc. and the main goal of these inventions was to provide more reliable and scalable ground for communication (Quality of Service). Meanwhile, rapid growth of traffic on the Internet forced network engineers and software reliability engineers to pay more attention on the optimization of data flow from both sides, developing network-oriented software and application-oriented network. To apply effective solutions on these tasks, engineers need to research specifics of the current network state. The more whole system evolves, more data about network traffic we gain, and now it helps us to make optimization and tuning of intermediate devices, rather than just scaling it up with more bare hardware. Which protocols are used the most? What types of applications loads the network bandwidth the most? and etc. Classification of packets can help resolve the answers, and there are different approaches to achieve this. In this study, I tried to explore already known tools and methods that can be applied to solve such tasks.*

Key words: *application identification, traffic characterization, advanced network management, convolutional Neural Networks, network traffic classification*

МЕТОДЫ И ИНСТРУМЕНТЫ ДЛЯ КЛАССИФИКАЦИИ СЕТЕВОГО ТРАФИКА

Аннотация: *Со времени первых выпусков моделей взаимодействия между конечными устройствами, таких как кольцевая топология, вся картина современной сетевой инфраструктуры изменилась до неузнаваемости. Современные сетевые системы состоят из множества сложных промежуточных модулей, таких как коммутаторы, маршрутизаторы, брандмауэры, концентраторы и т.д. И главная цель этих изобретений заключалась в том, чтобы обеспечить более надежную и масштабируемую основу для связи (качество обслуживания). Тем временем, стремительный рост трафика в интернете заставил сетевых инженеров и инженеров по надежности программного обеспечения уделять больше внимания оптимизации потока данных с обеих сторон, разрабатывая как сетевое программное обеспечение, так и сеть, ориентированную для прикладных программ. Для применения эффективных решений этих задач инженерам необходимо исследовать специфику текущего состояния сети. Чем больше развивается вся система, тем больше данных о сетевом трафике мы получаем, и теперь это помогает нам производить оптимизацию и настройку промежуточных устройств, а не просто масштабировать их с помощью большего количества голого оборудования. Какие протоколы используются чаще всего? Какие типы приложений больше всего загружают пропускную способность сети и так далее. Классификация пакетов может помочь решить ответы, и для этого существуют различные подходы. В этом исследовании я попытался изучить уже известные инструменты и методы, которые могут быть применены для решения подобных задач.*

Ключевые слова: *идентификация приложений, характеристика трафика, расширенное сетевое управление, сверхточные нейронные сети, классификация траффика в сети*

ЖЕЛІЛІК ТРАФИКТІ ЖІКТЕУ ӘДІСТЕРІ МЕН ҚҰРАЛДАРЫ

Аңдатпа: *Сақиналы топология сияқты, ең алғашқы ақпараттық құралдар байланысының модельдерінің пайда болғаннан бері, қазіргі заманғы желілік инфрақұрылымның көрінісі адам танымастай өзгерді. Заманауи желілік жүйелер, коммутаторлар, маршрутизаторлар, брандмауэрлер, концентраторлар және т.б. сияқты көптеген күрделі аралық модульдерден тұрады және осы өнертабыстардың басты мақсаты: байланыс үшін (қызмет көрсету сапасы) неғұрлым сенімді және масштабталатын негізді қамтамасыз ету болып табылады. Сонымен қатар интернеттегі трафиктің қарқынды өсуі желілік инженерлер мен бағдарламаны қамтамасыз етудің сенімділігі бойынша инженерлерді(ағыл. Software Reliability Engineers), деректер ағынын екі жағынан да оңтайландыруға көбірек көңіл бөлуге мәжбүр етті. Бұл мәселелерді тиімді шешу үшін инженерлер желінің ағымдағы жай-күйінің ерекшелігін зерттеуі қажет. Барлық жүйе тұтас дамып келе жатқан сайын, желі трафигі туралы мәліметтер соғұрлым көп болады, ал бұл бізге көп мардымсыз жабдықтың көмегімен оларды масштабтау ғана емес, аралық құрылғыларды оңтайландыруға және теңшеуге көмектеседі. Қандай хаттамалар жиі пайдаланылады? Желі өткізу қабілетін ең көп жүктеуге қандай қосымшалар түрлері бар? және т.б. Пакеттердің жіктелуі көптеген мәселелерді шешуге көмектесе алады және осыған орай әртүрлі тәсілдер бар. Бұл зерттеуде осындай міндеттерді шешу үшін қолданылуы тиіс белгілі құралдар мен әдістерді зерттеуге тырыстық.*

