Traffic Analysis of Mobile Broadband Networks

Geza Szabo, Daniel Orincsay, Balazs Peter Gero, Sandor Gyori, Tamas Borsos
TrafficLab, Ericsson Research, Budapest, Hungary
E-mail: {geza.szabo, daniel.orincsay, balazs.peter.gero, sandor.gyori, tamas.borsos}@ericsson.com

ABSTRACT

Detailed knowledge about the traffic mixture is essential for network operators and administrators, as it is a key input for numerous network management activities. Several traffic classification approaches co-exist in the literature, but none of them performs well for all different application traffic types present in the Internet. In this study we compare and benchmark the currently known traffic classification methods on network traces captured in an operational 3G mobile network. Utilizing the experiences about the strengths and weaknesses of the existing approaches, a novel combined method is proposed aiming at improving the completeness and accuracy of classification. The novel method is based on a complex decision mechanism, which can provide appropriate identification for each different application type. As a main contribution, with the help of the new method it is shown that applications previously used only in fixed access networks may appear in mobile broadband environment.

1. INTRODUCTION

With the evolution of mobile systems, the bandwidth capabilities of the packet switched services improved significantly. Currently, the access rates of 3G networks are comparable with the low segment of the access rates observed in fixed networks. As a consequence, applications that were present only in fixed broadband networks earlier are also appearing in mobile traffic. The change in the composition of traffic mixture may have high impact on the operation of the mobile access as well as the mobile core networks.

In [7] the authors reported that the traffic mix in volume is composed of two-third web browsing and the rest is mail, HTTPS, FTP. Today the traffic mix is different. Web traffic remained considerable, but P2P traffic gained relevance in some network situations. Such well-known services as e-mail, filetransfer, etc. together give only about 5% of the total traffic volume. HTTP traffic which is used for web browsing is about 40% of the total traffic volume. It is also interesting to note that, other applications than web browsing which use port 80 to communicate, e.g., for firewall traversing would deceive base traffic classification methods and would show that the proportion of web traffic is still the dominant.

In this paper we describe our experience with implementing and using existing traffic classification methods to analyze traffic traces captured in a live 3G mobile network. Further, we describe the set of rules and heuristics we devised and implemented in order to overcome the deficiencies of the individual methods available in the literature. As a result of continuous development, we now have a traffic classification tool prototype, which we use on a daily basis to analyze traces from various real networks, and to provide input to traffic modeling, network design and dimensioning as well as node design projects.

2. BACKGROUND

Currently, there are a couple of fundamentally different approaches for traffic classification. In the most common method the classification is based on associating a well-known *port number* to a given traffic type, e.g., web traffic is associated with TCP port 80 [2]. This method needs access only to the header of the packets. The port based method becomes insufficient in many cases, since no specific application can be associated to a dynamically allocated port number, or the traffic classified as web may easily be something else tunneled via HTTP.

In another method the classification is based on predefined byte signatures to identify the particular traffic types, e.g., web traffic contains the string 'GET'. The common feature of the *signature* (a.k.a. payload) *based methods* is that in addition to the packet header, they also need access to the payload of the packets. The payloads are processed by searching predefined byte signatures [1] in them. The main disadvantage of the signature based method is that the signatures have to be kept up to date, otherwise some applications can be missed, or the method can produce false positives. The other disadvantage is that this method cannot deal with encrypted content. The port and signature based methods can be referred as the classical traffic classification methods.

Another approach is the *connection pattern based method* presented in [4] (BLINC), where the basic idea is to look at the communication pattern generated by a particular host, and to compare it to the behavior patterns representing different activities/applications. The connection patterns describe network flow characteristics corresponding to different applications by capturing the relationship between the use

WICON 2007, October 22-24, Austin, USA
Copyright © 2007 978-963-9799-12-7
DOI 10.4108/wicon.2007.2157

of source and destination ports, the relative cardinality of the sets of unique destination ports and IPs as well as the magnitude of these sets. In the connection pattern based method the application specific behavior patterns are often difficult to find, especially if the network node uses multiple applications types simultaneously. In order to identify a communication pattern reliably, the method needs a lot of flows coming from and going to the host.

