**D. SHINGISSOV^{1*}, V. GOIKHMAN², A. LAVROVA², SH. SEILOV¹,
YE. ZHURSINBEK¹**

*¹L. N. Gumilyov Eurasian National University, Doctoral student in the specialty
"Information Systems", Nur-Sultan, Kazakhstan*

*²Ph.D., Associate Professor at the Department of Infocommunicational systems The
Bonch-Bruевич Saint-Petersburg State University of Telecommunications,
Saint-Petersburg, Russia*

*²Master of degree at the Department of Infocommunicational systems The Bonch-Bruевич
Saint-Petersburg State University of Telecommunications, Saint-Petersburg, Russia*

*¹Dean of the Faculty of Information Technologies at L. N. Gumilyov Eurasian National
University, Nur-Sultan, Kazakhstan*

*¹Master student in the specialty "Information Security" of the L. N. Gumilyov Eurasian
National University, Nur-Sultan, Kazakhstan*

ON STABILITY OF DISTRIBUTIONS OF WHATSAPP TRAFFIC CHARACTERISTICS

This paper deals with the main methods of traffic classification and describes the functional scheme of a test bench and the test procedure. It provides the results of verifying the hypothesis about the stability of distributions of WhatsApp traffic characteristics. The delivered test results in this paper emphasize the influence of certain traffic characteristics on the final traffic distribution form. In addition, the comparison of the results obtained for the entire set of tests and the results received for individual test sets reveals the absence of other critical traffic characteristics significantly influencing the distribution form concluding in the need for further research. The paper concludes that the stability pattern of distributions of WhatsApp traffic characteristics can be obtained and visualized after more critical traffic characteristics are revealed and processed in similar tests. This paper stands as a pioneer research in assessing the traffic analysis and implementing the results in applied science.

Keywords: traffic classification, traffic analysis, WhatsApp messaging app.

Network traffic classification is one of the critical tasks for the communication and Internet service providers, as they need to know the structure of the traffic passing through their networks.

Traffic classification plays a significant role in ensuring network security, timely detection of intrusion attempts, and quality of service management.

Identifying the traffic and the applications that generate it allows the service providers to differentiate the billing policy, manage the bandwidth, and provide flexible traffic prioritization.

There are many papers devoted to the issue of traffic classification using various methods [1-12].

There are two main approaches to traffic classification:

1. Payload-based classification.
2. Classification based on the statistical analysis of traffic probability characteristics.

* E-mail корреспондирующего автора: e-mail: shingissov@gmail.com

The first method – payload-based – can be further divided into a general and advanced payload analysis.

The *general payload analysis* is based on the information contained in the IP header; normally, the following information is analyzed:

IP addresses, port numbers, and protocols.

This classification method has its limitations because not all applications have officially fixed transport layer port numbers, and some applications use ports that were intended for other applications. In addition, this method is not applicable to the encrypted or tunneled traffic.

The *advanced payload analysis* relies on deep packet inspection (DPI), which is based on the signature analysis. Signatures are unique patterns that are predefined for each known application.

The signature analysis involves various methods, such as:

- Pattern analysis, which means searching for a known fixed combination of bytes, symbols, strings inherent to this application.
- Numeric analysis, which includes the determination of packets' numerical characteristics, such as payload size, number of response packets, bit shift, and so on.
- Behavior analysis, which is based on the traffic behavior of the traffic of the initialized application, on a certain packet exchange sequence, packet size, number of opened sessions, etc.

The disadvantage of the payload-based classification method is its high implementation cost because it requires high-performance and high-speed computers and highly qualified personnel. In addition, as new applications and modifications of existing ones continue to appear, new signatures and patterns should be constantly developed and the existing databases should be regularly updated.

The second traffic classification method is based on the analysis of statistical traffic characteristics, namely:

- Characteristics of packets in a flow (packet-based)
- Characteristics of a flow in general (flow-based)
- Characteristics of flows coming from one host in the network (host-based)
- Characteristics of directed graphs of flows (graph-based)

Both methods are based on the research data, according to which network traffic has different probability characteristics that are unique to certain classes of applications, which makes it possible to identify different source applications [11].

