Cluster Analysis of Internet Users Based on Hourly Traffic Utilization

M. Rosário de Oliveira[‡], Rui Valadas[†], António Pacheco[‡], Paulo Salvador[†]
[‡]Instituto Superior Técnico - UTL
Department of Mathematics and CEMAT
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
e-mail: rosario.oliveira@math.ist.utl.pt, apacheco@math.ist.utl.pt
[†]University of Aveiro / Institute of Telecommunications - Aveiro
Campus de Santiago, 3810-193 Aveiro, Portugal
e-mail: rv@det.ua.pt, salvador@av.it.pt

Abstract

Internet access traffic follows hourly patterns that depend on various factors, such as the periods users stay in the access point (e.g. at home or in the office) or their preferences for applications. The clustering of Internet users may provide important information for traffic engineering and tariffing. For example, it can be used to set up service differentiation according to hourly behavior, resource optimization based on multi-hour routing and definition of tariffs that promote Internet access in low busy hours.

In this work, we identify patterns of similar behavior by grouping Internet users of two distinct Portuguese ISPs, one using a CATV access network and the other an ADSL one and offering distinct traffic contracts. The grouping of the users is based on their traffic utilization measured every half-hour. Cluster analysis is used to identify the relevant Internet usage profiles, with the partitioning around medoids and Ward's method being the preferred clustering methods. For the two data sets, these clustering methods lead to 3 clusters with similar hourly traffic utilization profiles. The cluster structure is validated through discriminant analysis.

Having identified the clusters, the type of applications used as well as the flow duration and transfer rate are analyzed for each cluster resulting in coherent outcomes.

keywords: Access networks, cluster analysis, discriminant analysis, principal component analysis, Internet traffic characterization, traffic measurements.

1 Introduction

The behavior of Internet users is changing rapidly due to the continuous emergence of new applications and services and the introduction of broadband access network technologies, such as ADSL and CATV, offering large access bandwidths. An example of a recent user practice with a significant impact on access network traffic is the download of large files of music and films, using file sharing applications such as Kazaa. The characterization of Internet access traffic is important for both Internet Service Providers (ISPs) and access network operators. Due to the rapid changes in the traffic characteristics, this activity requires frequent traffic measurements.

Traffic measurements can be performed with various levels of detail. The finest possible detail level records information relative to every packet passing an observation point. Typically this information includes the arrival instant and selected header fields containing, e.g., the packet size and the origin and destination addresses. This can generate large amounts of data, placing a severe limit on the observation period. Measurements with lower resolution can be obtained by recording data at predefined times, e.g., at the (periodic) end of sampling intervals. In this case, an example of recorded data is the number of bytes and the number of packets observed in the sampling period. The resolution of a traffic measurement should match the particular task it is targeted for. For example, traffic measurements for accounting purposes only record information at the end of a login session, which is typically several hours long.

Traffic measurements can be performed over many different types of objects, e.g., all traffic downloaded by a user, the traffic aggregate on a link or all traffic generated by a specific application. An important object type is the so-called traffic flow, introduced in [1]. A flow is defined as a sequence of packets crossing an observation point in the network during a certain time interval. All packets belonging to a particular flow have a set of common properties (e.g. destination IP address, destination port number, next hop IP address, output interface). A flow is considered active as long as its packets remain separated in time by less then a specified timeout value. The IETF is developing a specification for measurement systems based on IP flows [2].

Internet traffic characterization has been the subject of several works [1, 3, 4, 5] and is addressed on a permanent basis by organizations such as CAIDA [6]. In particular, [3] analyzed the holding times and the call interarrival times of Internet access traffic at an ISDN central and [5] studied the influence of access speed on the Internet user behavior.

There are a number of network related tasks that are performed on an hourly basis. For example, off-line traffic engineering can be performed in this time scale, e.g., when updating routing periodically in order to optimize resource utilization. Another task is the definition and management of tariffing policies that may promote Internet access during the least busy hours. These resource and revenue management tasks require traffic measurements and characterization with a resolution close to one hour.

The need to provide good quality of service to Internet users calls for a detailed knowledge of the main types of individual user behavior, which may remain hidden if traffic is analyzed at an aggregate level. Moreover, effective ISP marketing strategies directed to its customers must also be based on detailed knowledge of their behavior. Aiming at the clustering of Internet users with similar statistical characteristics, we analyze in the paper the behavior of ISP customers (users) based on their hourly traffic utilization. In particular, a user is characterized by the average transfer rate of downloaded traffic in half-hour periods (over one day).

The users are grouped by means of cluster analysis [7], a set of techniques whose aim is to partition a set of objects into groups or clusters in such a way that profiles of objects in the same group are similar, whereas the profiles of objects in different clusters are distinct. Generally speaking, cluster analysis methods are of two types: hierarchical and partitioning methods. The hierarchical clustering techniques proceed by either a successive series of merges (agglomerative hierarchical methods) or by a series of successive divisions (divisive hierarchical methods). The partitioning methods start with a fixed number of clusters, each characterized by a specific object or representative, and progressively affect each of the objects to one of the clusters. In our analysis we use, in particular, Ward's method, an agglomerative hierarchical method, and the partitioning around medoids method, a partitioning method.

Any sound statistical analysis requires a preliminary data analysis aimed at highlighting the main characteristics and/or inconsistencies of the data as well as the need for transforming the data in some way. Accordingly, prior to performing the cluster analysis, we present an overview of the traffic traces analyzed and use principal component analysis (PCA) as an exploratory tool. PCA [8] is a multivariate technique whose aim is to transform a set of observed variables into a smaller number of uncorrelated variables, that maximize the explained variance.

Effective statistical analysis must be supported by validation procedures. Consequently, we validate the results of cluster analysis using discriminant analysis [9]. In particular, we compute a discriminant function that best separates the groups obtained from each cluster analysis. The validation of each obtained cluster structure is then done by evaluating the discriminant function as a classification mechanism using several associated measures, as explained in Section 5. After validation, the clusters obtained are analyzed based on various statistics of user flows.

The paper is organized as follows. Section 2 gives an overview of the traffic traces analyzed and in Section 3 principal components are used as a preliminary data analysis to the cluster analysis presented in Section 4. The results of cluster analysis are validated, using discriminant analysis, in Section 5 and the clusters obtained are characterized, based on several statistics, in Section 6. In Section 7 we describe several applications of hourly traffic profiles and, finally, in Section 8 we present our conclusions.

2 Overview of the traffic traces

Our analysis resorts to two data traces measured in two distinct ISPs, that will henceforth be designated by ISP1 and ISP2. ISP1 uses a CATV network and ISP2 an ADSL one. Both ISPs offer several types of services, characterized by maximum allowed transfer rates in the downstream/upstream directions. For ISP1 the services are (in Kbit/s) 128/64, 256/128 and 512/256 and, for ISP2, 512/128 and 1024/256. Both traces were measured on a Saturday: November 9, 2002 (ISP1) and October 19, 2002 (ISP2).

The measurements were detailed packet level measurements, where the arrival instant and the first 57 bytes of each packet were recorded. This includes information on the packet size, the origin and destination IP addresses, the origin and destination port numbers, and the IP protocol type.

