International Journal of Electronics, Communication & Soft Computing Science and Engineering
ISSN: 2277-9477, Volume 2, Issue 3

Identification and Investigation of the User Session for LAN Connectivity via Enhanced Partition Approach of Clustering Techniques

K.Gunasekaran

***Abstract*—This paper mainly presents some technical discussions on the identification and analyze of "LAN user-sessions". The identification of a user-session is non trivial. Classical methods approaches rely on threshold based mechanisms. Threshold based techniques are very sensitive to the value chosen for the threshold, which may be difficult to set correctly. Clustering techniques are used to define a novel methodology to identify LAN user-sessions without requiring an a priori definition of threshold values. We have defined a clustering based approach in detail, and also we discussed positive and negative of this approach, and we apply it to real traffic traces. The proposed methodology is applied to artificially generated traces to evaluate its benefits against traditional threshold based approaches. We also analyzed the characteristics of user-sessions extracted by the clustering methodology from real traces and study their statistical properties.**

***KeyTerms*— Clustering methods, traffic measurement, Partition techniques, Statistical threshold, user session.**

I. INTRODUCTION

Applications like telnet typically generate a single TCP connection per single user-session, whereas application layer protocols such as HTTP, IMAP/SMTP and X11 usually generate multiple TCP connections per user-session [1] [2] [3]. User-session identification and characterization play an important role both in Internet traffic modeling and in the proper dimensioning of network resources. Besides increasing the knowledge of network traffic and user behavior, they yield workload models which may be exploited for both performance evaluation and dimensioning of network elements. Furthermore, network dimensioning problems are usually based on simple assumptions to permit analytical formulations and solutions. The validation of these assumptions can only be obtained by checking the model against traffic measurements. Finally, the knowledge of user-session behavior [4] [5] [6] is important Operators are interested in monitoring these parameters, especially today that traffic demands change very quickly as new services are continuously proposed to customers. Thus, correct user-session identification and characterization are of fundamental importance and interest. A LAN user-session, simply named user-session, discussed in this paper, the goals are

(i) To devise a technique that permits to correctly identify the user-sessions

(ii) To determine their statistical properties by analyzing traces of measured data. A user-session informal definition can be obtained by describing the typical behavior of a user accessing the LAN. An activity (ON) period on the LAN alternates with a silent (OFF) period during which the user is inactive on the LAN. This activity period, named as session in this paper.

Clustering techniques are exploratory techniques used in many areas to analyze large data sets. Given a proper notion of similarity, they find groups of similar variables/objects by partitioning the data set in "similar subsets". Typically, several metrics over which a distance measure can be defined are associated with points (named samples) in the data set. Informally, the partitioning process tries to put neighboring samples in the same subset and distant samples in different subsets.

The aim of this paper is to define a clustering technique to identify user-sessions. Performance is compared with those of traditional threshold based approaches, which partition samples depending on a comparison between the sample to sample distance and a given threshold value. The main advantage of the clustering approach is avoiding the need to define *a priori* any threshold value to separate and group samples. Thus, this methodology is more robust than simpler threshold based mechanisms associated with points (named samples) in the data set. Informally, the partitioning process tries to put neighboring samples in the same subset and distant samples in different subsets. The aim of this paper is to define a clustering technique to identify user-sessions.

The main advantage of the clustering approach [7] is avoiding the need to define a priori any threshold value to separate and group samples. Thus, this methodology is more robust than simpler threshold based mechanisms.

As previously mentioned, a common definition of a user-session is given by a period of

time during which the user is generating traffic. A user-session is then terminated by a "long" inactivity period. Within an activity period, many TCP connections may be used to transfer data. Unfortunately, the identification of active and silent periods is not trivial and the definition of the user activity may also depend on the selected application. The threshold-based approach works well only if the threshold value is correctly matched to the values of connection and session inter-arrival times. Furthermore, different users may show different idle times, and even the same user may have different idle periods depending on the service,

If the threshold value is not correctly matched to the session statistical behavior, threshold based mechanisms are significantly error prone, to avoid this drawback, we propose a more robust algorithm to identify user-session. While the server log approach can be very effective, it does not scale well and, by leveraging on a specific application level protocol, can be hardly generalized. Our methodology is rather general, and is much more robust than any threshold based approach.