Түйінді сөздер: *қосымшаларды идентификациялау, трафик сипаттамасы, кеңейтілген желілік басқару, нейрондық желілер, желі трафигінің классификациясы*

Introduction

Since the first releases of intercommunication models between end-devices, like ring topology, the whole picture of nowadays network infrastructure was changed unrecognizable. Modern systems of networking, consists of many complex intermediate modules like switches, routers, firewalls, hubs etc. and the main goal of these inventions was to provide more reliable and scalable ground for communication(Quality of Service). The more whole system evolves, more data about network traffic we gain, and now it helps us to make optimization and tuning of intermediate devices, rather than just scaling it up with more bare hardware.

Many so-called "peer-to-peer (P2P)" shared applications, social networks, video streaming services, instant messaging services, online games, etc. have appeared on the Internet. This has led to a significant increase in the number of users and changes in their behavior. As a result, the volume of Internet traffic has significantly increased and its nature has changed. However, many different types of protocols are used on the Internet. In addition, network applications have different functional requirements, and most of these applications use TCP or UDP port numbers that are assigned by the IANA (Internet Assigned Numbers Authority) [1].

IANA has assigned specific port numbers for specific network applications, protocols, and services that change between 0 and 1023, and IANA has registered port numbers that change between 1024 and 49151. Even so, most applications do not have IANA-assigned port numbers, but use default port numbers, and these numbers often match the IANA port numbers. Therefore, it is often not possible to uniquely identify network applications with known or registered ports. So, in such conditions, it is very difficult to provide the required level of network performance and security, as well as QoS (quality of Service) for applications, services, etc.

However, research has shown that network traffic is a complex dynamic process and is a superposition of many threads with multiple interconnected characteristics that are generated by different protocols. First, it is traffic and related to the management of the network itself (for example, client initialization traffic, server traffic, etc.) that are generated periodically. Second, it is the traffic of network services, applications (for example, P2P, DNS, POP3, FTP, SMTP, ARP, NetBIOS session, HTTP, WINS requests, Telnet, etc.) and protocols that make up the bulk of network traffic [2].

Based on the above, effective methods of network monitoring, analysis, and evaluation are required to ensure the normal and safe operation of networks. To do this, first of all, it is necessary to accurately identify network traffic, which is a very difficult task and requires the development of adequate methods for identifying network traffic.

Knowledge about types of protocols and even better, about applications that network clients use, may help to construct better data flows and utilize resources properly. Traffic classification attracted a lot of interests from both industrial and academic activities related to advanced network management.

The purpose of this article is to analyze the methods of network traffic identification available in the literature in order to evaluate their capabilities for network traffic identification.

Common problems of traffic classification

The emergence of new applications, protocols and interactions between various endpoints in the Internet has totally increased the complexity of task of classifying traffic. Here are some of the critical challenges that we can face.

Encryption. Nowdays, most of the applications uses encryption of data, because big corporations like Google inc. forces software developers and organizations to use HTTPS instead of HTTP protocol. As a result, we got a lot of traffic with pseudo random payload and therefore classification of traffic become even harder in modern networks.

ISP. Most of the Internet Service Providers blocks peer-to-peer connections due to their overload of network bandwidth and copyright issues from authors. And now, these applications uses different techniques to bypass blocking by Internet traffic control from ISP. This is the most challenging task in network traffic classification.