The traffic analyzer was a 1.2 GHz AMD Athlon PC, with 1.5 Gbytes of RAM and running WinDump. No packet drops were reported by WinDump in both measurements.

Table 1: Statistics of the aggregate of applications.

	ISP1	ISP2
Capture date	Nov. 9, 2002	Oct. 19, 2002
Number of users	3432	875
Downloaded MBytes per user	85.6	489
Number of flows per user	1.80	1.75
Flow duration (hours)	4.9	8.6
Transfer rate (Kbits/sec)	23.9	81.1

Users were identified by matching IP addresses with accounting information. The data set of ISP1 includes 3432 users and the one of ISP2 includes 875 users.

In order to characterize the traffic traces we considered several statistics, including flow based statistics. A flow is defined as a sequence of packets with inter-arrival times smaller than 15 minutes. A flow starts upon arrival of the first packet and ends after the arrival of the last packet, that preceded the silence period greater than 15 minutes. Note that this is different from the notion of session considered in accounting systems such as RADIUS. In RADIUS, a session corresponds to the period of time a user is logged in the system, irrespective of its level of activity. In contrast, a flow reflects only packet level activity. Flows can be defined for the aggregate of applications or for each individual application of a user.

We first concentrate on several statistics for the aggregate of applications. In this case, a flow is first characterized by user, number of downloaded bytes, transfer rate and duration, without accounting for the actual applications. The addressed statistics are the averages of: downloaded MBytes per user, number of flows per user, flow duration and transfer rate. The main characteristics of the traffic traces are presented in Table 1. The downloaded MBytes per user is higher in ISP2, since this ISP offers service contracts with higher allowed transfer rates and, in addition, was not imposing any limits in the total amount of downloaded traffic by the time this measurements were carried out. This goes along with higher values of flow duration and transfer rate. However, the average number of flows per user are similar in the two ISPs. We note, in particular, the high values of flow durations, reaching average values as high as 8.6 hours.

We turn now our attention to the most typical applications. Applications were identified by port number. For ISP1 / ISP2, we enumerated all applications responsible for more than 0.1% / 0.05% of the downloaded traffic; the remaining applications were classified as “Other”. We used a higher percentage in ISP1 because a larger number of applications were observed in this ISP. Then, to allow an easier interpretation, these applications were grouped by type in the following way (we include the port number in parentheses).

- File sharing: Kazaa (TCP 1214), eDonkey2000 (TCP 4662), direct connect (TCP 412, 1412), WinMX (TCP 6699) and FTP (TCP 20, 21).
- HTTP: HTTP (80) and secure HTTP (443).
- Games: Operation FlashPoint (TCP 2234), Medal of Honor (UDP 12203), Half-Life (UDP 27005) and all traffic from the Game Connection Port (TCP 2346).
- IRC/news: IRC traffic (TCP 1025, 1026) and Newsgroups access (TCP 119).
- Mail: POP3 mail access (TCP 110).
- Streaming: MS-Streaming (TCP 1755) and RealPlayer (UDP 6970).
- Others: other applications and all unidentified traffic.

We characterize the groups of applications using again several statistics. In the case of flow based statistics, each flow was first characterized by user, application, number of downloaded bytes, duration and transfer rate. These statistics are:

- Relative (aggregate) utilization - Percentage of downloaded bytes in each application group.
- Average flow duration - Average duration of all flows belonging to an application group.
- Average flow transfer rate - Average transfer rate of all flows belonging to an application group.

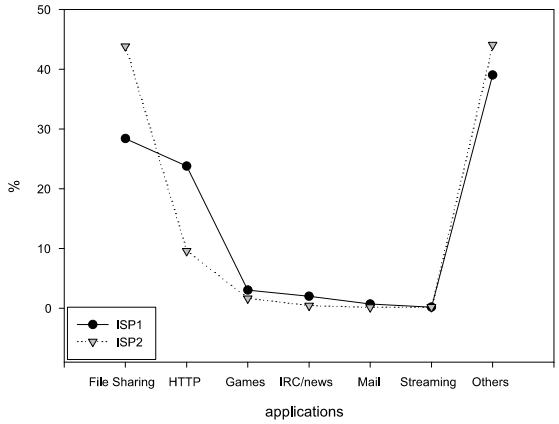

Figure 1: Relative (aggregate) utilization.

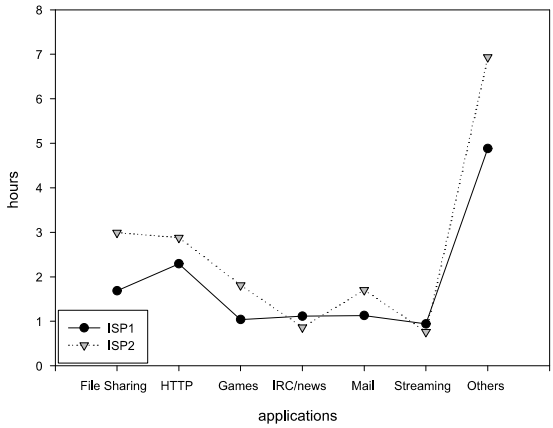

Figure 2: Average flow duration.

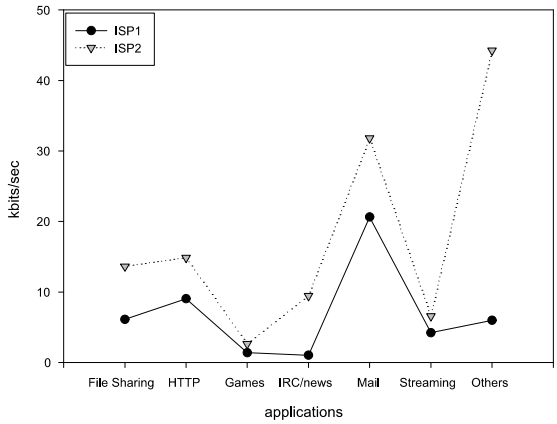

Figure 3: Average flow transfer rate.

The results are shown in figures 1, 2 and 3; we first comment the results for the identified application groups. Among these groups, file sharing and HTTP are the dominant ones in terms of relative utilization, and this is accompanied by higher flow durations. The transfer rates of these application groups are also high. Mail applications achieve relatively high transfer rates. This is due to the fact that mail downloads are mainly done from the mail server located in the ISP premises; therefore, mail transfer rate are only conditioned by the access network itself (and not by the Internet). The transfer rates are always higher in ISP2, a direct consequence of the type of service contracts offered by this ISP. In ISP1, there is not a strong difference between the utilization of file sharing and HTTP, as it is the case in ISP2. It seems that, in ISP1, a higher percentage of users is doing file transfer and sharing through HTTP. This can also explain the fact that, in ISP2, HTTP has higher flow durations than file sharing. The group called “Others” includes a significant percentage of the traffic, despite the fact that individually the port numbers assigned to this group generated few traffic (according to our criteria less than 0.1% of the downloaded bytes in ISP1 and 0.05% in ISP2). Given the high values of the flow durations and transfer rates we are led to suspect that most of these ports are being used by file sharing or video applications, that eventually distribute its traffic by a number of (non-standard) ports. This is more pronounced in ISP2 because, by the time the measurement was done, no limit was being imposed on the amount of downloaded traffic per-month free of charge.

