urn:nbn:de:gbv:ilm1-2011imeko-050:7

Joint International IMEKO TC1+ TC7+ TC13 Symposium
August 31st – September 2nd, 2011, Jena, Germany
urn:nbn:de:gbv:ilm1-2011imeko:2

INTERNET TRAFFIC CLUSTERING USING PACKET HEADER INFORMATION

Pekka Kumpulainen[1], *Kimmo Hätönen*[2], *Olli Knuuti*[3] *and Teemu Alapaholuoma*[1]

[1]Pori Unit, Tampere University of Technology, Pori, Finland
[2]CTO Research, Nokia Siemens Networks, Espoo, Finland
[3]KPMG Oy Ab, Helsinki Finland

Abstract – IP networks transfer huge amounts of data and information all over the word. The operator maintains the network and monitors it continuously to detect and eliminate any disturbances as soon as possible. Features and parameters included in the IP packet headers enable the operator to identify lots of information about the IP traffic and network users. The goal of this study is to analyze the measurements about the IP packet header information to support the operators in network management and optimization, trouble shooting, service creation and marketing. In this paper we propose multi-layer clustering that can reveal detailed description of traffic patterns and behaviour profiles of the network addresses. These descriptions can be enhanced by identifying abnormal traffic and analysing the reasons for such behaviour. Utilizing the measurement information included in the IP packet headers enables all these analysis tasks without violating privacy of the consumers.

Keywords: traffic pattern model, IP-traffic analysis, clustering anomaly detection, security monitoring

1. INTRODUCTION

Network traffic monitoring is an essential task in operating large-scale IP networks. The operator must ensure that the network is able to transfer sufficient traffic properly in all conditions and that the service level meets the requirements set in service level agreements. Any violations of policies, direct attacks or large deviations from the normal situation have to be detected and analysed as soon as possible. For example, old, but still growing and evolving threat for network are massive distributed denial of service (DOS) attacks against it, from it or through it. Large number of devices and the vast variety in their traffic behaviour introduce special challenges to detection of possible carefully disguised attacks or misuse of resources. Implementations of the detection and monitoring mechanisms require a lot of hardware and human resources.

Deep packet inspection (DPI) methods [1] provide information for policy enforcement, network protection and optimization purposes. DPI methods often require a lot of computation power and continuous effort to maintain the rules and patterns for identifying different higher level protocols and services used. Clustering and anomaly detection has also been suggested as methods to profile traffic to and from the servers and subscriber machines [2]. That approach was based on the use of information contained in the packet headers.

In this paper we show how multi-layer clustering can be used to monitor not only the current traffic patterns in the network, but also to characterize servers and devices that are generating the traffic. We use summarized time series data of selected variables (parameters) that describe the traffic. All the variables are computed by studying the IP packet header information. No DPI techniques need to be applied. The data are clustered in multiple phases. Clustering in the first phase divides the data into two groups of low and high levels of traffic. In the second phase these two groups are scaled and clustered separately to form a number of behavioural traffic patterns describing typical hourly behaviour. Behavioural profiles of the IP addresses are formed by studying how the traffic generated by each address is distributed between the traffic patterns. The addresses are combined to groups of similar behaviour in the third clustering phase using their proportions spent in each traffic pattern.

The information that is extracted can be used for multiple purposes ranging from intrusion and attack detection to traffic policy monitoring, service creation and marketing. For example, clustering can be used to analyse traffic and behavioural profiles of IP addresses accessing monitored services. This can be done so that the subscriber anonymity is maintained. We give also examples how clusters can support detection of anomalous traffic and changes in behaviour of addresses.

First, in the following section we introduce the application domain and provide descriptions of the measurements we use from the packet headers. In section 3 we present the clustering procedure of multiple phases. We provide descriptions to the properties of the resulting traffic patterns and behaviour profiles. We present the results of the anomaly detection in section 4. Concluding remarks are given in the last section.

2. APPLICATION DOMAIN AND MEASUREMENTS

The data were collected from a real network environment which consists of 7682 IP addresses. The privacy of the users must not be compromised, therefore the payload was dropped out and the actual IP addresses were anonymised. Information from individual packet headers of each address was summed within one hour time frames to form informative variables that describe the traffic. In this paper we use data from a period of 8 days, total of 1288770 observations of hourly traffic. The variables extracted from the packet headers are listed in Table 1. In the figures the variables are referred by their index.