II. Related Work

The earliest technique was presented by V. Paxson and S. Floyd [9] regarding "session" arrival process in. However, the focus was on Telnet and FTP sessions, where each session has related with a single TCP data connection. No measurements of HTTP sessions are reported.

Y. Fu et al. [8] proposed that an adaptation in a passive sniffing methodology to rebuild HTTP layer transactions to infer clients/users' behaviors. By crawling HTTP protocol headers, the sequence of objects referred by the initial request is rebuilt. This allows grouping several TCP connections to form a user-session.

F. D. Smith et al. [3] concluded that identification of HTTP user-sessions; by traditional approaches rely on the adoption of a threshold. TCP connections are aggregated in the same session if the inter-arrival time between two TCP connections is smaller than the threshold value.

In C. Nuzman et al. [2], η is selected to be 100s, while in [12] a threshold η =1 s is chosen. Results are obviously affected by the choice of η. Indeed, the threshold-based approach works well only if the threshold value is correctly matched to the values of connection and session inter-arrival times.

III. Clustering Techniques

In this chapter, we have briefly described the basics of clustering techniques to provide an overview of their main features in clusters, according to a notion of distance among objects. The objective is to exploit this property to group connections (objects) to identify user-sessions (clusters) in an automatic fashion.

Let us consider a metric space X, named sampling space and a set of samples $A=\{X_1, X2....X_n \in X\}$ which have to be grouped (clustered) into K subsets: The subsets in the partition are named *clusters*. Clusters contain "similar "samples, whereas samples associated with different clusters should be "dissimilar", the similarity being measured via the sample-to-sample and cluster-to-cluster distances. Now we introduce two clustering techniques that are,

A. The Hierarchical Agglomerative Approach

Each sample is initially associated with a different cluster. Then, on the basis of the definition of a cluster-to cluster distance, the clusters at minimum distance are merged to form a new cluster. The algorithm iterates this step until all samples belong to the same cluster. This procedure defines a merging sequence based on minimum distance between clusters. At each step, a *quality indicator* function is evaluated. The set is finally clustered by selecting the number of clusters such that is maximized. Intuitively, the quality indicator function measures the distance between the two closest clusters at step. A sharp increase in the value of is an indication that the merging procedure is merging two clusters which are too far apart, thus suggesting to adopt the previous partition as the best cluster configuration. This approach can be quite time consuming, especially when the data set is very large, since the initial number of clusters is equal to the number of samples in the data set . For this reason, non-hierarchical approaches, named partitional, are often preferred, since they show better scalability properties.

B. The Partitional Approach

This technique is used when the final number of clusters is known. The procedure starts with an initial configuration including clusters, selected according to some criteria. The final cluster definition is obtained through an iterative procedure. The cluster is represented by a subset of samples when measuring cluster-to-cluster distance. At procedure startup, clusters are created, with cluster centroids selected according to a given

rule in the measurement space. Each sample is associated with the closest cluster, according to the distance between the sample and the centroid of each cluster. When all samples are assigned to a cluster, new centroids are computed and the procedure iterates. The algorithm ends when either a prefixed number of iterations are reached, or the number of samples which are moved to a different cluster is negligible according to a predefined threshold. The final result may change depending on the chosen initial state.

IV. USE OF PARTITION TECHNIQUES ON THE MEASUREMENT OF DATA SET

To take the advantages and to avoid the drawbacks of both methodologies, we use a mix of them. Thus, for each, the following three-step algorithm is run to identify user-sessions:

1. An initial clustering is obtained using a partitional algorithm.
2. A hierarchical agglomerative algorithm is used to aggregate the clusters and to obtain a good estimation of the final number of clusters.
3. A partitional algorithm is used to obtain a fine definition of the clusters.