Earlier studies [13-21] suggest that some user applications, instant messengers, in particular, show stable (unique) distributions of some probability characteristics of the traffic they generate.

To verify the proposed hypothesis, we created a test bench according to the functional scheme shown in Figure 1.

The WhatsApp messaging app was chosen as the test application. The choice was driven by the wide use of this messaging app: today, WhatsApp users amount to over 1 billion people in more than 180 countries. In 2018, the daily number of active users exceeded 450 million. The app is available both for smartphones running Android, iOS, or Windows

Phone operating systems, and for computers running Windows or MacO; there is also a web-based version of the app for Linux/Unix operating systems.

Figure 1 – Test bench scheme

A smartphone connects to the Internet via a Wi-Fi wireless access point and a switch.

The traffic passing through one of the switch ports was mirrored to a server. This allowed us to use tcpdump for writing traffic dump to a file.

An Android-based smartphone with WhatsApp 2.19.274 installed was the traffic source. The Android operating system has an advantage over iOS, which does not allow clearing the memory and cache of some apps.

Thus, selecting Android OS, which has the above features, allowed us to conduct each test under the same conditions.

To simulate the other communication party, we used another smartphone connected to the Internet outside the test zone.

Each test started with bringing the test smartphone to its default state.

The default state was as follows: mobile data is off, Wi-Fi is off, no third-party apps are initialized, memory and cache are cleared.

The traffic generated by WhatsApp was studied in two modes: chat mode and file transfer mode.

In the chat mode, three messages were sent from the test smartphone one after another: "Hello", "How are you", "What are you doing; the chat screenshot is shown in Figure 2.

In the file transfer mode, a 1 MB file was sent.

The chat mode tests followed the below procedure:

- Bring the test equipment to its default state.
- Enable writing of the mirrored traffic on the server using tcpdump.
- Enable a Wi-Fi connection on the test smartphone and connect to an access point.
- Start the WhatsApp app.
- Selecting a contact in the chat list.
- Type and send "Hello".
- Type and send "How are you".

Figure 2 – Chat window

- Type and send "What are you doing".
- Note the check mark indicating that all three messages have been delivered.
- Close the WhatsApp app.
- Clear cache and memory.
- Disconnect from the Wi-Fi network -> default state.
- Stop tcpdump.

In the file transfer mode, the following procedure was used:

- Bring the test equipment to its default state.
- Enable writing of the mirrored traffic on the server using tcpdump.
- Enable a Wi-Fi connection on the test smartphone and connect to an access point.
- Start the WhatsApp.
- Selecting a contact in the chat list.
- Send a file.
- Note the check mark indicating that the file has been delivered.
- Close the WhatsApp app.
- Disconnect from the Wi-Fi network.
- Stop tcpdump.
- Connect to another Wi-Fi network with access to the Internet.
- Open the WhatsApp app.
- Delete the file from the chat for both conversation participants.
- Close the WhatsApp app.

To minimize the influence of random factors, the tests were performed on different days of the week and at different times of the day. In total, we performed six sets of tests and analyzed them both individually and collectively.

The first step in processing the test data was to identify the traffic directly related to WhatsApp among all the traffic flows. To do this, we used Wireshark, a traffic analyzer that, among other things, allows filtering the traffic based on the specified parameters.

To isolate the flows generated by WhatsApp, we used the following filter: "ip.address==<WhatsApp IP-address>". In addition, we further checked that we had correctly selected the traffic flow related to WhatsApp by analyzing the responses to DNS queries for the WhatsApp domain name. The IP address in the DNS response should have been the same as the IP address used for filtering the traffic.

After the filtering and DNS-based validation, the filtered traffic was exported to a CSV file and then processed using Microsoft Excel.