In *statistics based classification* some statistical feature of the trace is grabbed and used to classify the network traffic. To automatically discover the features of a specific kind of traffic, the statistical methods are combined with methods coming from the field of artificial intelligence. The most frequently discussed method is the Bayesian analysis technique as in [6], [10], [5]. The main problem with these techniques is that network traffic that had been previously hand-classified provides them with training and testing datasets, where the ratio of these data-sets are about 1:1.

A useful aid in traffic classification is introduced in [9] which is an *information theoretic approach* and can group the hosts into typical behaviors e.g., servers, attackers. The main idea is to look at the variability or randomness of the set of values that appear in the five-tuple of the flow identifiers, which belong to a particular source or destination IP address, source or destination port. The information theoretic approach can not be used for flow level traffic classification in the same way as the other methods. It is just an aid in traffic classification and arises the problem that it can only specify very broad application types but not capable of classifying specific applications.

One can draw the conclusion that none of the available traffic classification methods can provide a solution that is good enough on its own. Therefore, this study presents a novel traffic classification method which combines the existing methods described above, in order to improve the completeness as well as the accuracy of the traffic mixture identification process.

3. COMBINED CLASSIFICATION METHOD

It is difficult to construct one general method for traffic classification which would grab all the specific features of each application type, thus combining different approaches is very reasonable and definitely lacked yet. However, the way to do this can not be in an ad-hoc manner, thus, measurements of the reliability of the different methods are needed to make it possible to construct the decision mechanism in the most proper way. (The benchmark of the different traffic classification methods can be found in Section 4.)

3.1 The built-up of the system

The idea is to combine the results of multiple independent traffic classification modules with a decision mechanism in order to more accurately classify a flow.

The suggested system consists of several modules as it can be seen in Figure 1: there are three main modules: the signature based classification module, the port based classification module and the heuristics based classification module. These main modules have three additional submodules to preprocess their input data: the information theoretic classification submodule, the statistics based classification submodule and the connection pattern based classification submodule. (See Section 2 for further details.)

The input of the system is a packet level network traffic trace. This trace is the direct input of the statistics based classification method and the signature based method as information such as e.g., packet interarrival time, packet size distribution, packet payload, etc. is available in the packet level trace, but it is not available in the flow level trace. As our presumption is that no previous information such as preclassified trace segment is available during classification, we can not use Bayesian classification which requires an accurately classified training trace.

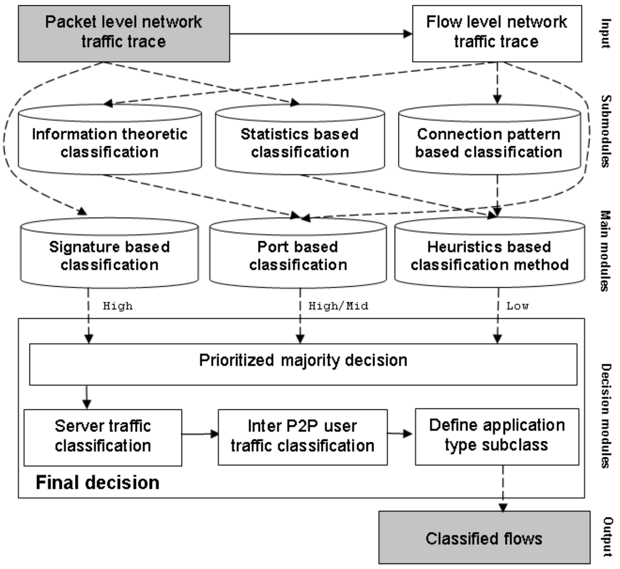

Packet level network traffic trace
Flow level network traffic trace
Input
Information theoretic classification
Statistics based classification
Connection pattern based classification
Submodules
Signature based classification
Port based classification
Heuristics based classification method
Main modules
High
High/Mid
Low
Prioritized majority decision
Server traffic classification
Inter P2P user traffic classification
Define application type subclass
Final decision
Decision modules
Classified flows
Output

Figure 1: Built-up of the system

During classification, the input of the port based method, the information theoretic approach and the connection pattern based method is the flow table, which contains a record for every flow present in the trace. It is possible to classify traces without payload but in that case the results of the signature based method can not be taken into account during the final classification mechanism, which can decrease the accuracy of the final decision.