So, despite the fairly active development of the field of network traffic classification, many works note a number of objective factors that hinder this development [3]. One of these factors is the lack of an open data set for testing, which is usually a saved and marked network routes. As a result, it is difficult to test the quality of the algorithm being developed, as well as to compare it with other algorithms. In particular, this leads to the need to solve two problems in the process of developing each new one algorithm:

- Getting your own network route on the internal network, from research partners, or from public sources. A complicating factor is the problem of privacy and emerging information security risks. To level out these factors, the resulting routes are usually pre-anonymized [4]. This, in turn, leads to the inapplicability of content analysis approaches, since the main method of anonymization, among other things, is to delete the content of the application-level package.
- Network trace marking by protocols and applications, for subsequent quality control of the developed algorithm, which can be performed in several ways, depending on whether the process of removing the network route is controlled or the route is obtained from an external source.

As a result, most research works use different trace snapshots, obtained at different points in different networks, under different scenarios, in particular - different time intervals.

On the other hand, the requirement of privacy leads to a more active development of the statistical direction of classification. This is due to the fact that this group does not require access to packet data, but only general characteristics such as size and timestamp are sufficient. Thus, a large number of values are suitable as input data and there are number of open network routes that have passed the anonymization procedure.

Fields of application

In addition to the question of the approach, another important factor is the applied problem and the solution by the specific system where the classification component is implemented. Depending on this, for example, the acceptable level of accuracy of classification results may differ markedly.

In addition, the set of groups into which many classified objects are divided may differ significantly. The roughest classification is usually used in traffic management systems, whose main task is to efficiently use available bandwidth. For example, an Internet provider can identify three main traffic groups:

• Sensitive – a type of traffic that is sensitive to delays and requires prompt delivery. This includes VoIP, video streaming, online game traffic, etc.

• Unwanted – spam and malicious traffic types.

• The rest – is the traffic that is allocated the remaining bandwidth servicing sensitive data streams.

Security systems and policy enforcement systems usually involve a much more precise classification – you need to identify the specific application that generates the corresponding traffic. In some cases, it is necessary to perform a complete analysis of traffic with the allocation of transmitted commands and high-level objects, such as web pages and other types of files. This may be required, for example, to detect potentially dangerous content. For roughness assessment for a specific approach, the term "granularity" is used.

The processing speed, i.e. the throughput of the algorithm, is a factor that affects the estimation of the approach to apply on specific task. This characteristic consists of two things: the amount of data that the algorithm must process to get the result and the complexity of the algorithm relative to the input length.

This characteristic is most relevant for DPI approaches that use the maximum amount of data to process – the entire payload contents of individual packages. This issue is studied in detail in a large number of papers, mainly in the context of choosing the type of automaton to search for signatures of various protocols: deterministic, nondeterministic, or some hybrid version [5-10].

Adopted methods of traffic classification

Existing methods for identifying network traffic are roughly divided into five categories: port-based identification methods; deep packet inspection [11], DPI identification methods, i.e. packet content analysis; identification methods based on network flow characteristics analysis; identification methods based on host behavior analysis; and machine learning algorithms-based identification methods.

Traditionally, simple methods based on analysis of network traffic characteristics were used to identify network traffic. These characteristics include packet characteristics such as port numbers, sender and recipient IP addresses, application and protocol types, packet contents, traffic statistics, and so on. Some of these methods are discussed in [12, 13].

Port-based approach is the most common and the oldest method used for traffic classification, which consists of the analysis of the communication ports defined in packet headers of the TCP/UDP network model. Since the usage of this ports are so wide, and almost became a standart in computer communications, IANA defined the list of well-known ports for different protocols, such as http, https, ssh, telnet, etc. Also, need to mention that this information is usually not affected by encryption and can be easily extracted from the packet data. Which makes the classification of network traffic based on port very fast and easy, and that's the reason why ACL rules and firewalls uses them to filter the incoming and outgoing data flow.

Nevertheless, not all protocols can be classified with the port-based approach. Indeed, protocols such as Peer-to-Peer (P2P) or passive FTP can use ephemeral or random ports. In addition, such applications can use ports associated with other protocols for masquerading purposes. Another example is the internet telephony where SIP is used to negotiate the terms for the call, e.g., port numbers, codecs among many others, which is then realized with RTP on random port numbers. Finally, this approach also fails on tunnels or Network Address Port Translation (NAPT). As described in [14] and [15] only 30%-70% of the traffic generated by certain protocols can be detected by evaluating the port numbers.