1. Initial Clustering Selection: Let's start with a partitional algorithm with clusters, with significantly smaller than the total number of samples. To efficiently position, in our uni-dimensional metric space, the representatives at procedure startup, evaluate the distance between any two adjacent samples .According to the distance metric, take the farthest couples and determine intervals. Each cluster is represented by a small subset of samples

2. The Hierarchical Agglomerative Procedure: In the second step, a hierarchical agglomerative algorithm is iteratively run, using only the representative samples to evaluate the distance between two clusters. Since the procedure starts with initial clusters, the number of steps is bounded. At each step the hierarchical agglomerative procedure merges the two closest clusters; then, distances among clusters are recomputed. After iterations, the process ends. The clustering quality indicator function permits to select the best clustering among those determined in the iterative process. Indeed, at each step, the clustering quality must be evaluated to determine if the optimal number of clusters has been found. Denote the j th cluster at step as $C_j^{(s)}$; at each step, the procedure evaluates the function $\gamma^{(s)}$

$$\gamma^{(s)} = d^{(s)}_{min} d^{-(s)}_{min} / d^{-(s)}_{min}$$

A sharp increase in the value of $\gamma^{(s)}$ is an indication that the merging procedure is artificially merging two clusters which are too far apart.

3. *Final Clustering Creation:* A partitional clustering procedure is run over the original data set, which includes all samples, using the optimal number of clusters determined so far and the same choice of cluster representatives adopted in the first step. A fixed number of iterations are run to obtain a final refinement of the clustering definition. This phase is not strictly required, since at the end of the hierarchical agglomerative procedure a partition is already available. However, it produces clusters of *real* samples instead of representatives. Furthermore, the computational cost of this phase is almost negligible if compared to the previous one.

V. PERFORMANCE ANALYSIS OF DATA SET

We have implemented the proposed technique on artificially generated traces with the following parameters.

i) To ensure its ability to correctly identify a set of TCP connections belonging to the same user-session.

ii) To assess the error performance of the proposed technique.

iii) To compare it with traditional threshold based mechanisms.

Analytical results are presented to determine the performance of threshold based mechanisms.

Finally, we run the algorithms over real traffic traces, to obtain statistical information on user-sessions, such as distributions of the following parameters.

i) Session duration

ii) Amount of data transferred in a single session

iii) Number of connections within a single session.

A study of the inter-arrival times of Web user sessions is also presented, from which it emerges that Web user-sessions tend to be Poisson, but correlation may arise due to network/ hosts anomalous behavior. Preliminary results on user-sessions statistical characterization were presented in [10].

Statistical properties of session inter-arrival times are investigated. A session arrival trace is obtained by superimposing in time all identified sessions during the same time period. This correlation is reflected by the session arrival process, which deviates significantly from the Poisson assumption [11]. To make sure that the anomalies were not introduced by our methodology, we also tried to fit the session inter-arrival distribution as identified by a threshold

procedure. The qualitative results are similar, even if the quantitative measurements are obviously different and strongly depending on the selected threshold.

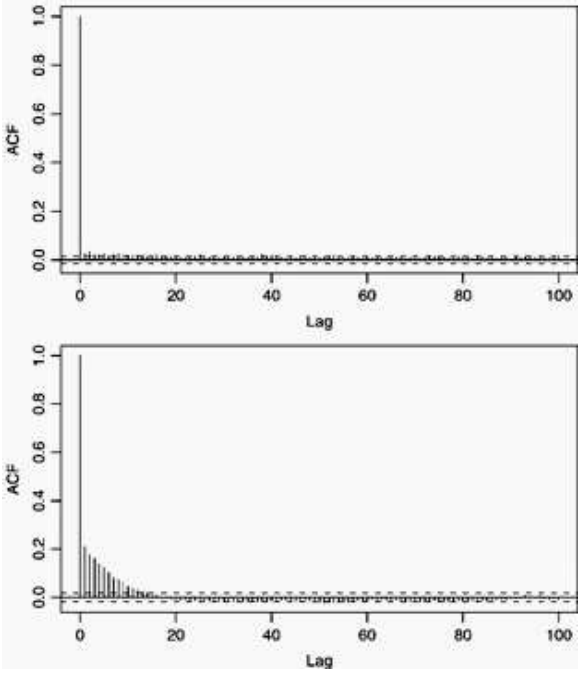

Fig1. Auto correlation function for session inter-arrival process. (Normal on top plot, warm attack on bottom plot)

The above plotted figure (Fig.1) reports the autocorrelation function evaluated on session inter-arrivals during a typical day on the top plot, while the bottom plot refers to the autocorrelation estimated during the day of the worm attack.