In this paper, we are going to perform cluster analysis based on the download transfer rates measured in half-hour intervals. Figure 4 shows the (total) transfer rate of the traffic aggregates of both ISP1 and ISP2, as a function of time period, along the day. The two ISPs exhibit coherent average hourly profiles, showing a quasi-sinusoidal shape, with the lowest utilization in the morning period and the highest one in the afternoon period.

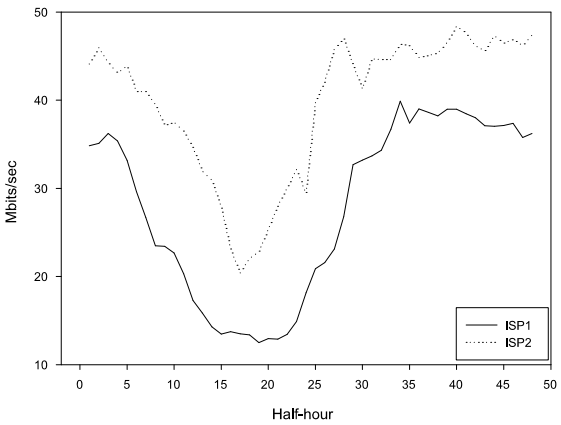

Figure 4: Total download transfer rate.

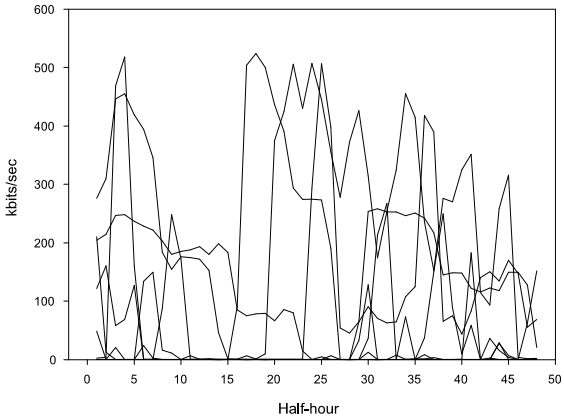

Figure 5: Download transfer rates of 10 selected ISP1 users.

The agreement between the two ISPs average hourly profiles may suggest that coherent user hourly profiles may be present across the two ISPs. However, it does not preclude that such an aggregate representation of the traffic may hide groups of users (clusters) with specific hourly profiles, markedly distinct from the average aggregate hourly profile. Evidence of the existence of very distinct user hourly profiles is given in Figure 5, where the transfer rates of 10 selected ISP1 users are plotted. The nonhomogeneous character of hourly user profiles motivates the use of cluster analysis to address the possible identification of groups of users (clusters) with similar hourly profiles, for each ISP. Of interest is also the comparison of the hourly profiles of clusters for the two ISPs, apart from the hourly traffic utilization rates.

3 Principal component analysis as an exploratory tool

A common goal in multivariate (statistical data) analysis is to explain a set of observations on several random variables using a smaller number of variables, with the new variables being function of the original ones. Principal component analysis (PCA) aims at maximizing the explained variance using as new variables, called principal components, uncorrelated linear combinations of the observed variables. Thus, PCA may be used to reduce the dimensionality of the original data. Moreover, the analysts are usually also interested in the interpretation of the principal components, i.e., the meaning of the new directions where the data is projected.

More precisely, given a set of n observations on the random variables $X_1, X_2, \ldots, X_p$, the k-th principal component (PC k) is defined as the linear combination,

$$Z_k = \alpha_{k1}X_1 + \alpha_{k2}X_2 + \ldots + \alpha_{kp}X_p \tag{1}$$

such that the loadings of Z_k, $\boldsymbol{\alpha}_k = (\alpha_{k1}, \alpha_{k2}, \ldots, \alpha_{kp})^t$, have unitary Euclidean norm, maximum variance and PC k, $k \geq 2$, is uncorrelated with the previous PCs, which in fact means that $\boldsymbol{\alpha}_i^t\boldsymbol{\alpha}_j = 0$ if $i \neq j$ and $\boldsymbol{\alpha}_i^t\boldsymbol{\alpha}_i = 1$. Thus, the first principal component is the linear combination of the observed variables with maximum variance. The second principal component verifies a similar optimal criteria and is uncorrelated with PC 1, and so on. As a result, the principal components are indexed by decreasing variance, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$, where λ_r denotes the variance of PC r and p is the maximum number of PCs.

It can be proved [8] that the vector of loadings of the k-th principal component, $\boldsymbol{\alpha}_k$, is the eigenvector associated with the k-th highest eigenvalue, λ_k, of the covariance matrix of the observed variables. Therefore, the k-th highest eigenvalue of the covariance matrix is the variance of PC k, i.e. $\lambda_k = \text{Var}(Z_k)$.

The proportion of the total variance explained by the first r principal components is

$$\frac{\lambda_1 + \ldots + \lambda_r}{\lambda_1 + \ldots + \lambda_p}. \tag{2}$$

If this proportion is close to one, than there is almost as much information in the first r principal components as in the original p variables. In practice, the number r of considered principal components should be chosen as small as possible, taking into account that the proportion of the explained variance, (2), should be large enough.

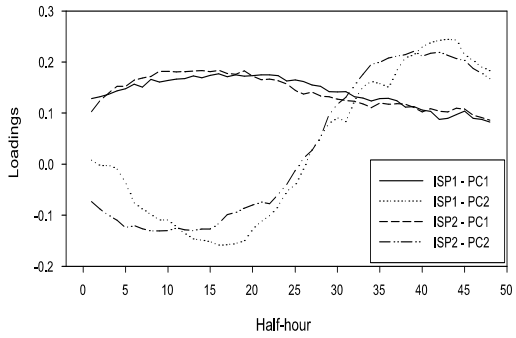

Figure 6: Loadings of the first two principal components for ISP1 and ISP2.

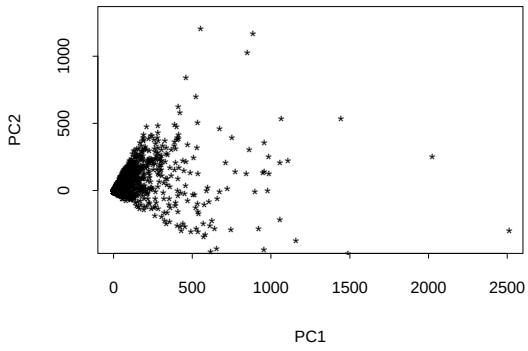

Figure 7: Scores of the first two principal components, ISP1.

Once the loadings of the principal components are obtained, the score of individual i on PC k is given by

$$z_{ik} = \alpha_{k1}x_{i1} + \alpha_{k2}x_{i2} + \ldots + \alpha_{kp}x_{ip} \tag{3}$$

where $(x_{i1}, \ldots, x_{ip})^t$ is the data corresponding to individual i. The scores (z_{i1}, z_{i2}) on the first two PCs give a graphical representation of the projection of the data on the first two PCs.