Processing the test data was aimed at verifying the hypothesis about the stability of distributions of such WhatsApp traffic parameters as the packet intervals and the packet length. Given that specific values of these parameters appear randomly, we are dealing with a probability distribution.

We treated the data obtained in similar tests as samples taken from a general set and drew corresponding histograms.

For the histograms of packet intervals in the chat mode, we selected bins with a 5 ms increment. Then we divided the entire range of possible values of packet intervals into 31 bins, for which we calculated how often the corresponding values occur and drew the histograms. Figure 3 shows an example of such a histogram.

We also drew histograms for packet length values, with a 10-byte increment for the bins of possible values. Figure 4 shows an example of such a histogram.

Figure 3 – Example of a histogram for packet intervals in the chat mode

Figure 4 – Example of a histogram for packet lengths in the chat mode

Similar calculations were made for the file transfer mode. Figures 5 and 6 show examples of packet interval and packet length histograms for the file transfer mode.

Figure 5 – Example of a histogram for packet intervals in the file transfer mode

Figure 6 – Example of a histogram for packet lengths in the file transfer mode.

Similar histograms were drawn for each test.

To verify the proposed hypothesis about the stability of distributions, it is essential to confirm that distributions of the obtained samples are uniform.

For this purpose, we used one of the goodness-of-fit tests, namely, the Wilcoxon signed-rank test. We checked the distributions' uniformity both for samples obtained in the chat mode and those obtained in the file transfer mode, with tests performed separately for the packet intervals and the packet length in each mode. The significance level was assumed to be 0.05.

Let's consider the adopted analysis procedure based on the example of packet intervals in the chat mode.

The random pairwise analysis of the corresponding samples to check if they pass the Wilcoxon signed-rank test did not allow us to confirm that the tested distributions were uniform, but it showed that a large number of samples lean towards uniformity of distributions.

To evaluate how close the distributions are to each other, we compared their two initial moments. The average value was used to estimate the first initial moment. Figure 7 shows the range of average values obtained for each sample.

Figure 7 – Range of average values of packet intervals for the entire set of tests

The average value in this range was 24.186 ms. To compare all values in the above range, the proximity threshold was set to 15% relative to the calculated average value. The limit values for the obtained deviation band will be 27.81 ms and 20.56 ms, which are shown with yellow lines in Figure 8.

If the values of the first initial moments of a sample are within the above limits, we can consider such samples suitable for the comparison of their second initial moments.

Figure 8 – 15% deviation band from the average value for the first initial moments.

The values shown in Figure 8 were divided into three subranges:

- a) less than the lower limit of the deviation
- b) greater than the lower limit and less than the upper limit of the deviation – within the deviation band
- c) greater than the upper limit of the deviation

The results are given in Table 1.

Table 1 – Number of values of the first initial moments that fall into the respective subranges

Subrange	Number	Percentage, %
<20,558	36	18%
20,558-27,814	126	64%
>27,814	34	17%
Total	196	100%

The obtained data show that only 64% of the values are within the deviation band.

Then we drew a graph of the second initial moments for the samples with the above values, on which the deviation band of 15% of the average value was overlaid – see Figure 9.

Figure 9 – Graph of the second initial moments for the packet intervals

This figure demonstrates that 45 samples out of 200 are within the deviation band. Thus, only 22.5% of samples differ from each other no more than by 15% in terms of their initial moments' values.

The above results describe the entire set of tests carried out in the chat mode.

However, the pattern is quite different if we address each of the six sets of tests separately, not taken as a whole, consider that the tests were conducted at different times.

A similar analysis performed within each set of tests allows us to conclude as follows.

- In the first set of tests, 66% of samples fall into the 15% deviation band.
- In the second set of tests – 52%.
- In the third set – 58%.
- In the fourth set – 67%.
- In the fifth set – 73%.
- In the sixth set – 50%.

The following results were obtained when we analyzed the samples of data for the packet length distribution in the chat mode.

