Large-scale Internet user behavior analysis of a nationwide K-12 education network based on DNS queries

Alexis Arriola$^{1[0000-0001-6937-6645]}$, Marcos Pastorini$^{1[0000-0003-0812-2419]}$, Germán Capdehourat$^{2[0000-0002-3975-2168]}$, Eduardo Grampín$^{1[0000-0001-6046-0023]}$, and Alberto Castro$^{1*[0000-0002-9174-398X]}$

[1] School of Engineering, Universidad de la República, Montevideo, Uruguay.
[2] Plan Ceibal, Montevideo, Uruguay
*acastro@fing.edu.uy

Abstract. To the best of our knowledge, this paper presents the first Internet Domain Name System (DNS) queries data study from a national K-12 Education Service Provider. This provider, called *Plan Ceibal*, supports a one-to-one computing program in Uruguay. Additionally, it has deployed an Information and Communications Technology (ICT) infrastructure in all of Uruguay's public schools and high-schools, in addition to many public spaces. The main development is wireless connectivity, which allows all the students (whose ages range between 6 and 18 years old) to connect to different resources, including Internet access. In this article, we use 9,125,888,714 DNS-query records, collected from March to May 2019, to study Plan Ceibal user's Internet behavior applying unsupervised machine learning techniques. Firstly, we conducted a statistical analysis aiming at depicting the distribution of the data. Then, to understand users' Internet behavior, we performed principal component analysis (PCA) and clustering methods. The results show that Internet use behavior is influenced by age-group and time of the day. However, it is independent of the geographical location of the users. Internet use behavior analysis is of paramount importance for evidence-based decision making by any education network provider, not only from the network-operator perspective but also for providing crucial information for learning analytics purposes.

Keywords: Machine Learning, Data Mining, Big Data.

1 Introduction

1.1 Motivation and Objective

The importance of Internet access to education has increased significantly in recent years. Today, in many places, it has become a relevant resource in formal school education, for various activities based on learning platforms and educational management systems. In order to support this new Internet-based educational paradigm, it is necessary to deploy and maintain the corresponding network infrastructure. This responsibility falls on the Education Service Providers (ESPs). These organizations help the education system to implement comprehensive reforms towards digitization. In particular, this includes providing teachers and students with reliable and high-quality access networks to avoid performance degradation, which impacts the quality of education.

From the network infrastructure design and deployment point of view, this is also quite a new challenge. The educational use case poses new challenges, both from the content and applications, as well as from the users, who are teachers and students (children and teenagers). Therefore, it is quite important to understand this novel scenario, with quite unique characteristics that are not common in other business environments. In this work, we try to shed some light on this last aspect, in particular, by analyzing the behavior of users in educational networks. For this purpose, we studied a large volume of Domain Name System (DNS) data collected during three months from a major nationwide ESP.

1.2 Background

During the 2005 World Economic Forum, Nicholas Negroponte, from MIT, proposed a novel project based on low-cost laptops to reduce the digital divide in less developed countries, known as One Laptop per Child (OLPC) [1]. Although in the following years, the one-to-one educational model gained importance within different regions of the world, the only country to implement it at national level was Uruguay. A Presidential Decree of April 18th, 2007 [2] formally kick-started Plan Ceibal, “a plan for inclusion and equal opportunities to support Uruguayan educational policies with technology,” as stated in its website [3].

The initial mission of Plan Ceibal was to promote digital inclusion. Its main goals were to deliver laptops to all the students and teachers and also to provide Internet access at every public educational center. These two overall objectives were achieved in the first three years of the program. Now, more than 10 years later, Plan Ceibal has greatly diversified its services to the education system. It has provided several educational resources and platforms, such as a content management system (CMS), an intelligent tutoring system (ITS) for math learning and a digital library. The increasing dependence on technology by the education system has also led to a greater technical support demand from Plan Ceibal.

