Non-Intrusive Identification of Peer-to-Peer Traffic

A.Ulliac and B.V.Ghita

Centre for Information Security and Network Research,
University of Plymouth, Plymouth, United Kingdom
e-mail: info@cscan.org

Abstract

This research study a new way of identifying hosts and connections involved in peer to peer traffic without requiring analysing the payload. The P2P use more and more encryption and port number randomization. Therefore the traditional classification based on signature identification used by deep packet inspection systems is not longer efficient. This study provides a new solution to identity connections of a host which are related to a peer to peer exchange. The final output is to make able to command a firewall able to block only connections of a host that are classified as using peer to peer without blocking all its traffic.

Keywords

Networking, peer to peer, detection, supervised neural network

1 Introduction

Since ten years, the usage of peer to peer protocols significantly grown to permit an easy way to exchange data with personal broadband connections. It was more and more required to monitor and control the usage of those protocols. This is because it generates a lot of overload on the network infrastructures and could be involved in illegal exchange of copyrighted materials. Therefore, it has been able to use signature detection to classify the packets. This increasing of monitoring has motivated the protocols designed to make them evolving in order to escape the classification tools. Therefore new versions of those protocols use port number randomization and encipher the packets payload.

In this paper will be shown how the new algorithm is design and an evaluation will be made in order to identify its benefits and its limitations.

2 Background of payload vs. non payload analysis

2.1 Payload analysis

The payload analysis is the traditional methodology to classify the packets over a network and has been demonstrated as being efficient (Sen et al., 2004). For instance, Snort, the IDS using Deep Packet Inspection, has rules making able to identify some commonly used peer to peer protocols. For instance bittorent exchanges are identified with a specific character string into the packets. This methodology has revealed some limitation over the time. First it is not flexible and

needs to write specific rules for every single protocol and rewrite rules following the evolution or variations of a protocol. Then it is inefficient to identify a protocol when the data is encrypted.

2.2 Non payload analysis

However, previous studies have already shown that statistical analysis can be an efficient response to perform an identification not based on packets payload. Various solutions have been experimented. The first one (Karagiannis et al., 2004) is based on the design of a state machine without requiring payload. However this solution shows limitations on flexibility because of constants variables into the algorithm. Then more studies based on statistical algorithm, using various forms of neural networks or Bayesian algorithm shown their efficiency (Fuke et al., 2007; Chen et al., 2009).

2.3 Discussion

The objective of this research is to provide a new solution, first to identify internal hosts using peer to peer clients, then for each of them to identify on by one which connection is involved in this peer to peer traffic.

3 Methodology and data

3.1 Methodology

The algorithm uses a supervised neural network system. It provides flexibility and is fast at classifying data once the learning stage has been realised. The principle of the learning stage is to provide couples of inputs and outputs to the neural network. Then it will build a network structure describing a pattern of the data in order to automatically classify to the right output considering the input. Then during the execution stage, the algorithm will process input parameters to send them to the trained neural network that will provide a classification with an estimation of the accuracy. This estimation could be use later in order to evaluate the efficiency of the algorithm or if a rule should be apply on a network firewall.

3.2 Data

Two main sets of data are used for the purpose of this research. The first one (traces 1) is a sample of two categories of traces containing respectively peer to peer or some other type of network flow. The peer to peer is from torrent and emule network which represent about 90% of the market share in Western Europe (ipoque, 2007). It is used to train the neural network and to perform the evaluation of the algorithm. The second set of data (traces 2) is made from an ISP in New Zealand, provided by the University of Waikato. It is used to perform evaluation on unknown traces in order to evaluate the efficiency of the algorithm on realistic data.

4 Supervised neural network to identify P2P traffic

4.1 First part of the analysis -- Overall peer to peer detection

The first part of the algorithm take every connection related to a host during a certain period of time in order to process them in their own context. The algorithm generates four inputs for the neural network, and then two outputs are obtained at the end of the processing.

4.1.1 Ratio between the number of external IPs and external ports contacted

Peer to peer protocols use port number randomization. Therefore this attribute can be used to identify them. It reviled that each remote host proposes to download on a different port number. With 65,535 different ports, there are 0.0015% of chances to have two remote hosts transferring data from the same port number. This result that when having peer to peer, the average ration between the number of different remote peers and the number of different ports used to connection to distant machine is about 1.

$$Ratio = \frac{number\ of\ remote\ IPs}{number\ of\ different\ remote\ \ ports}$$

4.1.2 Ratio between replies from peers compared to the number of requests from the internal host

The usage of peer to peer reviled that many connections from the internal host failed when using peer to peer clients. It shows that only 40% of requests receive replies. This rate is over 80% for users only connecting to central servers. This comes from that every peers registered on trackers files are not online on the same time, when the peer to peer client tries to contact them. The second reason could be explained by recent events. For instance The Pirate Bay (torrentfreak, 2008) claimed that they added random IP address onto trackers to make very difficult investigations on trackers illegally hosting copyrighted contents. So the client when starting to contact the IP from this list will also try to contact the randomly added addresses.