However, identification of network traffic based on port numbers is ineffective today [16]. This is mainly due to the emergence of more network applications and services that use non-standard TCP ports, as well as applications that tunnel HTTP and the widespread use of P2P applications on the Internet. As a result, some applications cannot be identified at all. The solution to this situation may be to analyze the contents of packages and create a signature for each application, but there are at least two problems: first legal and ethical, which is related to the user's privacy, and second is the inability to identify encrypted network traffic.

The idea of using statistical characteristics of network graphs to identify them or to describe their properties is not new. In [17, 18] for the first time, the issues of determining the characteristics of Internet traffic were considered and the relationship between the characteristics of flows and the application protocols that generate them was mainly determined. These studies show that analytical models of random variables can be used to describe the properties of several protocols.

Despite the fact that network traffic identification is a fairly specific area of research, the goals of existing work in this area are not identical. The purpose of some works is only to identify P2P traffic; the purpose of others is a detailed classification of network traffic, that is, the exact identification of the application that generates a specific traffic. In addition, with the rise of the new types of applications, the nature of existing network characteristics may change and other identification characteristics may be used to identify network traffic. For example, the emergence of some new applications, such as PPStream, BitTorrent, PPLive, etc., has led to the widespread use of the UDP Protocol.

In [6], methods for identifying network traffic with a detailed analysis of the contents of packets were proposed. The main disadvantage of these methods is that they require very large computing resources. At the same time, the accuracy of network traffic identification depends mostly on models based on the identified patterns and reflecting the main features of network traffic. However, despite the fairly high identification accuracy obtained in [h], traffics classified manually were used as input data for training the naive Bayesian algorithm.

In [8], proposed a method for classifying network graphs based on statistical analysis of host activity. However, packet contents are not analyzed, and host behavior patterns are mapped to one or more applications to classify network traffic.

A study of the disadvantages of network graph identification methods based on the analysis of port numbers and packet contents has shown that machine learning (ML) methods are more suitable for identifying network traffic [11].

Identification of network traffic based on machine learning algorithms

When identifying network traffic, one of the important areas of research is classification. The purpose of classification is to build classification models for predicting an unknown sample based on the study of a set of training data.

In the last decade, a significant part of the work on network traffic identification has been based on their classification using ML methods. These works can be classified as works that use ML methods with a teacher (supervised), without a teacher (unsupervised), and so-called semi-learning (hybrid) methods.

In network traffic classification based on ML, supervised learning methods is when training data is analyzed and an assumed function is output that can predict output classes from any test stream of data. However, it is very important to choose sufficiently well-founded training data. The methods of ML with the teacher include the following: Decision Trees - DT; Naive Bayesian Classification - NBC; Ordinary Least Squares Regression - OLSR; Logical Regression - LR; The method of support vectors - Support Vector Machine SVM, etc.

Using methods for classifying network traffic with ML algorithms without a teacher (i.e. clustering algorithms), clusters are found in unmarked traffic data and the data is detected in certain clusters. The unsupervised methods of ML without a teacher include the following: clustering algorithms; Principal Component Analysis - PCA; Independent Component Analysis; Singular Value Decomposition (SVD); Random Forest (RF); Self-Organizing Map - SOM, etc.

In [19], the researchers evaluated algorithms with a teacher, including a naive Bayesian algorithm with discretization, a naive Bayesian algorithm with an estimation of the density kernel, a C4.5 decision tree, a Bayesian trees and networks.

In [20], the authors proposed an approach to traffic classification based on real-time packet flow analysis. In [21], Bayesian neural networks are used to accurately classify traffic. In [22], the authors use unidirectional statistical functions to classify traffic. In [23], the authors used the probability density function to compactly express three statistical characteristics of traffic. In [24], the authors proposed using a single-class SVM (one class support vector machines) for traffic classification, and a simple optimization algorithm was proposed for each set of SVM operating parameters.

All these works used mod parametric algorithms, which require intensive training for classifier parameters and need to be re-trained when new applications are discovered.