TABLE I. Variables extracted from the packet headers

Index	Description of variables
1	Number of sending sequences
2	Number of sending sequences to different IP's
3	Number of receiving sequences
4	Num of receiving sequences from distinct IP's
5	Number of used port numbers <= 1024
6	Number of distinct port numbers <= 1024
7	Number of used port numbers > 1024
8	Number of distinct ports, number > 1024
9	Number of sent packets
10	Number of received packets
11	Number of sent data
12	Number of received data
13	Number of TCP connections
14	Number of UDP connections
15	Number of ICMP packets

3. TRAFFIC MONITORING

Traffic monitoring consists of several parts. First the data are divided into groups by clustering the hourly observations in two phases. The addresses are then divided to groups according to how their observations are distributed to the traffic clusters. The procedure is depicted in Fig. 1.

3.1. Clustering the traffic patterns

Clustering is a term for unsupervised methods that discover groups of similar observations in multivariate data [3]. A similarity metric is required to divide the observations to clusters. The most common distance metrics to use in clustering is Euclidean distance. It is sensitive to the scales of the variables and standardization or weighting is essential to even the effect of each variable [4, 5, 6]. The final results of clustering are strongly affected by the scaling: "*When done efficiently, weighting and selection can dramatically facilitate cluster recovery. When not, unfortunately, even obvious cluster structure can be easily missed*" [7].

The most common standardization procedure consists of subtracting the mean and dividing by the standard deviation [4]. We use a robust logarithm standardization (referred as "RLog Scaling" in Fig. 1), which has been found out to work well in mobile network monitoring data [8]. We first take a natural logarithm of the variable and then divide by a robust standard deviation: $x_{logs} = \ln(x+1) / s$, where $s = std\{\ln(x+1) \mid x > 0, x < q99\}$, and $q99$ refers to 0.99 quantile of the variable x. Finally, the remaining mean is subtracted.

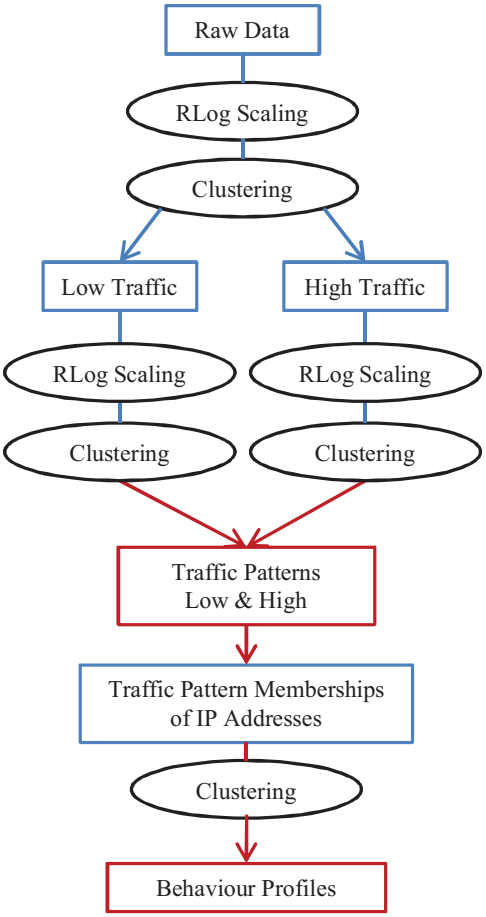

Fig. 1. Clustering procedure

Traffic patterns are created by clustering the data in two phases. The scaled data are first divided into two groups of low and high activity by k-means [9] algorithm. These groups are then scaled separately and clustered again. The optimal number of clusters is selected by Davies-Bouldin index [10]. Clustering in two phases produced traffic patterns that are more understandable and easier to interpret compared to directly clustering the data into larger number of clusters. Wide range of the volume in traffic obviously obscures the finer details that can be extracted when the high and low traffic are scaled and clustered separately. The resulting clusters are called traffic patterns as they represent the typical behaviour in the network.

The centroids of the second clustering phase presenting the mean values of the traffic and we call them traffic patterns. The patterns of both low and high traffic groups in the scaled space are depicted in the

following figures, 2 and 3. Both groups have 6 clusters, named L1 to L6 in the low, and H1 to H6 in the high traffic group. The profiles of the low and high traffic have distinct shapes. The high traffic patterns are nearly flat; all variables are on different level, except variables 5 and 6 that apply to port numbers below 1024. The patterns of the low traffic have a larger variety in their shapes.