In this paper, PCA is used as an exploratory tool and applied to the data corresponding to the transfer rate (in Kbits/s) in the k-th half-hour interval, X_k, $k = 1, 2, \ldots, 48$, for each of the two ISPs. The first two principal components explain 56.7% of the total variance for ISP1, and 54.5% for ISP2. The loadings of the first two principal components for each ISP are represented in Figure 6. Note that all PC 1 loadings are positive and of similar magnitude, while PC 2 loadings are negative for the first period of the day and positive for the second part of the day.

Taking into account not only the loadings but also the correlation between each observed variable and each of the first two principal components it can be said that, for both data sets, PC 1 is an average of the hourly traffic utilization along the day and PC 2 is a measure of contrast between the "morning" and "afternoon" utilizations. Thus, high values of PC 1 are associated with users with high Internet utilization rates along the day and low values represent users with low Internet transfer rates. Likewise, high (positive) values of PC 2 can be interpreted as describing users with high rates of utilization only in the last period of the day ("afternoon"), while very small (negative) values of PC 2 represent users with high rates of utilization in the first period of the day ("morning") and low rates of utilization during the second period of the day ("afternoon").

The scores of ISP1 users in the first two principal components are displayed in Figure 7; a similar shape is observed for the (not shown) scores of ISP2 users in the first two PCs. The peculiar pattern exhibited in Figure 7 leads to the conclusion that the variability of PC 2 increases with the value of PC 1. This suggests that the variability of Internet utilization in half-hour intervals along the day increases with the values of PC 1, i.e., with the daily average Internet utilization. Taking into account this fact, and in order to smooth this variability, the following transformation of the data was performed, in each set of data:

$$Y_j = \ln(1 + X_j) \tag{4}$$

for $j = 1, \ldots, 48$.

Repeating the PCA to the transformed data, Y_j, we conclude that the first two PCs explain 50.6% of the total variance for ISP1 and 60.4% for ISP2. These PCs have the following interpretation for both ISPs: PC 1 is an average of Internet utilization along the day and PC 2 is a measure of contrast between the "morning" utilization ("morning" means from 2:30 am until 1 pm for ISP1 and from 0 am until 1:30 pm for ISP2) and the "afternoon" utilization. Moreover, the transformation used succeeded in reducing the effect observed in the original data of increase in the variability of PC 2 with the value of PC 1. Thus, the transformed data is going to be used in sections 4–5.

In the next section, we use cluster analysis to define groups of users with similar traffic utilization, based on the data transformed by (4).

4 Cluster analysis

The aim of cluster analysis is to partition a set of objects into groups or clusters in such a way that profiles of objects in the same group are similar, whereas the profiles of objects in different clusters are distinct. The concept of cluster is linked with the concept of proximity or distance between objects and groups of objects [7]. Generally speaking, cluster analysis methods are of two types: hierarchical and partitioning methods, with the former being the most common approach.

The hierarchical clustering techniques proceed by either a successive series of merges (agglomerative hierarchical methods) or by a series of successive divisions (divisive hierarchical methods). The agglomerative methods start with as many clusters as objects and end with only one cluster, containing all the objects. The divisive methods work in the opposite direction. The results of hierarchical methods may be displayed as a form of a diagram tree, called dendogram (vide figures 8–9), that illustrates the merges or divisions which have been made. These methods are based on a measure of proximity between two objects and a criterion, relying on the distance between groups of objects being used, to define which clusters should be joined in each step.

In the present work, and since the observations are continuous, the chosen distance between two objects (users), $(x_{i1}, \ldots, x_{ip})^t$ and $(x_{j1}, \ldots, x_{jp})^t$, is the Euclidean distance:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}.$$

The merging criteria used leads to what is called *Ward's method*, next explained. Let n_r be the number of objects in group r and x_{ikr} be the value of the i-th object of group r on the k-th variable. The within group sum of squares is given by

$$\text{SSW}_r = \sum_{i=1}^{n_r} \sum_{k=1}^{p} (x_{ikr} - \bar{x}_{\cdot kr})^2,$$

where $\bar{x}_{\cdot kr}$ is the mean for the n_r users of group r in the k-th variable. Let (r, s) denote the group containing the users from groups r and s. At each step, Ward's method merges the two clusters r and s for which

$$\text{SSW}_{(r,s)} - (\text{SSW}_r + \text{SSW}_s)$$

is minimum and, whence, seeks to minimize the increase in total within sum of squares.

In addition to Ward's method we will also use the (partitioning around) *medoids method*, for which the analyst has to decide in advance how many clusters, say K, he wants to consider. The method starts by choosing the K medoids, here denoted by $m_1, m_2, \ldots, m_K$. These are representative objects that are chosen such that the total (Euclidean) distance of all objects to their nearest medoid is minimal, i.e., the algorithm finds a subset $\{m_1, \ldots, m_K\} \subset \{1, \ldots, n\}$ which minimizes the function

$$\sum_{i=1}^{n} \min_{t=1,\ldots,K} d_{im_t}.$$

Each object is then assigned to the cluster corresponding to the nearest medoid. That is, object i is assigned to cluster C_i whose associated medoid, m_{C_i}, is nearest to object i, i.e., $d_{im_{C_i}} \leq d_{im_j}$, for all $j \in \{1, 2, \ldots, K\}$.

In the present study, the users are the objects and the variables are the half-hour interval transfer rates along the day. Thus, we want to group users with similar pattern of Internet utilization, characterized by their transfer rates in each half hour interval. As clustering methods we have used Ward's and the partitioning around medoids methods.

For both ISPs, the two considered clustering methods lead to the choice of 3 clusters. The dendograms in figures 8–9 illustrate the merging of clusters using Ward's method while Figure 10 describes the medoids for each ISP using 3 clusters. The considered clusters using Ward's method were obtained by cutting the trees of figures 8–9 by a distance that leads to 3 branches. This choice was confirmed by the medoids method, since we obtained with 3 medoids the highest values of the overall average silhouette width: 0.49 for ISP1 and 0.32 for ISP2. The silhouette width is a quality index that measures how strong the cluster structure is and helps the analyst choosing the number of clusters to be considered. In [10] it is recommended that this index should be higher than 0.25. In both cases, other numbers of clusters as well as other clustering methods were considered leading to less clear and coherent results.

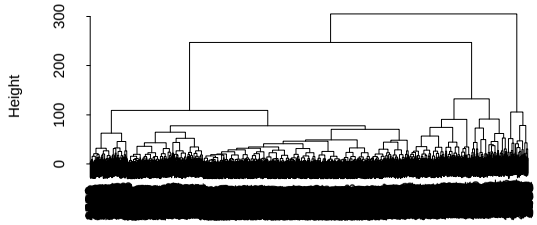

Figure 8: Dendogram, ISP1.

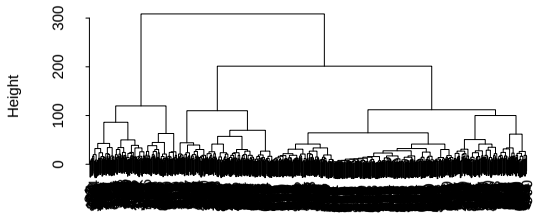

Figure 9: Dendogram, ISP2.