In the heuristics based submodule, we use the statistics based classification and connection pattern based classification methods together with further heuristics to combine their results and get as specific application type as possible without examining application specific port or bitstring information. As the statistical method is usually capable of classifying the traffic but the result frequently happens to be too general or fall into other application type which exhibits similar statistical properties as the one which has been selected, thus it needed to be further classified with BLINC and additional heuristics. The introduced heuristics can be found in [8] in details.

The flows in the flow table are classified using all of the methods separately. After the flow level classification with the classification modules, the final decision mechanism combines the results with an intelligent decision. Beside these modules the system can be extended with other classification modules as well, in order to handle the continuously emerging new applications. In that case the decision mechanism may need to be modified based on the information about the accuracy of the newly added specific classification module.

3.2 The final decision

After running of all the independent classification modules, and having managed to process the statistics based classification and connection pattern based classification output, we are ready to reach a final decision on the application type of every flow. In this section we introduce how the heuristic, the port and payload based classification results can be combined in the most appropriate way.

The novelty of the final decision mechanism is that it has been constructed on the basis of the empirical accuracy of the classification methods, where the more accurate method is taken into account with higher priority. This final decision mechanism involves majority decision and novel heuristics which select the most applicable result.

The operation of the final decision mechanism can be seen in Figure 1. The signature based method is capable of finding the most specific feature of an application, the signatures of the application layer protocol, so its results are considered with the highest priority during classification. The connection pattern based method is the less accurate one, that is why its results are considered the weakest sign during classification. The decision mechanism is constructed to use these signs from the strongest to the weakest direction. As there can be multiple output of a main classification module e.g., long flow contains multiple application protocol signatures, thus the final decision module decides based on the highest priority majority of the outputs.

If the classification result is a very common application type (e.g., web, P2P) then the port list and the signatures belonging to this common application type are checked. This step may reveal the specific application which is responsible for generating the flow (e.g., identifying Kazaa as the subtype within the P2P application type).

During classification, the most simple case is that if only the result of one method is available, then we have to accept that. As a preprocessing of the output of the signature based and port based classification modules we do the following: as several P2P protocols use the HTTP protocol for communication, if the payload based method result is P2P and the result of the port based method is web or vica versa, then we convert both of the results to P2P.

Packet payload available. We compare the results of the port based and the payload based methods and if several matches occur as flows may have been constructed from different types of packets, we decide on the strictest one. This case can occur e.g., when the inspection of the first packet of a P2P application suggests that it is a simple HTTP request with the 'GET' signature then it is revealed in later packets from the 'hash' signature that it is actually a P2P application. Thus two types of applications appeared on the same flow in different packets and we decide for the stricter one. If the result of the payload based method exists but the final class can not be decided, as the port based result and the payload based result differ, then we leave the decision to the advanced port based classification mechanism.

Server and dedicated port identification. The usual port based classification method has been extended with a decision strength to save the information on the accuracy of the port based classification for every decision. The decision strength can be introduced by the extension of the port based system with the result of the information theoretical classification method presented in [9]. This method can be used to search for servers and server ports as well. BLINC can find the servers as well: it uses simple thresholds, and therefore it can find servers only with plenty of flows while the information theoretic approach can work with smaller number of flows as well.