The top plot confirms that the Poisson assumption holds for normal days, being the autocorrelation function almost negligible except in the origin. On the contrary, as shown in the bottom plot, the autocorrelation function is quite relevant on days during which user activities are driven by external factors such as worm infection.

A. Parameter Sensitivity

We have initially evaluate the influence on performance results of the values chosen for i) the initial number of clusters K, used in the first clustering phase, and ii) the percentile , used in the hierarchical agglomerative clustering phase.

Considering the exponential connection inter-arrival scenario, in Fig. 2 the error probability is shown to be practically independent from the value of the initial number of clusters, since all curves overlap. Therefore is not a critical parameter, provided it is sufficiently larger than the number of sessions. In all the experiments, I choose for simplicity.

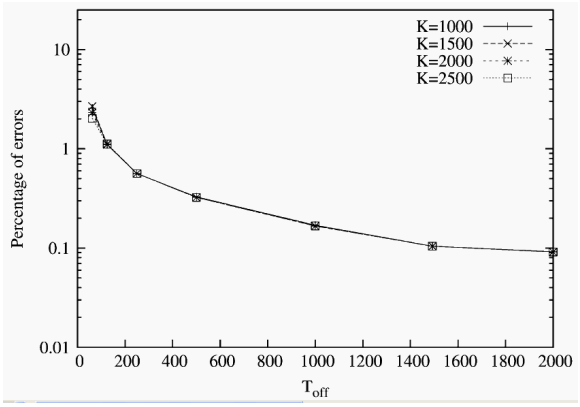

Fig2. Clustering sensitivity to the initial number of clusters K for exponential connection inter-arrival

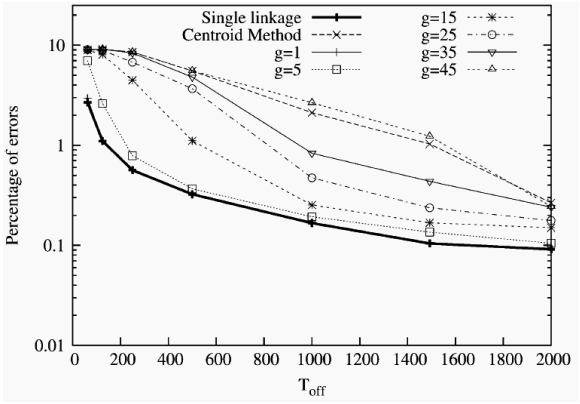

Fig3. Clustering sensitivity to the percentile $_g$ for exponential connection inter-arrival

Fig.3 shows instead the influence of the parameter that determines the value of the percentile used to select the cluster representatives in the cluster-to cluster distance.

The findings are following

1. The single linkage algorithm, which takes the two extreme values in the sample distribution as cluster representatives
2. The centroid algorithm, which uses the mean value of the sample distribution
3. The percentile algorithm, which uses the g^{th} percentiles, for variable values of $g \neq 0$. Regardless of the chosen distribution, the single linkage algorithm g=0 has the best performance, while the centroid algorithm is the worst.

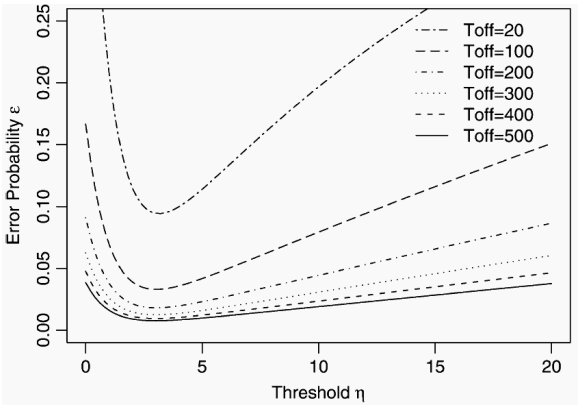

Fig4. Error probability as a function of the threshold η for exponential random variables.

Fig.4 shows that ε (η) considering different values of $\check{T}_{off}$. The percentage of errors grows for decreasing values of η. This is due to the larger probability of misidentifying a connection inter-arrival as a session inter-arrival. Similarly, large values η of lead to session inter arrivals being identified as connection inter-arrivals. Finally, when $\check{T}_{off}$ is small (i.e., comparable with $\check{T}_{on}$), the error probability becomes larger.