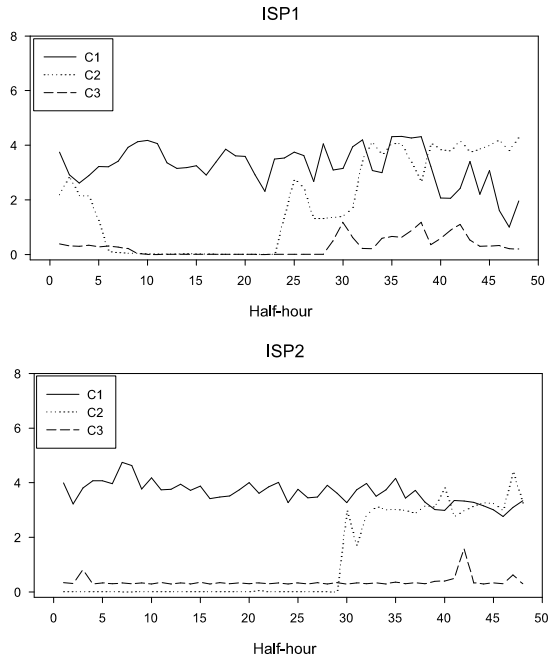

Figure 10: Medoids, ISP1 and ISP2.

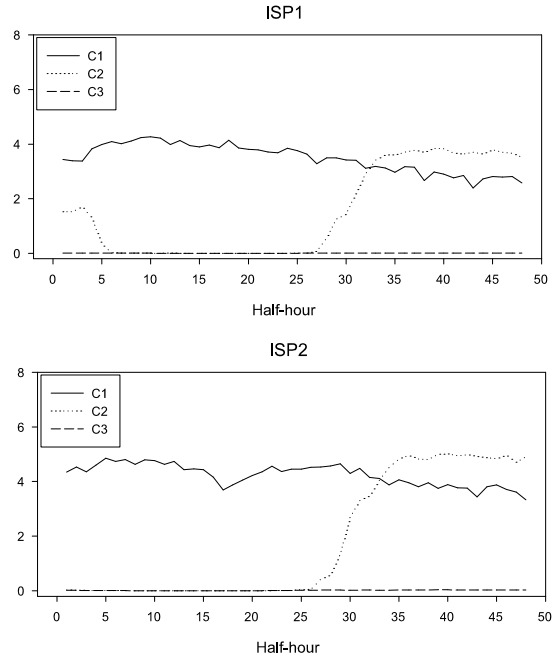

Figure 11: Half-hour medians for each cluster, ISP1 and ISP2.

The clusters formed through the medoids and Ward's methods lead to very high rates of in-between agreement of users partitioning, for both ISPs. In fact, 84.4% (84.0%) of the ISP1 (ISP2) users were assigned to the same cluster by both the medoids and Ward's methods. To interpret the 3 clusters obtained for each ISP, graphs of the median and average half-hour transfer rates along the day within each cluster were obtained. All graphs lead to similar conclusions, so only the graphs of the median, for each ISP, using the medoids method are presented; see Figure 11. In addition, the representation of the typical user in each cluster (i.e., the cluster's medoid) is given in Figure 10.

The clusters obtained can be interpreted in the same way for both ISPs and are described in Table 2. From Table 3 we can conclude that cluster C1 (characterized by high transfer rate in all periods) is the cluster with fewest users (with the exception of ISP2, medoids) and cluster C3 (characterized by low transfer rate in all periods) is the one with highest number of users. It is also worthwhile to mention that the discrepancy among cluster sizes is smaller for ISP2 than for ISP1. That is, proportionally, there are more ISP1 users characterized

Table 2: Interpretation of both ISP1 and ISP2 clusters.

Cluster	Interpretation
C1	high transfer rate in all periods
C2	low/high transfer rate in the morning/afternoon
C3	low transfer rate in all periods

Table 3: Cluster sizes for ISP1 and ISP2.

	ISP1		ISP2	
Cluster	Medoids	Ward	Medoids	Ward
C1	4.23%	4.87%	18.52%	20.00%
C2	7.75%	22.29%	14.17%	21.94%
C3	88.02%	72.84%	67.32%	58.06%

by low transfer rates and fewer users associated with high transfer rates all day long.

The structure of 3 clusters has to be validated. If the division of the population in clusters is in fact clear, the groups should also be well separated and it is expected that classification rules correctly assign most of the users. The cluster structure is going to be evaluated, in the next section, using discriminant analysis.

5 Validation using discriminant analysis

Discriminant analysis [9] is a multivariate technique concerned with the separation among different sets of objects and the classification of a new object into one of the previous defined groups. If, in fact, there is a cluster structure, then the separation among the clusters (groups) should be clear, and discriminant analysis be able to detected it. Moreover, if this separation is clear, it is anticipated that the estimated classification rules are able to allocate the majority of the users to the right cluster. If so, the cluster analysis is validated.

Generally speaking, linear discriminant analysis seeks for a linear function of the data, called discriminant function, that best separates g groups characterized by p random variables into linear combinations that achieve the best separation of the means relative to the variances.

R. A. Fisher suggested a sensible procedure to distinguish between groups [11]. The first discriminant function is the linear combination of the observed variables, $\mathbf{l}_1^t\mathbf{x} = l_{11}x_1 + \ldots + l_{1p}x_p$, that maximizes the ratio of the between-group sum of squares to the within-group sum of squares. Which means that the separation is made in such a way that within each group the objects are as similar as possible but, at the same time, the groups are as different as possible. A maximum of $s = \min(g-1, p)$ discriminant functions can be defined as linear functions of the observed variables, uncorrelated with the previous ones, which verify the same optimality criteria.

Let $\mathbf{x}_{ir}$ be the vector of observations on user i for group r (with n_r users) where $\mathbf{x}_{ir} = (x_{i1r}, \ldots, x_{ipr})^t$. The sample mean vector for group r is $\bar{\mathbf{x}}_{\cdot r} = (\bar{x}_{\cdot 1r}, \ldots, \bar{x}_{\cdot pr})^t$, where $\bar{x}_{\cdot jr} = \sum_{i=1}^{n_r} x_{ijr}/n_r$, $j = 1, \ldots, p$, $r = 1, \ldots, g$. The within-group sum of squares, W, is defined by

$$W = \sum_{r=1}^{g}\sum_{i=1}^{n_r} (\mathbf{x}_{ir} - \bar{\mathbf{x}}_{\cdot r})(\mathbf{x}_{ir} - \bar{\mathbf{x}}_{\cdot r})^t,$$

and the between-group sum of squares matrix is

$$B = \sum_{r=1}^{g}\sum_{i=1}^{n_r} (\bar{\mathbf{x}}_{\cdot r} - \bar{\mathbf{x}}_{\cdot\cdot})(\bar{\mathbf{x}}_{\cdot r} - \bar{\mathbf{x}}_{\cdot\cdot})^t = \sum_{r=1}^{g} n_r(\bar{\mathbf{x}}_{\cdot r} - \bar{\mathbf{x}}_{\cdot\cdot})(\bar{\mathbf{x}}_{\cdot r} - \bar{\mathbf{x}}_{\cdot\cdot})^t,$$

where the overall sample mean is $\bar{\mathbf{x}}_{\cdot\cdot} = \sum_{r=1}^{g}\sum_{i=1}^{n_r} \mathbf{x}_{ir}/n$, and $n = n_1 + \ldots + n_g$. It can be proved that $\mathbf{l}_j$ is the eigenvector associated with the j-th biggest eigenvalue of the matrix $W^{-1}B$, scaled such that $\mathbf{l}_j^t S_p \mathbf{l}_j = 1$, where $S_p = W/(n-g)$.