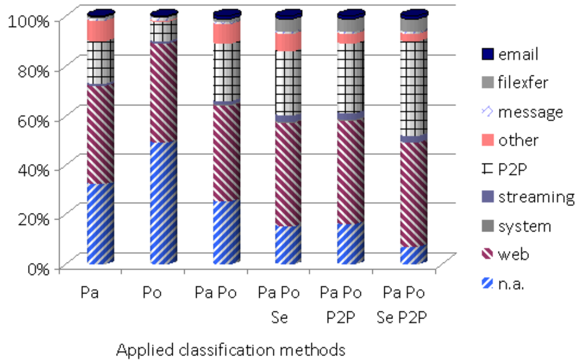


Figure 2: The effect of classification phases

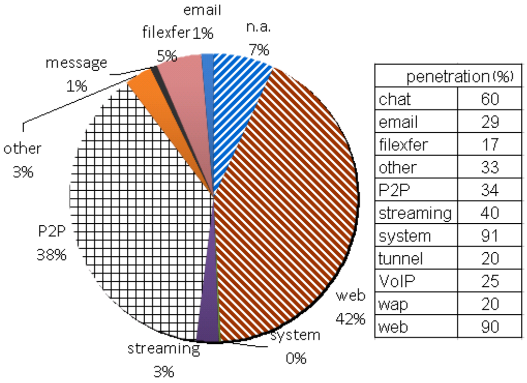
To find the application of a dedicated port, the execution of the port based method is followed by an additional heuristic which collects those flows where the port is selected by the outlying entropy based on the information theoretic method, but the application is not known based on the port number. For every IP, for every given port which has such a flow, the results of the payload based method is collected and the most specific application among the results of the payload based method carrying the highest transported data volume is selected. The result of the port based method is considered as the previously selected application with high confidence from now on. With this heuristic every dedicated port of an IP which is used by e.g. P2P application or Skype and have any flow classified by the payload based method can be marked as the port of that specific application.

After the previous steps, the next task is to loop through on all the bidirectional flows and if the two directions of a bidirectional flow are classified for different types of application then the stricter is chosen and both directions of the flow are set to that.

Inter P2P user traffic classification. In the case of P2P applications, Windows SMB service, and the passive FTP protocol the high volume of data goes through a flow having dynamically allocated port numbers. To classify them we would need to parse the whole protocol, that is why we use a simpler method: we collect those IPs where P2P occurred and the unknown flows which go among a group of these IPs are classified as the specific application that the IPs belong to. This procedure is similar to the one presented in [3] to mark the possible P2P peers. With this method we use again the concept that the unreliable heuristics are only accepted if it is corroborated by the outcome of other methods as well or if no other information is available at all.

4. BENCHMARK OF EXISTING SOLUTIONS

In [8] we systematically compared the different traffic classification methods to each other to see how they can perform in case of a specific application type. In this paper we show



penetration (%)	
chat	60
email	29
filexfer	17
other	33
P2P	34
streaming	40
system	91
tunnel	20
VoIP	25
wap	20
web	90

Figure 3: Application volume share

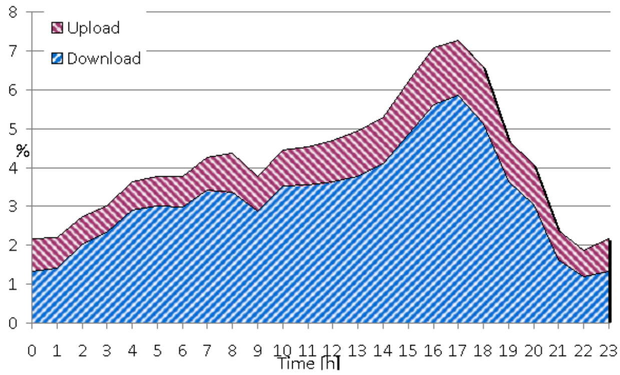


Figure 4: Upstream/downstream profile over time

the effects of the different methods on the traffic mix. From the final results both the classification completeness and accuracy can be examined and easily interpreted in case of the different application types.

In Figure 2 it can be seen that purely the byte signature based method [*Pa*], which is the most accurate, but less complete leaves 30% of the traffic unclassified. In the examined network, about 60% of the total P2P traffic is identified, and few percent P2P traffic is misclassified to web traffic.

The port based method [*Po*] on its own can classify smaller part of the traffic than the payload based method on its own, but its result can not be neglected mainly as we will see, that it is still able to classify such flows which the signature based method did not. In case of examined network, the web traffic ratio is decreased, which can be explained by the fact that many applications use HTTP protocol for other services: e.g., MSN Messenger applies it for the transmission of chat messages, which use other ports than 80.

Applying the result of both the port based and byte signature based method together [*PaPo*], we can see that the unclassified ratio decreased in the network. The increase of the P2P traffic ratio shows that plenty of traffic occurs in case of P2P when no byte signature can be applied on the flows: e.g., searching for other vanished peers with TCP SYNs, or in the case of a control flow existence like in the case of Directconnect, filetransfer flow contains only the data and not any control data which can be easily noticed.