The percentage of values of the first and second initial moments within the 15% deviation band is as follows:

- For all the tests – 38%.
- For the first set of tests – 84%.
- For the second set of tests – 94%.
- For the third set of tests – 78%.
- For the fourth set – 99%.
- For the fifth set – 79%.
- For the sixth set – 70%.

We obtained similar results for the analysis of samples in the file transfer mode.

The results allow us to conclude as follows:

- We cannot fully confirm that the distributions of the samples obtained during the Wilcoxon signed-rank tests are uniform as expected.
- We observed that the distributions of a large number of samples lean towards uniformity, which is confirmed by the comparison of estimates of the first two initial moments for the mentioned samples.
- The packet length distributions in the file transfer mode lean towards uniformity to the greatest extent.
- The observed difference between the results obtained for the entire set of tests and the results received for individual test sets indicates that there are still some factors, which have not yet been identified but influence the test results and, therefore, require further research.

REFERENCES

- 1 Zeba Atique Shaikh / An Overview of Network Traffic Classification Methods / Zeba Atique Shaikh, D.G.Harkut/ International Journal on Recent and Innovation Trends in Computing and Communication (IJRITCC) Volume:3 Issue:2 //2015;
- 2 Abirami Sivaprasad / Machine Learning based Traffic Classification using Statistical Analysis / Abirami Sivaprasad, Neha Ghawlkhar, Maitri Sanghavi, Srushti Hodge, Vidhuya Shindle/ (IJRITCC) Volume:6 Issue:3 // 2018;
- 3 Marcin Pietrzyk/ Methods and Algorithms for Network Traffic Classification/ M Pietrzyk/ Telecom Paris Tech// 2011;
- 4 Shahbaz Rezaei / Deep Learning for Encrypted Traffic Classification: An Overview / Shahbaz Rezaei, Xin Lui / SUBMITTED TO IEEE COMMUNICATIONS MAGAZINE //2019;
- 5 Silvio Valenti / Reviewing Traffic Classification / Silvio Valenti, Dario Rossi, Alberto Dainotti, Antonio Pescapè, Alessandro Finamore, Marco Melia / “Data Traffic Monitoring and Analysis: From Measurement, Classification, and Anomaly Detection to Quality of Experience” //2013;
- 6 Muhammad Shafiq/Network Traffic Classification Techniques and Comparative Analysis Using Machine Learning Algorithms/ M.Shafiq, X. Yu, L. Yao, F. Abdessamia/2016 2nd IEEE International Conference on Computer and Communications //2016;
- 7 A. Dainotti / Issues and future directions on traffic classification/ A. Dainotti, A. Pescapè / IEEE Network 26 //2012;
- 8 Manuel Crotti / Traffic Classification through Simple Statistical Fingerprinting / M. Crotti, M. Dusi, F.Gringoli / ACM SIGCOMM Computer Communication Review: Number 1Volume 37 //2007;

9 N. G. Shcherbakova / Analiz IP-trafika metodami Data-Mining. Problemy klassifikatsii [IP Traffic Analysis Using Data-Mining Methods. Classification Problems] / Vychislitelnye i setevye resursy //2012;

10 A. I. Getman / Obzor zadach i metodov ikh resheniya v oblasti klassifikatsii setevogo trafika (Overview of Problems and Solution Methods in the Network Traffic Classification Field) / A. I. Getman, Yu. V. Markin, E. F. Evstropov, D. O. Obydenkov / Proceedings of ISP RAS, Volume 29, Issue 3//2017;

11 R. Suguna / A survey on Network Traffic Classification Techniques / R. Suguna / International Journal of Pure and Applied Mathematics // 2017;

Cisco: Cisco WAN and Application Optimization Solution Guide [online]. http://www.cisco.com/c/en/us/td/docs/nsite/enterprise/wan/wan_optimization/wan_opt_sg/chap05.html ;

12 M. V. Avramenko, V. Yu. Goikhman "M2M RESHENIE DLYA EKOLOGICHESKOGO MONITORINGAOKRUZHAYUSHCHEYSREDY"["M2MSOLUTIONFORENVIRONMENTAL MONITORING"], Collected Papers of the V International Scientific, Technical, and Methodological Conference on Topical Problems of Infotelecommunications in Science and Education. 2016.