The Fisher's discriminant functions were derived to obtained a representation of the data that separates the population as much as possible. However, it can be used to produce a discrimination rule [12]. Let $(\mathbf{l}_1^t\bar{\mathbf{x}}_{\cdot r}, \ldots, \mathbf{l}_s^t\bar{\mathbf{x}}_{\cdot r})^t$ be the sample mean vector of the discriminant scores associated with group r. To classified the object $\mathbf{x}_0$ we have to evaluate the discriminant functions on this object $(\mathbf{l}_1^t\mathbf{x}_0, \ldots, \mathbf{l}_s^t\mathbf{x}_0)^t$. The object should be allocated to the group for which its square Euclidean distance to the group sample mean vector $(\mathbf{l}_1^t\bar{\mathbf{x}}_{\cdot r}, \ldots, \mathbf{l}_s^t\bar{\mathbf{x}}_{\cdot r})^t$ is the smallest, i.e. to the group associated to the nearest centroid.

If the groups have very different sizes, prior probabilities associated with each group can be used to obtain a better classification rule. If the prior is such that $p_1 = \ldots = p_g = 1/g$, this rule is equivalent to the

Table 4: Error rates, for ISP1 and ISP2.

Error rate	ISP1		ISP2	
	Medoids	Ward	Medoids	Ward
Apparent	1.31%	9.47%	3.09%	6.74%
Jackknife	1.81%	9.99%	5.71%	9.49%
Cross-valid.	1.78%	10.05%	6.59%	9.57%
Double cross	2.18%	10.02%	7.77%	10.86%

classification rule obtained when the observations have normal distribution and the groups have different mean vectors but equal covariance matrices.

Discriminant analysis can be used to validate the obtained clusters. In fact, if we consider that each user belongs to a group, determined by the cluster he was classified in, we can obtain a discriminant function that best separates these groups (clusters). If the groups are well separated, then it is expected that the estimated classification rules should produce lower rates of misclassification, validating the cluster structure. Once again, several procedures to estimate the discriminant functions were considered, however the one that lead to the lowest error rates was the Fisher's discriminant function, modified in order to consider prior probabilities [12]. A discriminant rule can be evaluated by reclassifying the users, and calculating the number of individuals badly classified in each group. Classifying the individuals used to estimate the discriminant function determines what is called the apparent error rate. The apparent error rate tends to underestimate the true error rate, thus misleading the conclusions. Accordingly, 3 other error rates were considered:

- The first error rate is obtained by a procedure called leave-one-out or Jackknife. Each user is left out, a classification rule using all the other $(n-1)$ users is estimated and the user left out is classified.
- The second error rate is obtained by cross-validation. To be more precise, each data set is divided in two. The first data set, containing 80% of the data (called the training set) is used to estimate the classification rule and the rest of the data is classified, leading to an error rate.
- The third error rate is obtained by double cross-validation. Once again, the data set is divided in two groups, each of them containing 50% of the data. Take the first group to estimate the classification rule, use it to classify the second data set and compute the associated error rate. Then use the second group to estimate the classification rule and classify the first, calculate a second error rate. The final error rate is the mean of the two previous ones.

Given that the obtained clusters have very different sizes (vide Table 3) the classification rules, as well as the error rates were obtained taking into account the proportion of users belonging to each group, i.e. $p_r = n_r/n$. If all these error rates are low it is expected that the groups are well separate, thus the cluster structure is well defined and therefore validated. The error rates are presented in Table 4. For both ISPs, the clusters obtained using the partition around medoids lead to lower error rates. But in general, all the rates are low, validating the cluster structure proposed.

In the evaluation of the classification rules, the actual group that each user belongs to is known. Therefore, a matrix crossing the number of users from each actual group with the number of predicted users in each group can be very helpful and is called confusion matrix.

Having analysed all the obtained confusion matrices we can conclude that, with the exception of ISP1 with clusters formed by medoids, a higher proportions of users from cluster C2 have been misclassified, and clusters C2 and C3 are the ones that all discriminant functions have more difficulty to distinguish. As an example of this statement consider the confusion matrix presented in Table 5. This matrix indicates the number of ISP2 users correctly classified for each of the clusters obtained by the partitioning around medoids method, resulting from the classification of 20% of the users, when the rest of the data were used to estimate the discriminant function (cross-validation). The highest rate of misclassification (14.95%) occurs with C2, with the majority of C2 misclassified users being assigned to C3 (15 in 16 badly classified). This result can be justified by the comparison of the average transfer rates of C2 and C3 users in each period. We can conclude that, at a 5% significant level, the average transfer rates of C2 and C3 users are equal on the period: between 4:30 am and 11 am, for ISP1, and between 0 am and 0:30 pm, for ISP2. So since there is a period in the morning where,

Table 5: Confusion matrix, for ISP2. Clusters were obtained by partitioning around medoids (cross-validation).

Actual cluster	Predicted cluster C1	C2	C3	Percentage of misclassif.
C1	127	11	5	11.19%
C2	1	91	15	14.95%
C3	5	11	434	3.56%

Table 6: Averages for the aggregate of applications, clustered users, ISP1 (left) and ISP2 (right).

	Downloaded MBytes	Number of flows	Flow duration (hours)	Transfer rate (Kbits/sec)
C1	695 / 1450	1.23 / 1.28	19.0 / 19.2	71.9 / 145.9
C2	312 / 764	1.82 / 1.64	8.0 / 8.8	56.8 / 123.1
C3	36.5 / 167	1.83 / 1.90	4.0 / 5.7	18.7 / 54.5
All	85.6 / 489	1.80 / 1.75	4.9 / 8.6	23.9 / 81.1

in average, the two clusters are equal, then the classification rules have difficulty in assigning users to the right clusters.

With ISP1, when the clusters were formed using medoids, the higher rate of misclassification is associated with C1 (C2 is the second worst), and the majority of badly classified users are assigned to C2 (11 in the total of 16 badly classified). In fact, in a small period of the afternoon (between 4 pm and 6 pm) the C1 and C2 are also not statistically different in average.

6 Evaluation of clusters

The goal here is to assess if the clusters obtained in previous sections are also meaningful from the point of view of other traffic characteristics that were used in the cluster analysis. In particular, we are interested in the way the relative utilization of applications, the flow durations and the flow transfer rates map into the clusters C1, C2 and C3. We restrict our attention to the clusters obtained using the medoids method, since this method leaded to the lowest misclassification rates (*vide* Table 4). We consider again the statistics of Section 2, that are now computed over each cluster: relative (aggregate) utilization (of application groups), average flow duration and average flow transfer rate. In addition, we consider the average relative user utilization (i.e., the average of the relative utilizations per user) for groups of applications. The relative utilization of a group of applications by an user is the percentage of downloaded bytes associated to all applications of the group made by the user with respect to the total number of downloaded bytes by the user.