Therefore, still today, one of the most relevant responsibilities of Plan Ceibal is to provide connectivity at all educational centers throughout the country. The typical networking solution deployed consists of a router at each school with fiber Internet access from the local ISP and several access points providing Wi-Fi connectivity. With more than 1,500 educational centers (including schools, high-schools, and Domain Name System (DNS)) and more than 8,000 Wi-Fi access points, Plan Ceibal is one of the nation’s largest wireless Internet provider, reaching a number of devices comparable to the number of subscribers of local mobile network operators.

Two years ago, Plan Ceibal has incorporated a novel cloud-based DNS solution, Cisco Umbrella [4], which also serves for security purposes and content filtering.

Among the advantages of this new solution, it is possible to access in a very simple way all the records of the DNS queries processed, as logging can be configured so that logs are stored to an Amazon S3 bucket. By having the logs uploaded to an S3 bucket (where data is stored for a maximum of 30 days), then they can be automatically downloaded to keep them in perpetuity in backup storage outside of Umbrella's data warehouse storage system. In this way, centralized access to the logs of the DNS queries of the entire network is quite easy. They can then be analyzed and integrated with other data in a specialized data analysis platform.

1.3 Related Work

While there is a vast literature of previous work on DNS data analysis, the majority of them focus on security applications [5]. In all cases, the problems addressed correspond to malicious activities detection based on DNS, such as malicious URLs, botnets, phishing, and web-spam [6].

Concerning behavior analysis with DNS data, several previous works are worth mentioning. Plonka et al. propose a context-aware clustering method [7], where the type of the DNS queries is pre-classified as canonical (i.e., RFC-intended behaviors), overloaded (e.g., black-list services), or unwanted (i.e., queries that will never succeed). In [8], Gao et al. present the analysis of a large dataset of DNS performance measurements. The data from 600 different recursive DNS resolvers, which were globally distributed, is used to compare them, and find out differences and similarities between them (e.g., query success rate and lookup failures causes). A data mining methodology based on different clustering techniques is developed by Ruana et al. in [9], aiming to learn the behavior patterns of DNS queries and detect anomalies. The impact of DNS cache and the wide use of NATs is tackled by Su et al. in [10], for the analysis of a DNS dataset from a major ISP in China. The work by Schomp et al. [11] addresses the study of DNS behavior of individual clients, looking forward to developing an analytical model. In the same way, Li et al. [12] seek to understand and profile what people are doing from DNS traffic, also focused on network behavior analysis. Finally, Jia et al. [13] developed an accessing behavior model for each user, by analyzing network DNS log in a campus network.

Our work is not security-oriented but instead seeks to analyze the observed behavior and recognize relevant patterns. The goal is to understand the typical characteristics and main trends of the DNS queries, in this particular educational context. To the best of our knowledge, this work presents the first study of DNS data in a K-12 education scenario, where most of the users are between 6 and 18 years old.

2 Dataset description

2.1 Data collection

The data used in this study were collected from Plan Ceibal's network infrastructure. This network, with 8,587 Wi-Fi access points (APs) located in 1,878 educational buildings (covering more than 95% of the K-12 students in Uruguay), it is one of the largest communication networks in the country. Of those 1,878 educational buildings, 70% are schools and 30% are high-schools (Fig. 1).

In particular, each record (i.e., data point) in our dataset corresponds to a DNS-query request. Plan Ceibal, as part of its network infrastructure, has deployed the Cisco Umbrella system [4] as its DNS solution. All the DNS-query requests performed by users connected to Plan Ceibal's network are recorded and categorized by the Cisco Umbrella system. There are two types of records: DNS logs and proxy logs. The former corresponds to the standard DNS queries, and the latter corresponds to dubious queries that are sent to a proxy for further inspection. For this work, only the DNS logs were considered. Each DNS-log record has the following fields [14]: *i*) *Timestamp*: when the request was made in UTC; *ii*) *InternalIp*: the IP address of the device (e.g., laptop, cellphone) that made the query; *iii*) *ExternalIp*: the IP address of the router that receives the query (i.e., Wi-Fi AP at the educational building); *iv*) *Action*: whether the DNS-query request was allowed or blocked; *v*) *Query Type*: the type of DNS request made; *vi*) *Response Code*: the DNS return code for the query; *vii*) *Domain*: the domain (e.g., youtube.com) that was requested; and *viii*) *Categories*: the system assigns one or more categories (e.g., Video Sharing) to the query, depending on the content of the destination. As an example of a DNS-log record, a typical one looks as follows: "2019-03-10 17:48:41", "10.10.1.100", "24.123.132.133", "Allowed", "1(A)", "NOERROR", "instagram.com", "Photo Sharing, Social Networking".