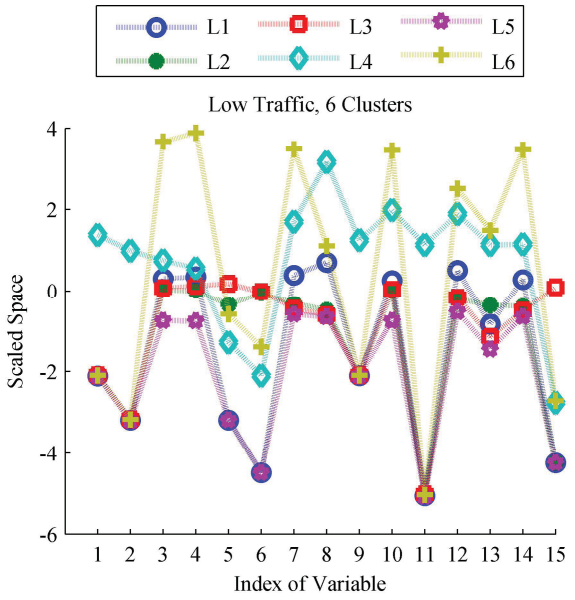

Fig. 2. Traffic patterns of the low traffic.

Fig.2. presents the six traffic patterns formed from the hourly traffic observations that were clustered to the low traffic group in the first phase. As can be seen, there is quite a lot of variation between patterns. Two most distinctive patterns seem to be L4 and L6. L4 is the only pattern in the small traffic group that sends relatively much, uses larger port numbers and also receives quite a lot of data. L6, on the other hand, receives lots of packets from several sources, doesn't send much and uses relatively much UDP protocol.

An expert can deduce what kind of service usage each traffic pattern represents. For example in pattern L4 it might be a question of P2P traffic where the client receives more data than it sends. In L6, the client equivalently sends and receives data from P2P network. Increased use of UDP protocol refers also to active use of voice over IP solutions like Skype.

Patterns L2 and L3 seem to be quite similar to each other. They differ only with regard to three variables; 5, 13 and 15. Variable 5 is 'Number of used port numbers <= 1024', where L3 has larger values. Variable 13 is 'Number of TCP connections', where L3 gets a lower value; and index 15, 'Number of ICMP packets' where L3 has relatively the highest value. Pattern L2 might refer to DNS traffic that is used to translate human readable domain names into IP addresses. Respectively, L3 refers to traffic using ICMP protocol, which is used to clarify network problems.

L1 and L5 use only few of the small port numbers, which refers to traditional client-server network traffic type of usage, where server offers quite a static set of services to the clients. This set may include services like, email, Web and FTP. L1 receives more data, uses more frequently large port numbers and uses more UDP protocol.

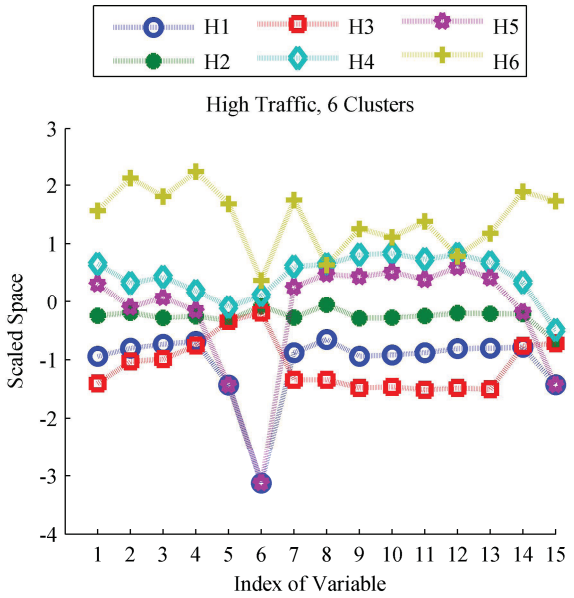

Fig. 3. Traffic patterns of the high traffic.

Patterns derived from the observations in the high traffic group are presented in Fig 3. Pattern H6 is the most distinctive one. It contains traffic samples with very high activity: lots of data sent and received both to and from several addresses and ports using all the monitored protocols.