As before, we first consider the flow statistics of the aggregate of applications. The results are shown in Table 6. To ease the analysis we repeat the values corresponding to the aggregate of users (line ”All” in the Table).

For both ISPs, the downloaded MBytes per user, average flow duration and average transfer rate decrease from C1 to C3; the number of flows per user increases from C1 to C3. Thus, C1 users are the most intensive ones, using the Internet during longer periods at higher transfer rates. We note, in particular, the very high flow durations, on the order of 19 hours in both ISPs, for cluster C1. On the opposite end, C3 users have significantly lower transfer rates and flow durations. The flow durations in ISP1 and ISP2 have now relatively close values in each cluster.

These results also highlight the way cluster analysis can enhance traffic characterization. Clearly, the values for the aggregate of users (line “All” in the Table) are close to the C3 values since the majority of users belong to C3. Whence, the values for the aggregate of users hide the fact that a small number of users consumes most of the resources.

We consider now the most typical applications within each cluster. In figures 12–13 we depict the relative (aggregate) utilization and in in figures 14–15 the average relative user utilization, for both ISP1 and ISP2.

The user utilization statistics (figures 14– 15) show that the main application group of C1 and C2 users is file sharing followed by HTTP. Conversely, C3 users prefer HTTP against file sharing. There is no marked dif-

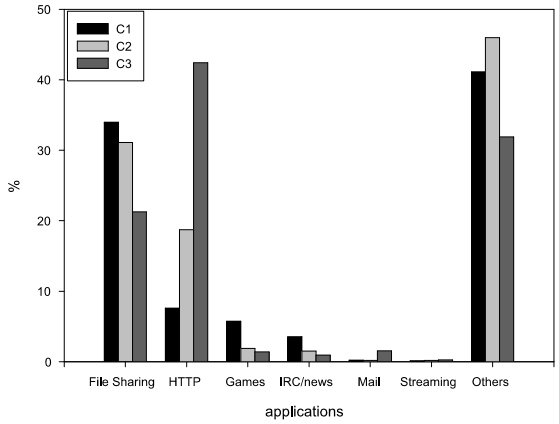

Figure 12: Relative (aggregate) utilization, clustered users, ISP1.

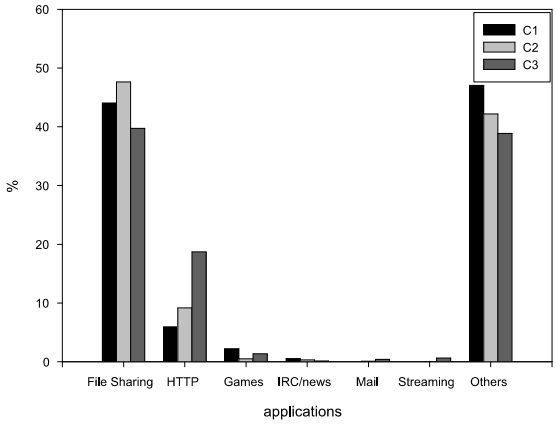

Figure 13: Relative (aggregate) utilization, clustered users, ISP2.

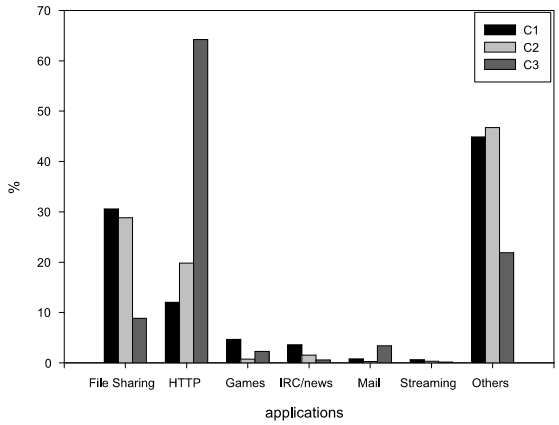

Figure 14: Average relative user utilization, clustered users, ISP1.

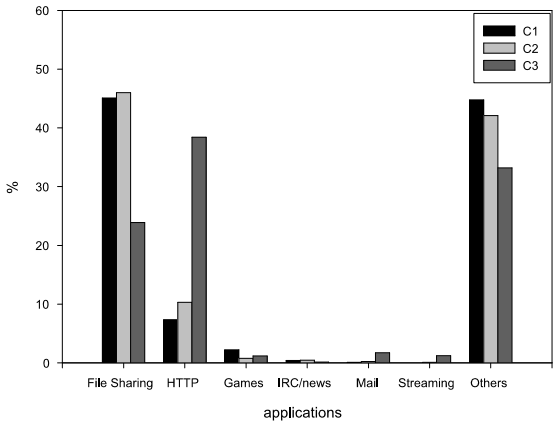

Figure 15: Average relative user utilization, clustered users, ISP2.

ferences between C1 and C2 in what concerns these two application groups. Looking at secondary applications it can be seen that C1 users use more on-line games.

The statistics of aggregate utilization (figures 12–13) confirm these later conclusions, except that the most used application group in cluster C3 of ISP2 is file sharing in spite of the fact that most users use HTTP more often. This indicates that in cluster C3 of ISP2 a few users have a relatively high utilization of file sharing applications. This may be due to users that may have been classified in C3, that have a high utilizations in the “morning” period or in short time periods of the “afternoon”. It is also worth noting that C3 users have a relatively higher utilization of mail applications.

Figures 16–17 show the transfer rate statistics. C1 and C2 users transfer file sharing applications at a higher rate than C3 users. It is also noticeable that the transfer rate of HTTP is approximately the same irrespective of cluster.

Flow durations, figures 18–19, can be used to distinguish users from clusters C1 and C2: the flows durations of file sharing applications are higher for C1 users. Note also that this characteristic is also observed for almost all applications.

Thus, the main characteristics that can be used to discriminate the clusters are the following: C1 and C2 use more file sharing and on-line games and C3 uses more HTTP and mail; C1 and C2 use file sharing applications at a higher transfer rate; C1 and C2 users are mainly distinguished by the flow durations, with C1 users having higher flow durations.

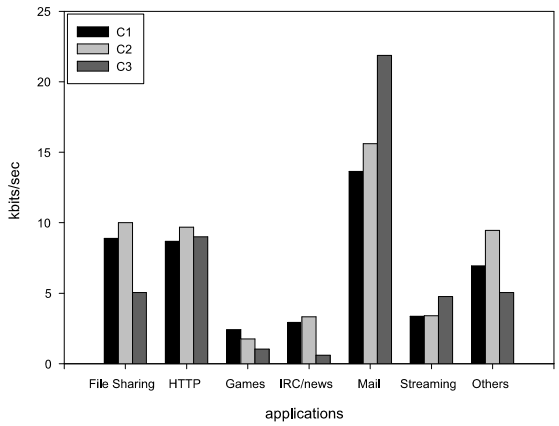

Figure 16: Average transfer rate, clustered users, ISP1.