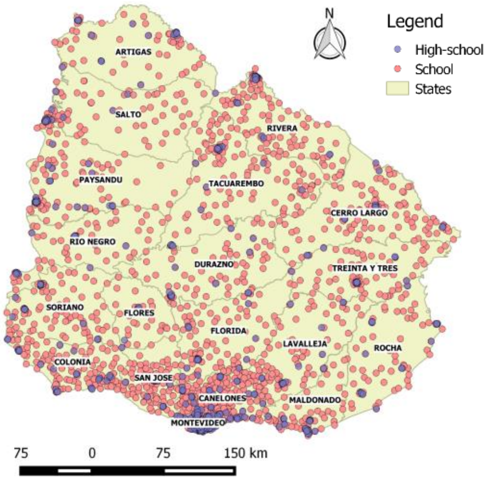

Fig. 1. Educational buildings in Uruguay.

Given that each Wi-Fi AP has a unique IP within the Plan Ceibal's network and that we know where each Wi-Fi AP is located, we used the DNS-log field *ExternalIp* to join a DNS query with an educational building. Particularly, from each education building we have information about the state, area (rural/urban), and the type of educational center (school, high-school, or others).

In this study, we analyzed 9,125,888,714 DNS-query records collected from March to May 2019. It is worth mentioning that, being Uruguay a Southern Hemisphere country, the academic year aligns with the calendar year, lasting from March to December. It is important to highlight that this study was approved by the ethical and data privacy committee of Plan Ceibal. All the datasets are de-identified and handled according to the Uruguayan privacy protection legislation.

2.2 Data statistics

As shown in Fig. 2, approximately 40% and 53% of the DNS queries correspond to schools and high-schools respectively; while less than 8% corresponds to other educational centers (e.g., technical schools). For this study, we only considered the records from these two large groups.

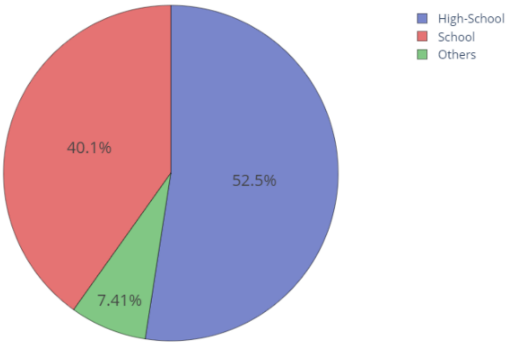

Fig. 2. Percentage of total DNS queries requested group by educational building.

As mentioned above, the Cisco Umbrella system assigns to each DNS query one or more categories [15] that describes the content of the requested domain. Fig. 3 shows the distribution in percentage of the 14 most popular categories (in terms of the number of DNS queries that they were assigned to):

- *Search Engines*: Sites that offer results based on keywords.
- *Infrastructure*: Infrastructure for delivering content and dynamically generated content, websites that cannot be classified more accurately because they are safe or otherwise difficult to classify.
- *Non-Category*: Sites to which the system could not assign a category.
- *Social Networking*: Sites that promote interaction and networking among people.
- *Business* Services: Sites for corporations and businesses of all sizes, especially corporate websites.
- *Chat*: Sites where you can chat in real-time with groups of people. It includes video chat sites.
- *Video-Sharing*: Sites to share video content.
- *Software/Technology*: Sites on computing, hardware, and technology, including news, information, code and provider information.
- *Photo Sharing*: Sites to share photos, images, galleries, and albums.
- *Games*: Sites that offer gameplay and game information.