All the other traffic patterns are more or less flat differentiating only with regard to general activity level. Three remarks can be made though.

Pattern H3 has the lowest activity but it uses relatively large number of small (<1024) ports and also UDP and ICMP protocols. Patterns H1 and H5 use relatively small amount of small ports and ICMP protocol, although otherwise they differ from each other only on general activity. Also patterns H2 and H4 differ from each other basically on the general activity, but use more small ports than patterns H1 and H5.

When comparing the low traffic patterns to the high traffic patterns, the latter are more difficult to analyze, especially by using the scaled information. In the low traffic there are clear differences between patterns but in the high traffic, scaled patterns are somewhat flat and only the differences in activity levels separate them. This can actually be considered as a sign of successful scaling: used variables are in balance and as the used scaling is logarithmic, the traffic pattern centroids are quite far from each other

in the original multivariate space. Note that the low and high groups were separately scaled and thus the values on y axis are not comparable between those groups.

The high traffic patterns H4 and H6 may refer to P2P traffic. They also most probably contain some server type traffic. Together they are responsible of 82% of the data transferred in the whole network. Patterns H1 and H5 contain traffic according to traditional client-server model because they use only a limited set of small port numbers (variables 5 and 6).

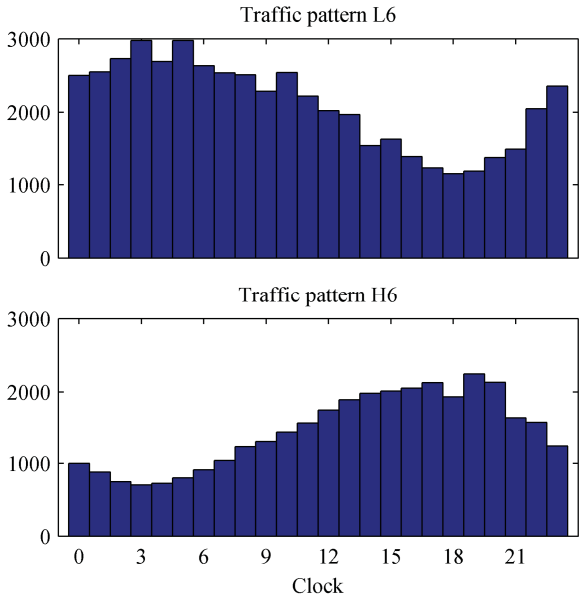

Fig. 4. Histograms of events along the time of the day in two traffic patterns.

An interesting aspect of the traffic is its distribution over the day. The activity of traffic patterns varies over the day as depicted in Fig. 4. The daily distributions of traffic patterns L6 and H6 are quite opposite. This kind of information, especially about the voluminous traffic that is represented by pattern H6, can be used in many ways in network and service planning and optimization. For example, all the regular data transfers over the network can be timed to the hours, where there is more available capacity in the network. The operator can also offer some discounts, e.g., for the customer that is using the network only outside the busy hours.

3.2. Behavioural profiles of the devices

Behavioural profiles describe how the observations of each IP address are distributed to the traffic patterns. Number of observations in the 12 traffic patterns are counted for each IP address and divided by the total number of observations to obtain proportions of activity in each traffic pattern.

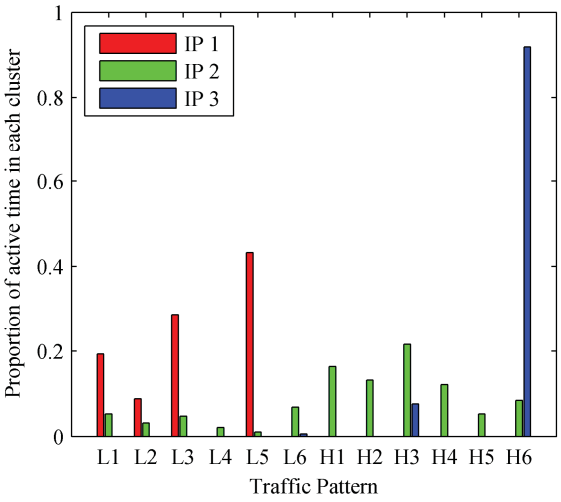

Fig. 5. Distribution of 3 IP addresses in the traffic patterns.