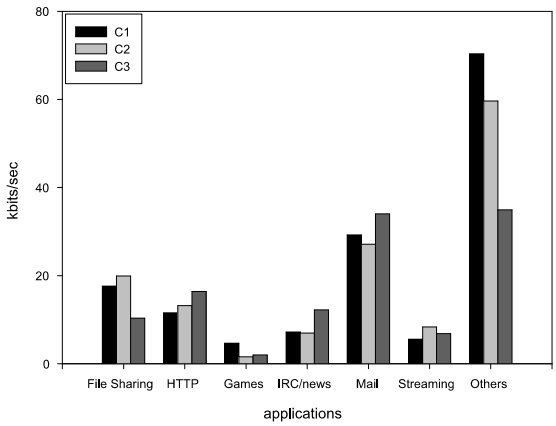

Figure 17: Average transfer rate, clustered users, ISP2.

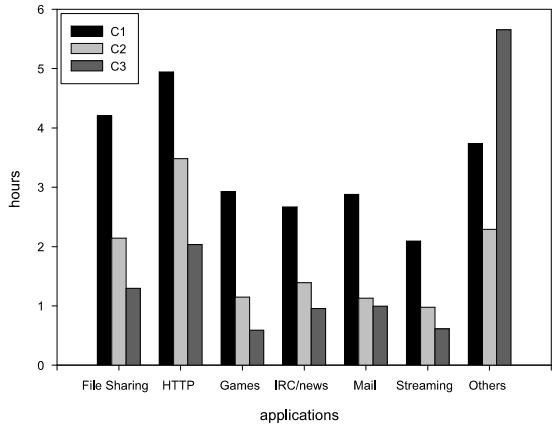

Figure 18: Average flow duration, clustered users, ISP1.

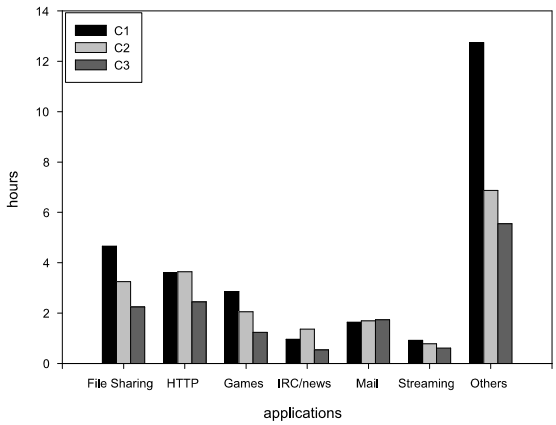

Figure 19: Average flow duration, clustered users, ISP2.

7 Applications of cluster analysis

The cluster analysis of Internet users based on the hourly traffic utilization, besides allowing a deeper understanding of the traffic characteristics, can be applied in tasks that require assessment of the hourly behavior of traffic, such as resource management and tariffing. We give some examples in the remainder of this section.

Resource management - Cluster analysis of Internet users can be employed by ISPs and network operators in resource management tasks that are performed on an hourly basis. For example, when routing is allowed to change several times per day, a feature sometimes called multi-hour routing and used in telephony, ATM and MPLS networks, cluster analysis can be used to delimit the time periods when routing is kept fixed and to identify the groups of users with similar routing. Cluster analysis also can be employed by ISPs in differentiating service given to users based on their hourly utilization, e.g., through traffic shaping.

Tariffing - Hourly based tariffs can be used, for example, to promote Internet access in the least busy hours. Cluster analysis of Internet users allows ISPs to assess whether or not hourly based tariffs are advantageous. This may be the case when a significant number of users follows a non-flat usage profile, as in our C2 cluster. In this situation, cluster analysis can help deciding on the number and type of hourly based tariffs, and to delimit the time periods of each tariff. Moreover, in case an ISP offers hourly based tariffs, cluster and discriminant analysis can be used to advise users on the type of tariff they should select, based on their previous usage. Cluster analysis can also be used when ISPs need to lease circuits from access network operators for the transport of traffic aggregates, a common situation in ADSL networks. On one hand, access network operators have to do similar studies as ISPs in previous example, to assess if it is beneficial to offer circuits with hourly

based tariffs and to define the characteristics of these tariffs. On the other hand, ISPs can use cluster and discriminant analysis to decide on what tariffs to choose.

8 Conclusions

In this paper, we have partitioned Internet users based on their hourly traffic utilization using cluster analysis. In particular, we have used two clustering methods, the partitioning around medoids and Ward's methods. The analysis resorted to two traffic traces measured at distinct Portuguese ISPs offering distinct traffic contracts, one using a CATV access network and the other an ADSL one. In spite of the ISP's differences and the use of different clustering methods, based on hourly traffic utilization, 3 clusters, with similar interpretations, were obtained for both ISPs. The typical user of each cluster (C1, C2 and C3) is characterized as follows: a high utilization rate in all time periods, for C1; a low utilization rate in the first half of the day and a high utilization rate in the second half, for C2; and, finally, a low utilization rate in all time periods, for C3. This cluster structure was validated by discriminant analysis as the computed misclassification rates were low. These results show that the users of both ISPs have a typical residential profile.

The 3 clusters were also evaluated in terms of several characteristics not used in the cluster analysis, such as relative utilization of applications, average flow duration and average transfer rate. The results showed that the clusters have distinctive characteristics: users from C1 and C2 mainly use file sharing applications and users from C3 mainly use HTTP; users from C1 have higher flow durations than C2 users. We have highlighted the importance of hourly Internet utilization profiles in the context of traffic engineering and tariffing applications.

References

[1] K. Claffy, H. Braun, and G. Polyzos, "A parameterizable methodology for Internet traffic flow profiling," *IEEE Journal of Selected Areas in Communications*, Mar. 1995.

[2] J. Quittek, T. Zseby, B. Claise, and S. Zander, "Requirements for IP flow information export," Internet-draft, IETF, http://www.ietf.org/internet-drafts/draft-ietf-ipfix-reqs-09.txt, Feb. 2003.

[3] J. Färber, S. Bodamer, and J.Charzinski, "Statistical evaluation and modelling of Internet dial-up traffic," in *Proc. SPIE Photonics East Conf. "Performance and Control of Network Systems III"*, Boston, MA, USA, Sep. 1999.

[4] P. Barford, A. Bestavros, A. Bradley, and M. E. Crovella, "Changes in web client access patterns: Characteristics and caching implications," *World Wide Web, Special Issue on Characterization and Performance Evaluation*, vol. 2, pp. 15–28, 1999.

[5] N. Vicari, S. Kohler, and J. Charzinski, "The dependence of Internet user traffic characteristics on access speed," in *Proceedings of the 25th Local Computer Networks (LCN) Conference*, Tampa, USA, Nov. 2000.

[6] http://www.caida.org/, "Caida," .

[7] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, 1990.

[8] I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, 1986.

[9] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, 1992.

[10] A. Struyf, H. Mia, and P. J. Rousseeuw, "Integrating robust clustering techniques is S-PLUS," *Computational Statistics and Data Analysis*, vol. 26, pp. 17–37, 1997.

[11] J. D. Jobson, *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods*, Springer-Verlag, 1992.

[12] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, Inc, 1982.