Also, there are several papers based on nonparametric ML algorithms. In [25], the authors used the methods of nearest neighbors and linear discriminant analysis to classify traffic. Five statistical characteristics were used for classification. In [26] a so-called BLINC method is proposed for traffic classification, which uses the behavior of hosts. Although nonparametric methods have some advantages over parametric methods, for some reason, they are not widely used for traffic classification.

In [27], the authors proposed using the EM algorithm (Expectation Maximization Algorithm) to group traffic flows in a small number of clusters, and each cluster is marked manually. In [28], the AutoClass algorithm was used for clustering traffic flow, and a metric of intra-class homogeneity was proposed for evaluating clusters. In [29], the K-means algorithm was used for clustering traffic and clusters for applications were marked using the analysis of useful information. In [30], the authors evaluated K-means, DBSCAN, and AutoClass algorithms for clustering traffic based on two sets of empirically collected data.

In General, these clusterization methods can be used to identify traffic from previously unknown applications. In study [31], authors proposed integrating clusterization, based on statistical flow characteristics with a method for comparing the signature of useful information, which eliminates the need to use training data sets. In [32], the authors proposed combination of clusterization, based on statistical characteristics of the flow and clusterization, based on statistical characteristics of useful information for detecting unknown traffic.

However, clusterization methods have the problem of mapping a large number of clusters to real applications. This problem is very difficult to solve if there is no information about real applications. To solve this issue, a new nonparametric approach is proposed in [33]. This approach consists of including a correlative information of flows in the classification process.

Semi-trained or hybrid methods for classifying network traffic use both marked and unmarked flow statistics [34]. Because of this approach, these methods provide more accurate and faster traffic classification, as well as allow you to identify unknown applications and applications with dynamic behavior. In study [35], authors proposed using a set of training data in the ML algorithm without a teacher. However, if the training data is too small, the main part of the display is made up of "unknown" clusters.

In [36] for identifying a TCP and UDP Protocol traffic, author proposed a classification method based on the use of the support vector method (Support Vector Machine - SVM). In this approach, a genetic algorithm is used to select a subset of the best characteristics, and the Particle

Swarm Optimization (PSO) method is used to calculate the weights of each characteristic. At the same time, the traditional SVM algorithm is used to classify the different traffic flows with optimiziation using the PSO algorithm, which can effectively improve the performance of the SVM algorithm. The proposed approach allows you to classify internet traffic, based on statistical characteristics of traffic flows without using port or host information, and there is no need to check the application signature.

In [37], to identify network traffic, authors proposed a hybrid model that uses the Apriori algorithm for atomic generation of associative rules and a self-organizing Koohonen (SOM). This proposed approach allows you to identify network traffic without using content and port numbers, as well as generate associative rules for identifying unknown applications. At the same time, the Apriori algorithm allows you to choose the most typical rules for each type of traffic, while the SOM-based algorithm allows you to group the characteristics of similar protocols and applications.

The author in [38] proposed an approach to identifying P2P traffic based on the random forest algorithm. The random forest algorithm is a combination of decision trees. Building a random forest allows you to increase the accuracy and efficiency of p2p traffic identification.

Conclusion

Classification of Internet traffic has been an area of intensive research since the creation of the Internet itself. Over the years, several methodologies have been proposed to solve existing technological problems. Thus, we can conclude that the evolution of approaches of traffic classification has directly affected the evolution of the international network itself. Surveys then become a valuable tool for understanding and analyzing this evolution. Several reviews have been published to provide an overview of this ever-evolving field of research. However, such surveys focused only on the analysis of statistical work on traffic classification and were limited to reporting and comparing published results.

For this purpose, the best solution is to combine existing classification mechanisms using the supervised and unsupervised ML methods, as well as using an ensemble of classifiers. This will significantly improve the accuracy and completeness of network traffic identification.