The dedicated port identification [*PaPoSe*] contributes to the completeness by reducing the unclassified traffic volume with a few percent. Actually it is main contribution is the accuracy improvement which made those flows exactly classifiable which can not be classified neither based on a simple port based method nor the payload based method.

The inter P2P traffic heuristic [*PaPoP2P*] reduces the unclassified traffic to about 10% and obtains a few percent in volume from the web traffic as well. This means that web-like traffic which would be classified as HTTP based on its used port number or byte signature is actually goes among P2P applications.

Applying all the methods [*PaPoSeP2P*] gives the final results.

5. ANALYSIS OF THE TRAFFIC MIX

We worked on a three day long trace collected in an operational mobile broadband network. In Figure 3 the traffic mix in volume can be seen. The types of applications used by subscribers have a large influence on the network traffic. It can be seen that in the network the web browsing and P2P application take most of the bandwidth. Due to the fact that web browsing is downlink dominant, it contributes with lesser amount to the uplink direction. In the network, the share of the P2P traffic and the web traffic stays constant during daytime and the share of the P2P traffic comparing to other traffics grows for the night and early morning hours. The hours of the growth of P2P traffic volume share coincide the hours of the growth of the uplink traffic volume share. Apart from the web and the P2P traffic, there is a considerable amount of streaming traffic. Due to the fact that the streaming traffic is downlink dominant, it only contributes to the downlink traffic share. Other applications such as e-mail or FTP gives only a few percent of the total traffic.

In Figure 4 the upload/download traffic ratio over time can be seen. The uplink/downlink volume share shows the ratio of data sent and received by subscribers. The share of the uplink traffic stays constant slightly above 1% of the total traffic during daytime and grows for the night hours. This is the effect of the missing of web application during night which is downlink dominant, thus its traffic does not contribute to the downlink traffic at night.

Activity can be examined, e.g., on a daily or busy-hour basis. The number of subscribers who used packet services during an average weekday divided by the total number of subscribers is the activity ratio. This ratio can be used as an activity factor for HSPA subscribers. The ratio of active users comparing to the total number of users over time has similar characteristics as the upstream/downstream profile over time in Figure 4 with a 23% peak at the 17-19 period. A subscriber is considered active user if it sends some amount of data in a certain time period. In particular we considered a user active if it has at least 1 Mbyte traffic. About 20% of the subscribers in the network send at least a packet during an hour. During the active user identification only uplink packets were considered as activities. Downlink packets were not considered because they can be misleading as port scans would raise the activity values significantly. In particular, port scans raise the number of active users by roughly a factor of two, because one Internet source scans a lot of client IP addresses by sending them a packet one by one. The reason for the high user activity ratio in the examined network can be the higher ratio of PC cards compared to the

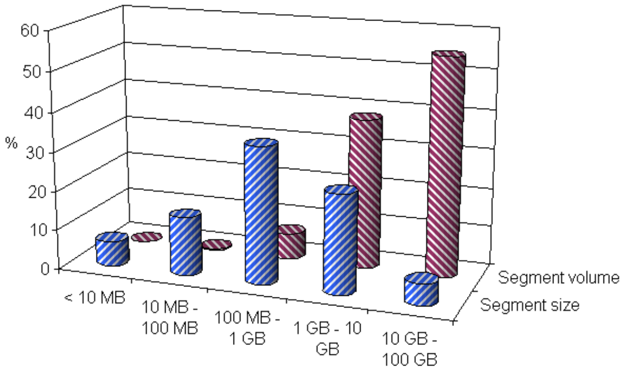


Figure 5: User segments and their contribution to the total traffic

handheld terminals. Based on the number of subscribers and the daily traffic volumes, the average per subscriber daily traffic demand can be calculated. With the assumption that the average per subscriber daily traffic demand stays the same for a month, the average monthly traffic demand can be estimated.

Besides the application volume share, it is also important to investigate what portion of the whole subscriber set uses a certain application regularly. In the right side of Figure 3 the application penetration can be seen. Browsing web is very common among the users, and this type of traffic generates system messages as visiting web pages generates DNS queries. Chat applications are very common among the users. It is interesting to note that only about 30% use e-mail applications, which is due to the popularity of instant messaging applications which are also capable to store the messages and send to the users when they become online, the other reason is the popularity of webmail services. In the consequence of these factors, the traditional e-mail protocols popularity has fallen back. From the figures we can notice that DNS (system) traffic is in correlation with web browsing, P2P traffic in does not generate significant DNS traffic.