Fig.5 presents three IP addresses with different types of behaviour. Address IP1 has complied with four traffic patterns during the measured period. All the traffic of IP1 has been in the patterns of the small traffic group. On the contrary, a set of traffic patterns of address IP2 is very large and heterogeneous. The address has spent some time with all the traffic patterns identified from the data. The most homogeneously behaving address IP3 has complied only 3 traffic patterns – L6, H3 and H6 – during the measurement period. A small portion of its time (0.5%) that it has spent in L6 could be considered as an anomaly and the reason for it could be examined if considered necessary by the operator.

The third clustering phase divides these proportions into groups of similar behaviour. Centroids of the clusters illustrated in Fig. 6 represent the typical behaviour profiles for the addresses. Number of addresses in each behaviour profile group is given in the legend. X-axis contains the traffic clusters, 6 low traffic (L) and 6 high traffic (H).

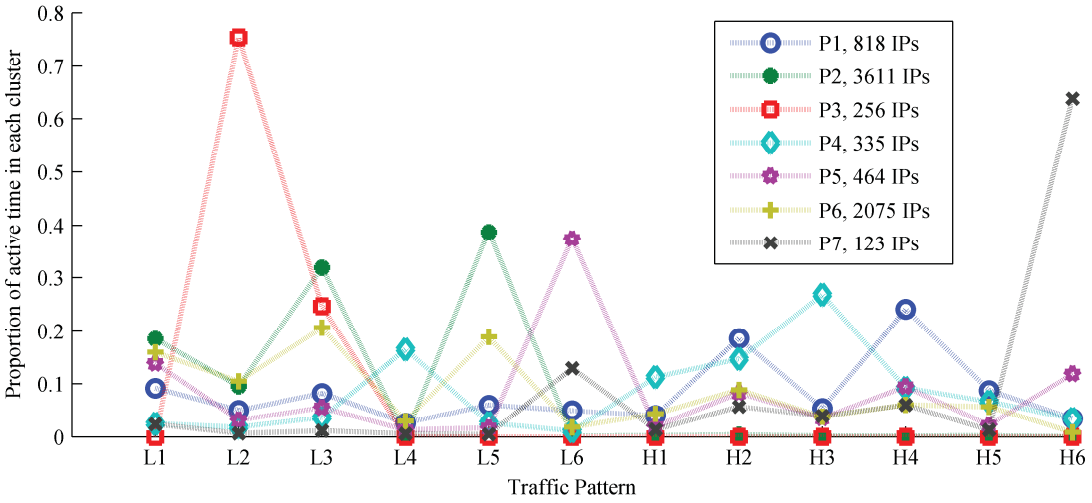

Fig. 6 Behavioural profiles of the devices

These behaviour profiles that describe the distribution of the traffic to and from of an IP address, seem to be more separated from each other than the traffic patterns. Two profiles resemble each others, namely P2 and P6. The only difference is that some of the activity of P6 has been mapped to the high traffic patterns while P2 stays purely in the patterns of the low traffic group. Both of them seem to connect traffic patterns L1, L2, L3 and L5 together, which suggest that these addresses are dynamically changing their behaviour between these patterns.

Profiles P2 and P6 are also very important and interesting as together they contain majority of the IP addresses in the monitored address space.

Profile P3, on the other hand, connects traffic patterns L2 and L3 together, as its traffic pattern is always in either one of them. Behaviour profile P3 has also a unique distribution over the time of the day. Its appearance seems to be limited to the working hours. Even the lunch break is visible in Fig.7.

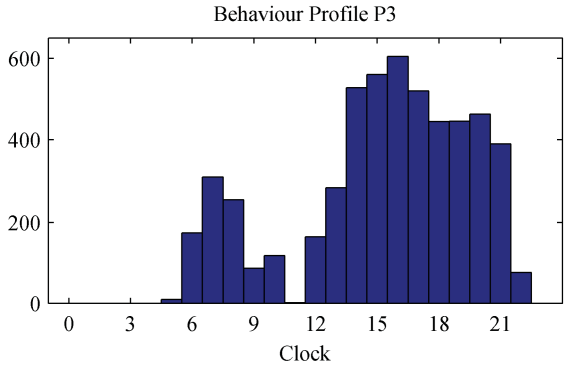

Fig. 7. Histogram of events along the time of the day in behaviour profile P3.