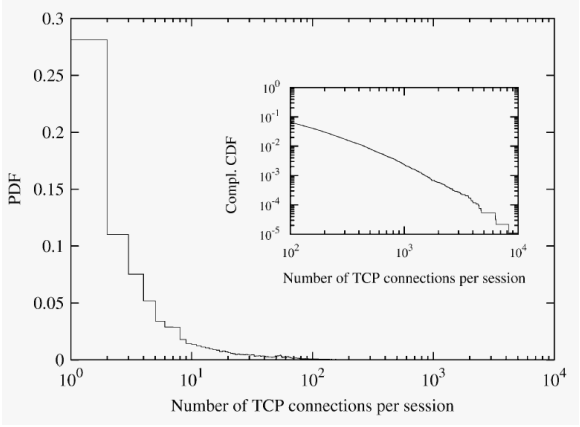

Fig5. PDF of the number of TCP connection in each session. Complementary CDF in the inset.

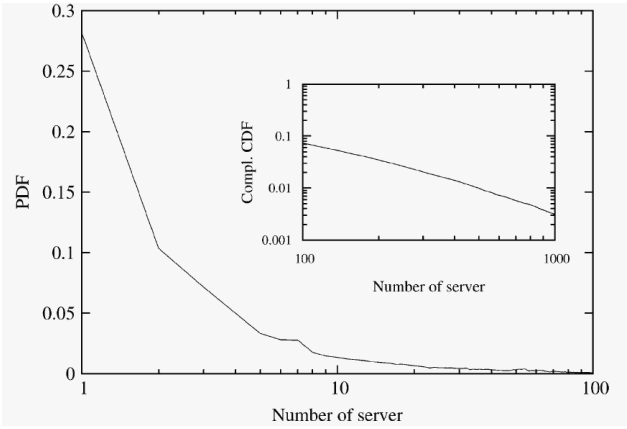

Fig6. PDF of the number of different server IP addresses per session. Complementary CDF in the inset.

Fig. 5, which reports the number of TCP connections per user-session. Indeed, more than 25% of sessions include only one TCP connection. Furthermore, most of the identified sessions are composed by very few connections (about 50% by 4 connections or less).

This demonstrates that the following three logics.

i. The client is usually able to obtain all the required data using few TCP connections

ii. The number of required external objects is limited

iii. The time spent by the users over one Web page is large enough to define each Web transaction as a session.

The CDF, reported in the inset, shows a linear trend, highlighting that the distribution has a heavy-tail.

Fig. 6 reports the PDF of the number of different server IP addresses in each session. Roughly 27% of session's aggregate connections from a single server, and about 10% of sessions refer to only two servers. However, the PDF has a heavy-tail, as highlighted by the complementary CDF, which shows that the percentage of sessions contacting more than 100 different servers is not negligible.

VI. CONCLUSION

Clustering techniques were applied to real traffic traces in order to identify LAN user-sessions. Clustering techniques can be used to identify the web user session. A novel clustering methodology was proposed and compared with the classical threshold based scheme. The effectiveness and robustness of the proposed clustering methodology was first assessed by applying it to an artificial data set, and showing its ability in the identification of Web user sessions without requiring any *a priori* definition of threshold values. Then, the proposed clustering methodology was applied to measured data sets to study the characteristics of LAN user sessions.

The analysis of the identified user-sessions shows a wide range of diverse behaviors that cannot be captured by any threshold based scheme. The clustering partition techniques proposed in this paper can be helpful in studying traffic properties at the user level, and could be easily extended to deal with other types of user-sessions, not necessarily related to Web traffic.

REFERENCES

[1] Andrea Bianco, Gianluca Mardente, Marco Mellia, , Maurizio Munafò, , and Luca Muscariello,"Web User-Session Inference by Means of Clustering Techniques" IEEE/ACM Transactions On

Networking, VOL. 17, NO. 2, pp. 405-416, APRIL 2009.

[2] C. Nuzman, I. Saniee, W. Sweldens, and A. Weiss, "A compound model for TCP connection arrivals, with applications to LAN and WAN," Computer Networks, Special Issue on Long- Range Dependent Traffic, vol. 40, no. 3, pp. 319–337, Oct. 2002.