- *SaaS and B2B*: Web portals for online commercial services.
- *Blogs*: Personal or group journal sites, diaries sites, or publications.
- *Educational Institutions*: Sites for schools that cover all levels and age ranges (including Plan Ceibal educational platform).
- *Non-Profits*: Sites for non-profit or charity organizations and services.

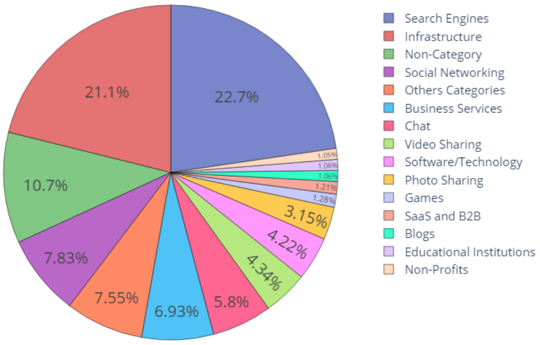

Fig. 3. Percentage of total DNS queries requested group by category

In Fig. 4, a representation of the total amount of DNS queries grouped by state is reported. It is possible to see that Montevideo alone (capital of the country) contains almost 26% of the records and Canelones nearly 17%, which was expected since they are account for the 55% of the total Uruguayan population.

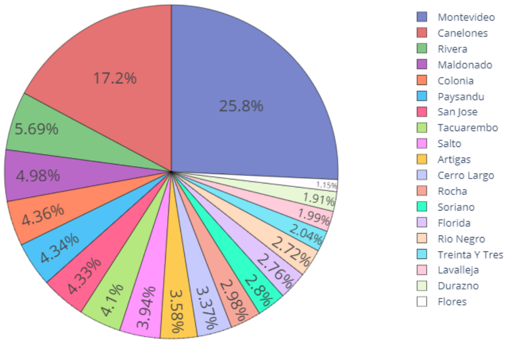

Fig. 4. Representation of the total amount of queries grouped per Uruguayan state.

Fig. 5 shows the histogram of the daily number of queries. Focusing on the number of requests per day in the period of the data availability (March-May 2019), we can see that the days with the most significant activity precisely correspond to the days of student activity (Monday-Friday). On March 6th, it is possible to see a growth in the number of queries in the network, corresponding to the beginning of the academic year. On the other hand, during April, there is a week of very low activity, which corresponds to Easter week holidays, when there are no classes. Since the main objective of this work is the study of the students' Internet behavior, weekends and holidays will not be considered in the data analysis. In addition, since the first three weeks of classes (March 6th – March 24th) show a different pattern from the rest of the weeks, they will not considered.

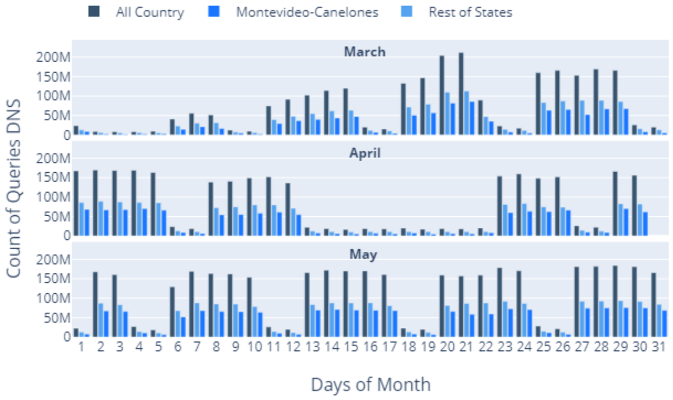

Fig. 5. Histogram of the daily number of queries for all the country, for Montevideo-Canelones, and for the rest of the states.