Profiles P5 and P7 represent similar type of activity but on distinct activity levels. P5 spends most of its time in traffic pattern L6 while P7 does the same in H6. Other parts of their behaviour they spend in the same set of traffic patterns.

Profiles P1 and P4 share their time across several traffic patterns. Addresses in profile P1 spend majority of their time in traffic patterns H2 and H4, which basically represent two different activity levels. Profile P4 contains addresses dividing their time between traffic clusters L4, H1, H2 and H3, L4 and H3 being the most frequent ones.

Altogether, there are 127 addresses that have activity in high traffic patterns only and almost half, 3465 addresses that have only low traffic behaviour.

The example addresses IP1, IP2 and IP3 in Fig.5 were assigned to the behaviour profiles P2, P4 and P7 respectively.

4. ANOMALY DETECTION

Anomaly detection (AD) is one of the core tasks in data mining. Anomalies in the internet traffic data can reveal malfunctioning equipments, new attractive network service, intrusion attempts, attacks or misuse of the resources or just some rare ways of using the net.

We use an AD method that is based on Kohonen self organizing maps (SOM) and is able to detect anomalies in local neighbourhoods [8]. The whole data set is scaled using the robust logarithm scaling before applying the AD algorithm.

It has been claimed that up to five thousand intrusion alerts per day can be handled manually by a big operator [11]. However, most operators prefer the information to be summarised to a reasonable level. The number of detected anomalies is relatively large and therefore it is necessary to summarise the information contained in the anomalies. Clustering the detected anomalies has been appreciated by the end users [12].

Anomaly clustering supports identification of possible cause of the anomalies either automatically by the system or manually by an expert. This simplifies and speeds up analysis and selection of counter measures to reveal and fix the sources of the threatening or disturbing deviation.

In this paper we selected 1000 most severe anomalies, i.e. those that deviate most from their local common behaviour. These observations in the scaled space were clustered to reveal the common patterns in anomalous behaviour. Davies-Bouldin index suggested four clusters. The Fig. 8 depicts logarithms of the means calculated from observations in the four anomaly clusters.

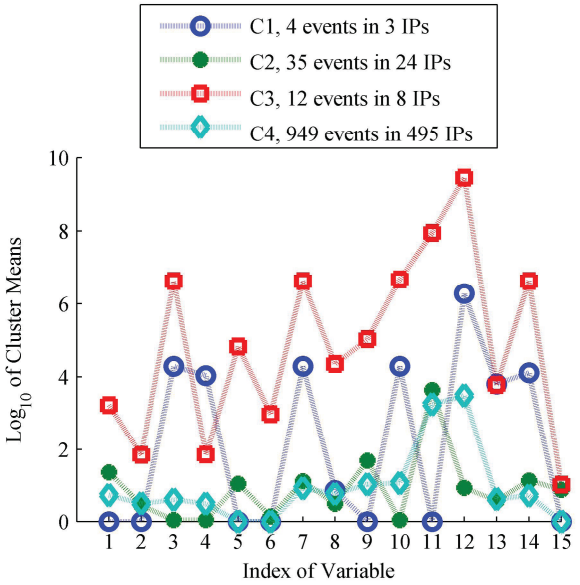

Fig. 8. Means of the anomalies in 4 clusters.

The largest one of the clusters could represent the classical client-server usage of the network but on the very low activity level. Such behaviour can be either a sign of small scale web browsing or a beginning or an end of a more active session just separated from the high activity session by the one hour summarization interval in these measurements. However, if this kind of low activity behaviour appears to the address that has been inactive for some time, it might be worth to monitor it for, e.g. botnet activity.

As shown in Fig. 8, clustering is a useful tool for grouping the anomalies and effectively decrease the work load of technicians monitoring the network. From these four groups, they can, for example, select anomalies in the clusters C1, C2 and C3 and find out what kind of addresses are responsible for them and what kind of traffic behaviour they present. Such a report would be simple to construct automatically also by the monitoring software.

5. CONCLUSION

In this paper we present a method for monitoring internet traffic utilizing the information in the package headers without detailed prior knowledge of the devices in the network. The method is based on multiple levels of clustering. First the data are divided into low and high traffic groups. They are scaled separately and clustered to identify generic traffic patterns and further behavioural profiles for individual IP addresses. The introduced knowledge about traffic and behaviour clusters can be used in several operator tasks including network management and optimization, trouble shooting, service creation and marketing.