$$Ratio = \frac{\sum Replies\ from\ externals\ hosts}{\sum Requests\ from\ the\ internal\ host}$$

4.1.3 Mean value of port numbers used to communicate with external hosts

Most of commons network services use well identified ports number. Most of them are under 1024. Therefore, peer to peer which is using random ports number have a large distribution of values. So the average value of every external ports used by internal hosts to download from peers is high.

$$mean = \frac{\sum remote\ hosts\ ports\ number\ value\ for\ each\ connection}{total\ of\ connections}$$

4.1.4 Standard deviation value of port numbers used to communicate with external hosts

The randomization of ports number also makes the standard deviation of the port number growing. If the ports number are distributed on every ports that can be allocated, the standard deviation will be about 30.000 which is what shown the experiments.

$$SD = \frac{\sum_{i=1}^{i=n}(X_i - mean)^2}{total\ of\ connections}\ ;\ X_i\ is\ current\ the\ i^{th}\ port\ number$$

4.1.5 Reporting of the results

Figure 1 shows the various relations explained in the section 4.1.1, 4.1.2, 4.1.3 and 4.1.4.

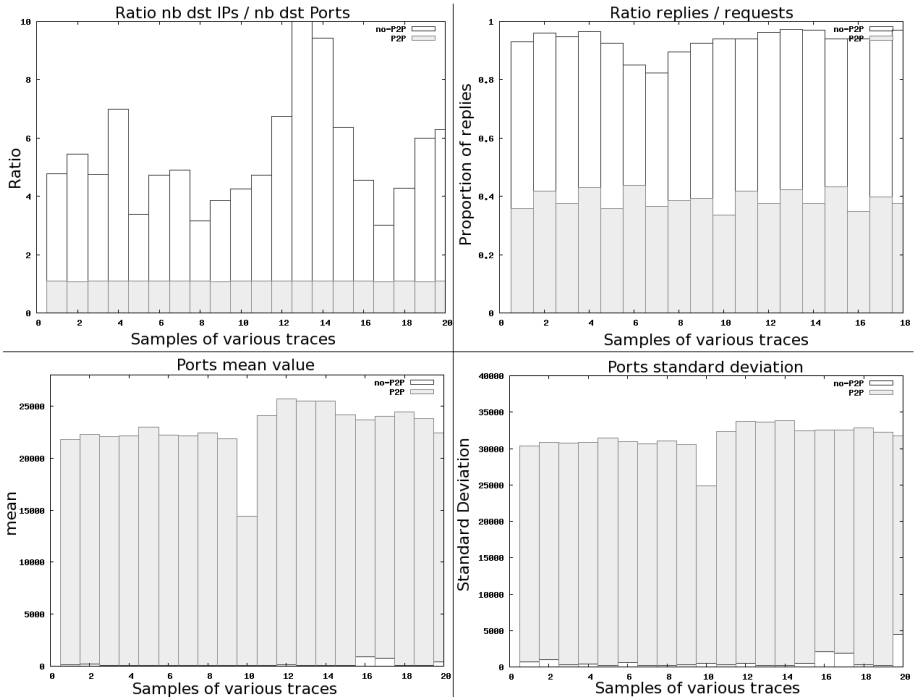

Figure 1 Analysis of the 4 parameters of the overall classification

4.1.6 Output

The output will determinate if the next part of the algorithm will be processed. It provides the probability to have to peer to peer within the group of connections that just have been processed.

4.2 Second part of the analysis – One by One connections classification

This part of the algorithm takes one by one, each connection related to a host in order to mark the one using peer to peer. The aim is to restrict specifics connections of a host without blocking all its traffic. This part of the algorithm generates 3 inputs to the neural network, which are the download and upload throughput, and the outgoing packet frequency of the connection.

4.2.1 Throughput of the connection between the two peers

The throughput between to hosts gives interesting information about the type of host involved in the exchange. Personal broadband access provides lower bandwidth and usually an asymmetric connection. For example, in UK, 80% users access the internet with ADSL (broadband-finder, 2009). So it will explain that the throughput of the exchange between two peers will never be at maximum at the upload bandwidth of the remote peer and not at the download bandwidth of the local host. So the average download throughput will be defined by the average upload bandwidth provided by ISPs. Throughput estimation formula:

$$Throughput = \frac{\sum_m^n s_i}{\Delta T} \text{ with } \Delta T = t_n - t_m$$

4.2.2 Average outgoing requests packet frequency from the internal host

The frequency of packet sent from the internal host to external peers is a good indication if the connection carries download or upload of files. It has been made the observation, that the frequency of packets is lower when there is a human interaction. So this parameter will exclude every connections involving human control like for instance web browsing or instant messaging. The frequency estimation formula is:

$$Freq = \frac{1}{\Delta T} \text{ with } \Delta T = \frac{\sum_1^N t_n - t_{n-1}}{N}$$

4.2.3 Output

The final output is an estimation of the probability of having P2P on the current analysed connection from the host. It provides a couple of values (defining a connection) which are an external IP and its remote port associated to it if the estimation reveals that it is part of P2P traffic.