13 M. V. Avramenko, V. Yu. Goikhman. "Vybor discipliny peredachi dannykh sensornogo uzla [Selecting the Sensor Node Data Transfer Discipline] – Pervaya Milya, No. 2, 2017, pp. 34-39.

14 D. D. Bekmambetova "Issledovanie Internet-trafika kvartirnogo sektora" [Internet Traffic Research in the Residential Sector], Proceedings of APINO, 2018

15 M. A. Buranova, V. G. Kartashevsky, M. S. Samoilov "Analiz statisticheskikh kharakteristik multimedijnogo trafika uzla agregatsii v multiservisnoy seti" [Analysis of Statistical Characteristics of Multimedia Traffic of an Aggregation Node in a Multiservice Network], Radiotekhnicheskie i telekommunikatsionnye sistemy, Murom, 2014.

16 A. I. Getman, E. F. Evstropov, Yu. V. Markin, Analiz setevogo trafika v rezhime realnogo vremeni: obzor prikladnykh zadach, podkhodov i resheniy [Real-Time Network Traffic Analysis: Overview of Applications, Approaches, and Solutions], http://www.ispras.ru/preprints/docs/prep_28_2015.pdf;

17 V. Yu. Goikhman, T. Yu. Lushnikova, Yu. A. Zinenko "Analiz trafika studencheskogo obshchezhitia [analysis of traffic in a dormitory]", Vestnik Kibernetiki. 2016. No. 1 (21).

18 V. YU. Goikhman, T. YU. Lushnikova "Kharakternye cherty kampusnogo trafika" [TYPICAL CHARACTERISTIC OF CAMPUS TRAFFIC], ELEKTROSVYAZ, NO. 4, 2017 PP. 51-55

19 V. Yu. Goikhman, A. Dremina "Statisticheskii analiz setevogo trafika, vznikayushchego pri zagruzke fotografii v socialnye seti" [STATISTICAL ANALYSIS OF THE NETWORK TRAFFIC WHEN UPLOADING PHOTOS TO SOCIAL NETWORKS] Collected Papers: "Topical Problems of Infotelecommunications in Science and Education", Collected Papers of the V International Scientific, Technical, and Methodological Conference. 2016.

20 V. Yu. Goikhman, N. A. Sokolov "Otsenka rosta intensivnosti vkhodyashchego trafika" [Estimation of the Incoming Traffic Intensity Growth], Elektrosvyaz, No. 3, 2018, pp. 75-77.

**Д. С. ШИНГИСОВ¹, В. Ю. ГОЙХМАН², А. Р. ЛАВРОВА², Ш. Ж. СЕИЛОВ¹,
Е. Ш. ЖҮРСІНБЕК¹**

¹Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық жүйелер мамандығының докторанты, Нұр-Сұлтан, Қазақстан

²Федералды мемлекеттік білім беру бюджеттік жоғары кәсіптік білім беру мекемесінің инфокоммуникация жүйелері кафедрасының доценті, т.ғ.к., «проф.

М.А. Бонч-Бруевич атындағы Санкт -Петербург мемлекеттік телекоммуникация университеті», Санкт -Петербург, Ресей

²Федералды мемлекеттік білім беру бюджеттік жоғары кәсіптік білім беру мекемесінің инфокоммуникация жүйелері кафедрасының магистрі, «проф. М.А. Бонч-Бруевич атындағы Санкт -Петербург мемлекеттік телекоммуникация университеті», Санкт-Петербург, Ресей

¹Л. Н. Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық технологиялар факультетінің деканы, Нұр-Сұлтан, Қазақстан

¹Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Ақпараттық қауіпсіздік мамандығының магистранты, Нұр-Сұлтан, Қазақстан