We also apply anomaly detection for detection of abnormal behaviour in the network. The information of the detected anomalies is summarized for monitoring purposes by clustering. This enhances the network monitoring and enables the operator to detect and solve problems in the network more efficiently.

Further development is targeted towards increased utilization of anomaly detection in intrusion detection as suggested by Lippmann at al. [13]. The robustness of the identification of the traffic profiles will also be studied.

REFERENCES

[1] Bendrath, R. “Global Technology Trends, Transnational Market Forces, and National Regulation: The Case of Internet Traffic Monitoring by Deep Packet Inspection”. *International Studies Annual Convention* New York City, 15-18 February 2009

[2] O. Knuuti, T. Seppälä, T. Alapaholuoma, J. Ylinen, P. Loula, P. Kumpulainen, K, Hätönen, “Construction communication profiles by clustering selected network traffic attributes” *Proceedings, The Fifth International Conference on Internet Monitoring and Protection ICIMP* 2010, 9-15 May 2010, Barcelona, Spain. pp. 105-109, 2010.

[3] B. Everitt, S. Landau, M. Leese, “Cluster analysis. Edition:4”, *Arnold*, London, 2001.

[4] R.O. Duda, P.E.Hart, D.G. Stork, “Pattern Classification, 2nd Edition”, *John Wiley & Sons*, 2001.

[5] Kumpulainen, P., Kylväjä, M., Hätönen, K: “Importance of Scaling in Unsupervised Distance-Based Anomaly Detection” In: *Proceedings of IMEKO XIX World Congress. Fundamental and Applied Metrology.* September 6-11, 2009, Lisbon, Portugal. pp. 2411—2416, 2009

[6] Hätönen, K., Laine, S., Similä T. “Using the LogSig-function to integrate expert knowledge to Self-Organising Map (SOM) based analysis” *IEEE International Workshop on Soft Computing in Industrial Applications*, Birmingham University, New York, June 23-25, 2003. pp. 145- 150, 2003

[7] R. Gnanadesikan, J.R. Kettenring, S.L. Tsao, “Weighting and selection of variables for cluster analysis” *Journal of Classification.* Vol. 12, no. 1, pp. 113-136, March 1995.

[8] P. Kumpulainen, K. Hätönen, "Local anomaly detection for mobile network monitoring" *Information Sciences*, vol. 178, issue 20, pp. 3840-3859, 2008.

[9] J. Macqueen, "Some methods for classification and analysis of multivariate observations" *In: Proc. Fifth Berkeley Symp. on Math. Statist. and Prob*. Vol. 1. pp. 281-297, 1967.

[10] D.L. Davies, D.W. Bouldin, "A cluster separation measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1 (2). pp. 224–227, 1979.

[11] C. Kruegel, G. Vigna, "Anomaly detection of web-based attacks" *Proceedings of the 10th ACM conference on Computer and communications security*, Washington D.C., USA. pp. 251-261, 2003.

[12] M. Kylväjä, P. Kumpulainen, K. Hätönen, "Information Summarization for Network Performance Management", In: M. Laszlo, J.V. Zsolt, (eds.). *Proceedings of the 10th IMEKO TC10 International Conference on Technical Diagnostics*, Budapest, Hungary, pp. 167-172, 2005.

[13] R.P. Lippmann, D.J. Fried, I. Graf, J.W. Haines, K.R. Kendall, D. McClung, D. Weber, S.E. Webster, D. Wyschogrod, R.K. Cunningham, M.A. Zissman,. "Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection Evaluation" *DARPA Information Survivability Conference and Exposition. DISCEX '00. Proceedings,* Vol 2, pp. 12 – 26, 2000

Authors: Pekka Kumpulainen, Pori Unit, Tampere University of Technology, TUT/ASE P.O. Box 599, FI-33101 Tampere, FINLAND. P: +358 408490930, pekka.kumpulainen@tut.fi.
Kimmo Hätönen, Security research, Nokia Siemens Networks, Research, Espoo, Finland, kimmo.hatonen@nsn.com
Olli Knuuti, KPMG Oy Ab. olli.knuuti@kpmg.fi
Teemu Alapaholuoma, Pori Unit, Tampere University of Technology, Pori, Finland, teemu.alapaholuoma@tut.fi