4.2.4 Reporting of the results

The figure 2 illustrates the comparison of throughput and bandwidth values.

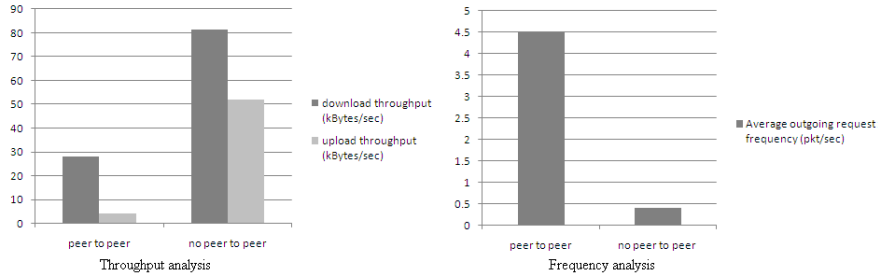

Figure 2: Analysis of the parameters of the connection by connection analysis

4.3 Summary of the algorithm

```
Capture N packets from a host

set i 11 with ratio dst IPs / dst ports on the external host
set i_12 with ratio replies / requests from the internal host
set i_13 with mean of the destination ports number to the external host
set i 14 with standard deviation of the destination ports number to the
external host

do neural network analysis with i_11 i_12 i_13 i_14

if p1 < threathold 1 //p1 is the neural network output probability on
p2p
   Stop
else
   // Here start the second part of the algorithm
   for each connection

      if connection has only request packets without reply from remote
host
         skip analysis

      set i21 with ingoing average throughput from the internal host
      set i22 with outgoing average throughput from the internal host
      set i23 with outgoing average packet per seconds from the
internal host

      do neural network analysis with i_21 i_22 i_23

      if p2 > threathold_2 //p2 is the neural network output
probability on p2p
         connection contain peer to peer
         apply filtering rule
      else
         connection does not contain peer to peer
```

5 Evaluation

5.1.1 Buffer optimisation

In order to process the input parameters of the neural network, the size of the packets buffer should be identified. This size will influent on the accuracy of the detection, on the memory size and on the processing power required. Experiments were made by gradually increasing the size of the buffer, from 50 to 2,000. It shows that the

reliability increases with the buffer size. According to the result reported in the Figure 3 the size of the buffer reach a accuracy limit around 1,000 packets.

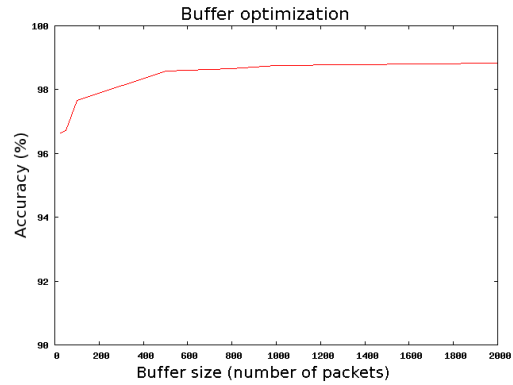

Figure 3: Classification accuracy depending on the packet buffer size

5.1.2 Evaluation of the algorithm

A first evaluation was made with the traces obtain from a 1Gbyte trace obtain from a personal gateway in front of two hosts. The host using peer to peer is detected with an average accuracy of about 98% with a standard deviation of 0.7. Then the host not using peer to peer is also classified with an accuracy of about 98% with the same standard deviation of 0.7.

The reliability of the classification of the second part of the algorithm has been processed with the same data than previously. This analysis is done on connections without requiring the context, so the same sample traces than the previous evaluation have been reused. The result is lower with about 84% of true positives. It has a larger variation around the average which is higher, with a standard deviation equal to 14.5. The Figure 4 shows this accuracy on 30 various sets of data coming from the traces 1.