In Figure 5 the user clusters and their contribution to the total traffic can be seen. Various subscriber segments can be differentiated based on their monthly traffic volume. Basically, subscriber segments can be presented in two different ways: a) Segment sizes: it is shown what portion of the whole subscriber set falls into the given category predefined by a monthly traffic range. b) Volume share: In the case of volume shares, it is shown what portion of the whole traffic volume of the network (in terms of bytes) is generated by the segments. The two metrics together give a complete picture about the population and traffic contribution of heavy, light, and intermediate subscribers. Figure 5 shows that 80% of the subscribers in Network B transfer less than the average. This is caused by the fact that there are subscribers with significantly higher usage.

6. SUMMARY

The identification of the network traffic mixture is essential for network operators and administrators. In this study the currently known traffic classification methods are benchmarked on network traces captured in operational broadband mobile networks. Examining the results of the different traffic classification methods, their accuracy varies for different application types. Basically, none of them can provide a solution that is capable of identifying correctly all traffic types present in various networks. We have showed the advantages and drawbacks of the different types of traffic classification methods. Using this knowledge we introduced a novel traffic classification method. The novel traffic classification method combines the different traffic classification methods and introduces a set of new heuristics, aiming at improving both the completeness and the accuracy of the traffic mixture identification process. The main benefit of the novel approach from the point of view of operators and administrators is that the ratio of the unclassified traffic decreases significantly. As another advantage, the reliability of the classification process improves, since the various methods can confirm the results of each other. Moreover, the application types can be identified more specifically, by applying the various methods successively. The new combined traffic classification method was applied on several packet traces in order to demonstrate the improvement compared to the existing methods. Moreover, the traffic of a real 3G network was analyzed with the help of the proposed method. During the analysis it turned out, that in accordance with the increased access capacity of mobile broadband networks, applications that were previously present only in fixed access networks appear and consume considerable portion of the total network traffic volume.

7. REFERENCES

[1] Application specific bit strings, http://www.cs.ucr.edu/tkarag/papers/strings.txt.

[2] IANA.TCP and UDP port numbers, http://www.iana.org/assignments/port-numbers.

[3] T. Karagiannis, A. Broido, M. Faloutsos, and K. Claffy. Transport Layer Identification of P2P Traffic. In *Proc. IMC*, Taormina, Sicily, Italy, October 2004.

[4] T. Karagiannis, K. Papagiannaki, and M. Faloutsos. BLINC: Multilevel Traffic Classification in the Dark. In *Proc. ACM SIGCOMM*, Philadelphia, Pennsylvania, USA, August 2005.

[5] A. McGregor, M. Hall, P. Lorier, and A. Brunskill. Flow Clustering Using Machine Learning Techniques. In *Proc. PAM*, Antibes Juan-les-Pins, France, April 2004.

[6] A. W. Moore and D. Zuev. Internet Traffic Classification Using Bayesian Analysis Techniques. In *Proc. SIGMETRICS*, Banff, Alberta, Canada, June 2005.

[7] F. Ricciato, P. Svoboda, E. Hasenleithner, and W. Fleischer. On the impact of unwanted traffic onto a 3g network. In *2nd International Workshop on Security and Trust in pervasive and Ubiquitous Computing*, Lyon, Frankreich, June 2006.

[8] G. Szabó, I. Szabó, and D. Orincsay. Accurate traffic classification. In *Proc. IEEE WOWMOM*, Helsinki, Finnland, June 2007.

[9] K. Xu, Z. Zhang, and S. Bhattacharyya. Profiling Internet Backbone Traffic: Behavior Models and Applications. In *Proc. ACM SIGCOMM*, Philadelphia, Pennsylvania, USA, August 2005.

[10] S. Zander, T. Nguyen, and G. Armitage. Automated Traffic Classification and Application Identification Using Machine Learning. In *Proc. IEEE LCN*, Sydney, Australia, November 2005.