3 Methodology

Conventional multivariate data-analysis techniques, principal component analysis (PCA) and cluster analysis (CA), were used in this study to understand users' Internet behavior. These techniques are unsupervised methods, meaning that no information about other response variables, or cluster belonging, are used to obtain the outcomes. This makes these methods appropriate for exploratory data analysis, where the goal is hypothesis generation rather than hypothesis verification [16].

3.1 Principal Component Analysis (PCA)

PCA is a method that reduces and transforms measured data into new uncorrelated variables known as principal components (PCs) [17]. The raw measured data are considered as independent variables. There are as many PCs as the number of variables and every single PC is a linear combination of the original variables. The PCs make the basis of the respective vector space and they are organized by decreasing variance. Consequently, PC1 describes the largest data variance, PC2 the next largest data variance, and so on. Most of the variance is accounted for in the first few PCs [18]. Each object is identified by a score and each variable by a loading value or weighting. In its graphical representation (PCA biplot), vectors representing parameters that form an acute angle are considered as correlated parameters, while those that are perpendicular are considered as uncorrelated.

3.2 Clustering

In this study, the CA was not used as a separate methodological approach, but rather as a complementary method for validating and supporting previous PCA results, as well as for providing a better insight into the differences in users' Internet behavior all over the country.

The purpose of the clustering algorithm is to partition a complex dataset into homogeneous clusters, such that the between-group similarities are smaller than the within-group similarities [19]. These clusters can reveal trends and/or patterns related to the phenomenon under study. The similarity between observations is measured by a distance function. Firstly, the similarity is calculated between the data points. Once the data points begin to be clustered, the similarity is computed between the groups as well. Several metrics could be used to calculate it, and the choice of the similarity measure could influence the results. *A priori* and arbitrary decision of the number of groups (*k*) is also required in particular for the k-means CA.

In this study, the silhouette method was used to calculate the best *k* [20]. This analysis measures how close each point in a cluster is to the points in its neighboring groups:

$$s(i) = \frac{b(i)-a(i)}{\max(b(i),a(i))} \tag{1}$$

where $a(i)$ is the average distance of the point (i) to all the other points in the cluster it is assigned (A), $b(i)$ is the average distance of the point (i) to other points to its closest neighboring cluster (B). Silhouette values lie in the range of [-1, 1]; the higher the value, the better is the cluster configuration. In particular, the value 1 is ideal since it indicates that the sample is far away from its neighboring cluster and very close to the group it is assigned to. Similarly, the value -1 is the least preferred since it indicates that the point is closer to its neighboring cluster than to the cluster it is assigned to.

Hierarchical cluster analysis (HCA) was adopted in this study to run a heat map analysis. Unlike k-means, HCA starts with each of the *n* data points being their own cluster. In the next step, the two most similar data points are joined to form one cluster giving in all *n*-1 clusters. Afterward, the two most similar groups are joined to form one cluster, giving in all *n*-2 clusters. The process is repeated until every data point is in the same cluster that occurs at step *n*-1. The result is a hierarchical classification tree called dendrogram [21].

Heat map analysis is a false-color image with two dendrograms for two different objects and can divide these two objects into several clusters [22]. The different influence features in these two objects were reordered according to their similarity based on HCA.

4 Results

4.1 Computing Infrastructure

To perform the data analysis, we utilized the Hortonworks Data Platform (HDP) [23] deployed in a single cluster configuration with 40 CPUs Intel Xeon and 256 TB of RAM. HDP is an open-source framework specialized in the storage and distributed processing of large volumes of data (BigData) whose core is based on Apache Hadoop. Apache Hadoop [24] allows the storage and processing of large data sets.

To process our dataset and build the input matrices for the PCA and CA, we implemented specific SQL queries in Apache Spark [25] to be run on Apache Hive [26]. Apache Hive is based on Apache MapReduce. Specifically, it is a data warehouse that allows ad-hoc queries through a SQL-like interface for large data sets stored in an Apache Hadoop Data File System (HDFS). Apache Spark is a programming framework for distributed data processing.