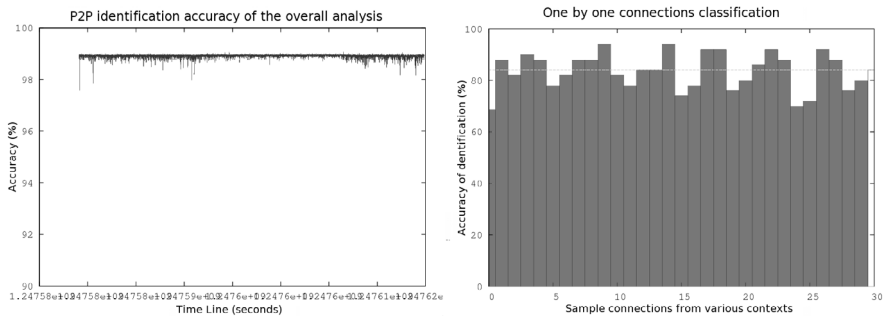

Figure 4: True positive rate on overall and one by one connection identification

5.2 Evaluation with New Zealand ISP traces

An evaluation has been made with unknown data retrieved from an ISP. It comes from a more realistic environment with more noise. The detection accuracy of peer to peer on overall detection is lower with about 85% of true positives than with the samples containing mainly peer to peer. However the true negative rate is higher, with 98%, and closer to the previous experiments. The difference can be explained by the more important quantity of noise on the network.

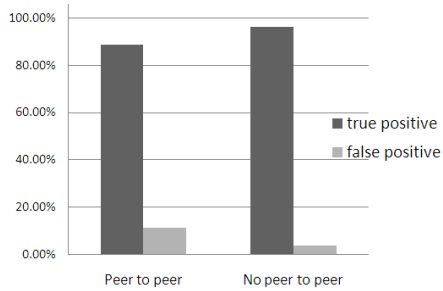

Figure 5: Evaluation of true and false positive on the traces from the New Zealand ISP

6 Discussion and implementation

6.1.1 Discussion and analysis of limitations

The algorithm is efficient when used on a gateway to identify which hosts are exchanging data on peer to peer networks. According to the evaluation tests it can vary between 85 and 98 % of accuracy depending on the noise on the network.

6.1.2 Implementation

An experimental prototype has been realised for the purpose of the research. All the analysis and output graphs are made from its output. Basically there are two ways to perform the analysis. One is offline, by analysing traces after they have been saved, and the second one is operating online. The offline analysis is useful for reporting users connected to peer to peer networks and is not intrusive. The second part of the algorithm identifying single connections is less useful in this case. But the second capability is to be able to perform online analysis to make able to stop, monitor or apply quality of service to the connections of a host (e.g. it is possible to build iptables or iproute2 rules). For this the output of the algorithm provides a couple of information which is a remote host IP with the destination port associated to it.

7 Conclusion and future work

This research shown that with a supervised neural network, it is possible to identify hosts using peer to peer on an internal network. Then the identification of peer to peer on a single connection has been made available.

On a future work it might be interesting to improve the classification methodology on single connections. This study tries to use the supervised neural network for both parts of the detection, but new statistical methodologies can be experiment to detect the peer to peer protocol on a connection.

8 References

Chen, Z., Yang, B., Chen, Y., Abraham, A., Grosan, C. and Peng, L. (2009) "Online hybrid traffic classifier for Peer-to-Peer systems based on network processors", *Applied Soft Computing,* Volume 9, Issue 2, 2009, Elsevier Science Publishers B. V., ISSN: 1568-4946.

Fuke, S., Pan, C. and Xiaoli, R (2007) "Research of P2P Traffic Identification Based on BP Neural Network", *Proceedings of the Third International Conference on International Information Hiding and Multimedia Signal Processing*, Volume 2, 2007, IEEE Computer Society, pp75-78, ISBN: 0-7695-2994-1.

ipoque (2007) "Internet Study 2007", http://www.ipoque.com/resources/internet-studies/internet-study-2007 (Accessed 4 August 2009).

Karagiannis, T., Broido, C., Faloutsos, M., and Claffy, K. (2004) "Transport Layer Identification of P2P Traffic", *Internet Measurement Conference,* 2004, ACM, pp121-134, ISBN:1-58113-821-0.

Maurizio Dusi, Manuel Crotti, Francesco Gringoli, Luca Salgarelli (2008) "Detection of Encrypted Tunnels across Network Boundaries", *IEEE International Conference on Communication*, 2008, IEEE Computer Society, ISBN: 978-1-4244-2075-9.

Robert L., H. (1994) *Neural network principles*, 1994, Prentice-Hall, ISBN: 0131121944.

Sen, S., Spatscheck, O., and Wang. D. (2004) "Accurate, Scalable In-Network Identification of P2P Traffic Using Application Signatures", *International World Wide Web Conference*, 2004, ACM, pp512-521, ISBN:1-58113-844-X.

TorretntFreak (2008) http://torrentfreak.com/the-pirate-bay-tricks-anti-pirates-with-fake-peers-081020/ (Accessed 4 August 2009).

WAND Network Research Group (2009) "WITS: Waikato Internet Traffic Storage", http://www.wand.net.nz/wits/ (Accessed 15 August 2009).