WHATSAPP ТРАФИГІНІҢ СИПАТТАМАЛАРЫН БӨЛҮДІҢ ТҰРАҚТЫЛЫҒЫ ТУРАЛЫ

Мақалада трафикті жіктеудің негізгі әдістері талқыланады, сынақ стендінің функционалдық диаграммасы мен тестілеу процедурасы сипатталған. WhatsApp трафигінің таралу тұрақтылығы туралы гипотезаны тексеру нәтижелері ұсынылады. Бұл мақалада ұсынылған тест нәтижелері трафиктің кейбір жеке сипаттамаларының оның таралуында тұрақты динамиканың пайда болуына елеулі әсер көрсететінін анықтайды. Сонымен қатар, барлық тест жинағы үшін алынған нәтижелер мен жеке тест жинағы үшін алынған нәтижелерді салыстыру барысында трафиктің таралуын толық тұрақтандыруға септігін тигізетін трафиктің басқа маңызды сипаттамаларының жоқтығын көрсетеді. Демек, жоғарыда ескерілген ақпарат, бұл тақырып қосымша зерттеулерді қажет ететінін білдіреді. Мақалада, WhatsApp -тың трафиктік сипаттамаларының таралуының тұрақты заңдылығын анықтап оны бейнелей алатын дәрежеге жету үшін трафиктің таралуын толық тұрақтандыруға елеулі септігін тигізетін трафиктің сипаттамалары анықтап, сынақ барысында өңдеу қажет деген қорытынды жасалады. Бұл мақала трафикті талдауды бағалаудағы және нәтижелерді қолданбалы ғылымға қолданудағы алғашқы зерттеу болып табылады.

Түйін сөздер: трафик классификациялау, трафикті талдау, WhatsApp хабар алмасу қосымшасы

Д. С. ШИНГИСОВ, В. Ю. ГОЙХМАН, А. Р. ЛАВРОВА, Ш. Ж. СЕИЛОВ, Е. Ш. ЖУРСИНБЕК

¹докторант по специальности «Информационные системы» Евразийского национального университета им. Л.Н.Гумилева, Нур-Султан, Казахстан

²Доцент кафедры инфокоммуникационных систем федерального государственного образовательного бюджетного учреждения высшего профессионального образования «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича», Санкт-Петербург, Россия

²Магистр кафедры инфокоммуникационных систем федерального государственного образовательного бюджетного учреждения высшего профессионального образования «Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича», Санкт-Петербург, Россия

¹Декан факультета информационных технологий, Евразийский национальный университет им. Л.Н.Гумилева, Нур-Султан, Казахстан

¹Магистрант по специальности «Информационная безопасность» Евразийского национального университета им. Л.Н.Гумилева, Нур-Султан, Казахстан

ОБ УСТОЙЧИВОСТИ РАСПРЕДЕЛЕНИЙ ХАРАКТЕРИСТИК ТРАФИКА WHATSAPP

В статье рассмотрены основные методы классификации трафика, описаны функциональная схема испытательного стенда и процедура тестирования. Приведены результаты проверки гипотезы о стабильности распределений характеристик трафика WhatsApp. Результаты тестирования, представленные в этой статье, подчеркивают влияние определенных характеристик трафика на конечную форму распределения трафика. Вдобавок, сравнение результатов, полученных для всего набора тестов, и результатов, полученных для отдельных наборов тестов, показывает недостаток других критических характеристик трафика, значительно влияющих на форму конечного распределения для выявления стабильного поведения. Следовательно, упомянутое выше приводит к необходимости дальнейших исследований. В статье делается вывод о том, что стабильное поведение распределения характеристик трафика WhatsApp может быть получено и визуализировано в том случае, когда будут выявлены и обработаны многие критичные характеристики трафика в аналогичных тестах. Эта статья представляет собой пионерское исследование в области оценки анализа трафика и внедрения результатов в прикладную науку.

Ключевые слова: классификация трафика, анализ трафика, приложение для обмена сообщениями WhatsApp.