[3] F. D. Smith, F. H. Campos, K. Jeffay, and D. Ott, "What TCP/IP protocol headers can tell us about the web," SIGMETRICS Perform. Eval. Rev., vol. 29, no. 1, pp. 245–256, 2001.

[4] M. Pioro and D. Medhi, "Routing, flow, and capacity design in communication and computer networks," The Morgan Kaufmann Series in Networking, 2004.

[5] Universal Mobile Telecommunications System (UMTS), Selection Procedures for the Choice of Radio Transmission Technologies of the UMTS (UMTS 30.03 Version 3.2.0), ETSI TR 101 112 V3.2.0 (1998-04), (1998-04).

[6] S. Ben Fredj, T. Bonald, A. Proutière, G. Régnié, and J. W. Roberts, "Statistical bandwidth sharing: A study of congestion at flow level," in Proc. ACMSIGCOMM 2001, San Diego, CA, Aug. 2001, pp. 111–122.

[7] D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining. Cambridge, MA: MIT Press, 2001.

[8] Y. Fu, A. Vahdat, L. Cherkasova, and W. Tang, "EtE: Passive end-to-end Internet service performance monitoring," in Proc. General Track: 2002 USENIX Annu. Tech. Conf., Monterey, CA, 2002, pp. 115–130.

[9] W.Willinger, V. Paxson, and M. S. Taqqu, "Self-similarity and heavytails: Structural modelling of network traffic," in A Practical Guide to Heavy Tails: Statistical Techniques and Applications, R. Adler, R. Feldman, and M. S. Taqqu, Eds. Boston, MA: Birkhauser, 1998.

[10] A. Bianco, G. Mardente, M. Mellia, M. Munafò, and L. Muscariello, "Web user session characterization via clustering techniques," in Proc. IEEE GLOBECOM 2005, St. Louis, MO, Vol. 2, pp. 1102–1107, Nov. 2005.

[11] T. Bonald, A. Proutière, G. Régnié, and J. W. Roberts, "Insensitivity results in statistical bandwidth sharing," in Proc. Int. Teletraffic Congr. (ITC) 17, Salvador, Brazil, Dec. 2001.

[12] M. E. Crovella and A. Bestavros, "Self similarity in World Wide Web traffic: Evidence and possible causes," IEEE/ACM Trans. Netw., vol. 5, no. 6, pp. 835–846, Dec. 1997.

[13] R. Caceres, P. Danzig, S. Jamin, and D. Mitzel, "Characteristics of wide-area TCP/IP conversations," in Proc. ACM SIGCOMM'91, Aug. 1991, pp. 101–112.

[14] P. Danzig and S. Jamin, "Tcplib: A library of TCP Internetwork traffic characteristics," USC, Tech. rep., 1991.

[15] P. Danzig, S. Jamin, R. Caceres, D. Mitzel, and D. Mestrin, "An empirical workload model for driving wide-area TCP/IP network simulations," Internetworking: Research and Experience, vol. 3, no. 1, pp. 1–26, 1992.

[16] V. Paxson, "Empirically derived analytic models of wide-area TCP connections," IEEE/ACMTrans. Netw., vol. 2, no. 4, pp. 316–336, Aug. 1994.

[17] P. Barford and M. Crovella, "Generating representative web workloads for network and server performance evaluation," in Proc. SIGMETRICS' 98/PERFORMANCE'98, 1998, pp. 151–160.

[18] L. Cherkasova and P. Phaal, "Session-based admission control: A mechanism for peak load management of commercial Web sites," IEEE Trans. Comput., vol. 51, no. 6, pp. 669–685, Jun. 2002.

[19] M. F. Arlitt and C. L. Williamson, "Web server workload characterization: The search for invariants," in Proc. ACM SIGMETRICS'96, Philadelphia, PA, 1996, pp. 126–137.

[20] The GARR Network Topology. 2005 [Online]. Available: http://www.garr.it/reteGARR/mappe.php

AUTHOR'S PROFILE

K.Gunasekaran,
Assistant Professor,
Department of CSE,
Pavendar Bharathidasan College of Engg & Tech,
Trichy, Tamilnadu, India.
E – Mail ID: - gunaasekarann@yahoo.co.in