REFERENCES

1. https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml
2. McGarty, Terrence. (2002). The Internet Protocol (IP) and Global Telecommunications Transformation.
3. M.Mellia, A. Pescapè, L. Salgarelli. Traffic classification and its applications to modern networks. Elsevier Computer Networks, Dec. 2008
4. T. Farah, L. Trajkovic. Anonym: A tool for anonymization of the Internet traffic. In IEEE 2013 International Conference on Cybernetics (CYBCONF), 2013, pp. 261-266.
5. Cascarano N, Ciminiera L, Risso F. Optimizing deep packet inspection for high-speed traffic analysis. Network System Manager. 2011 19(1), pp. 7-31.
6. S. Kumar and P. Crowley. Algorithms to Accelerate Multiple Regular Expressions Matching for Deep Packet Inspection. In Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM '06), 2006, New York, USA, pp. 339-350.
7. D. Ficara, S. Giordano, G. Procissi, F.Vitucci, G.Antichi, A. Di Pietro. An Improved DFA for Fast Regular Expression Matching. SIGCOMM Comput. Commun. Rev. 38, 5 (September 2008), pp. 29-40.

8. F. Yu, Z. Chen, Y. Diao, T. V. Lakshman, and R. H. Katz. Fast and Memory-Efficient Regular Expression Matching for Deep Packet Inspection. In Proceedings of the ACM/IEEE symposium on Architecture for networking and communications systems (ANCS '06). 2006, New York, USA, pp. 93-102.
9. S. Kumar, B. Chandrasekaran, J. Turner, and G. Varghese. Curing Regular Expressions Matching Algorithms From Insomnia. In Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and communications systems (ANCS '07). 2007,New York, USA, pp. 155-164
10. R. Smith, C. Estan, S. Jha, and S. Kong. Deflating the Big Bang: Fast and Scalable Deep Packet Inspection with Extended Finite Automata. In Proceedings of the ACM SIGCOMM conference on Data communication (SIGCOMM '08). 2008, New York, USA, pp. 207-218.
11. El-Maghraby, Reham & Mostafa, Nada & Bahaa-Eldin, Ayman. (2017). A survey on deep packet inspection. 188-197. 10.1109/ICCES.2017.8275301.
12. P. Gupta and N.McKeown, Algorithms for packet classification, IEEE Network Magazine. vol.15, no.2, pp. 24-32, 2001.
13. M.L. Bailey, B. Gopal, M.A. Pagels, L.L. Peterson, and P. Sarkar, PathFinder: A pattern-based packet classifier, Proceedings of the First Symposium on Operating Systems Design and Implementation, pp. 115- 123, 1994.
14. Moore AW, Papagiannaki K (2005) Toward the accurate identification of network applications. In: PAM, Springer, vol 5, pp 41–54
15. Madhukar A, Williamson C (2006) A longitudinal study of p2p traffic classification. In: Modeling, Analysis, and Simulation of Computer and Telecommuni- cation Systems, 2006. MASCOTS 2006. 14th IEEE International Symposium on, IEEE, pp 179–188
16. Hullár, Béla & Laki, Sandor & György, András & Vattay, Gábor. (2010). New Methods in the Payload Based Network Traffic Classification.
17. V. Paxson Empirically derived analytic models of wide-area TCP connections, IEEE/ACM Trans. Netw., vol.2, no.4, pp. 316-336,1994.
18. V. Paxson and S. Floyd Wide area traffic: the failure of Poisson modeling, IEEE/ACM Trans. Netw., vol.3, no.3, pp. 226-244, 1995.
19. J. Nilsson Introduction to Machine Learning http://robotics.stanford.edu/people/nilsson/MLDraftBook/MLBOOK.pdf
20. Alshammari R, Zincir-Heywood AN (2011) Can encrypted traffic be identified without port numbers, ip addresses and payload inspection? Computer networks 55(6):1326–1350
21. Auld T, Moore AW, Gull SF (2007) Bayesian neural networks for internet traffic classification. IEEE Transactions on neural networks
22. Bagui S, Fang X, Kalaimannan E, Bagui SC, Sheehan J (2017) Comparison of machine-learning algorithms for classification of vpn network traffic flow using time-related features. Journal of Cyber Security Technology
23. Bengio Y (2009) Learning deep architectures for AI. Found Trends of Machine Learning
24. Crotti M, Dusi M, Gringoli F, Salgarelli L (2007) Traffic classification through simple statistical fingerprinting. ACM SIGCOMM Computer Communication Review