PCA, silhouette method, and CA were programmed in Phyton 3.7 using the scikit-learn library [27]. The heatmap analysis was also coded in Python 3.7 using the seaborn library [28].

4.2 Data Analysis

Two different analyses with two different aims were performed.

The purpose of the first study was to identify the different age-group behavior of Ceibal's users. For this reason, we built a matrix characterized by 14 columns (categories described in section 2.2) and 38 rows (the type of educational centers, school or high-school, per state). In particular, for the two rows per state considered, the sum of queries by schools and high-schools was calculated respectively. In addition, since states have different numbers of schools and high-schools, the number of queries per state was averaged by the number of educational centers that are located in that particular state. Since the importance of columns (features) is independent from its own variance, we first centered each feature by subtracting its observed values the column's mean, and then we standardized it by dividing each value by the column's standard deviation.

This data matrix (38×14) was the input of the PCA, performed to decrease the dimensionality and have a better visualization of the observations, and k-means clustering, tackled to identify different group behaviors. To select the best number of clusters, the silhouette method was used. The average silhouette score of all the considered values in the dataset was calculated and represented by boxenplot (Fig. 6).

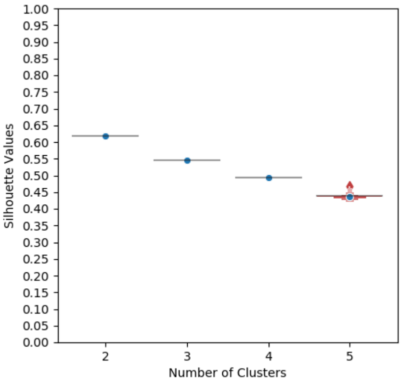

Fig. 6. Boxenplots of silhouette values (or scores).

The boxenplots in Fig. 6 represent an almost null dispersion of the silhouette values. Therefore, the number of clusters corresponding to the highest silhouette value was considered (*k*=2).

The result of PCA and k-means on this first data matrix is reported in Fig. 7. The first two principal components (PCs) were selected since they represented more than 96% of the variance (respectively 86.5% and 9.7%).

Each line of the input matrix (the type of educational center per state) is represented by a data point in the plot, and their location indicates their score for the PCs. Two net clusters were identified: one for school and the other for high-school-data points. This representation proves the different pattern between these two age groups. In other words, kids and teenagers clearly have a different Internet-use behavior. Furthermore, it is interesting to see that the high-school cluster shows a more significant variation in the PC1 compared to the school cluster.

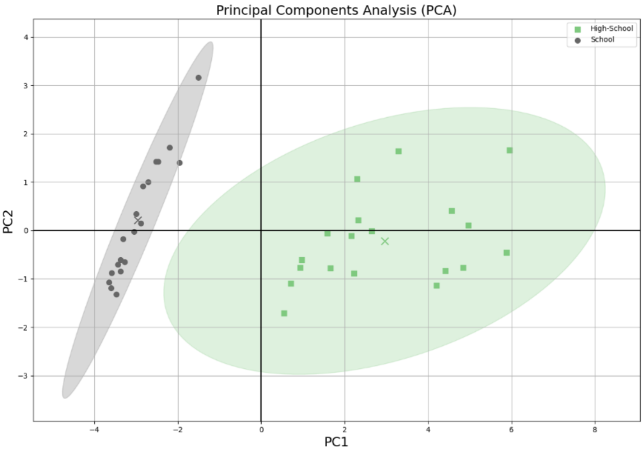

Fig. 7. PCA scores plot with school and high-school data points, and k-means clustering.

The objective of the second analysis was the detection of users' Internet behavior during the day hours. Considering the results of the previous study, it was decided to separate the dataset and examine the school and high-school observations independently. For this reason, two data matrices were built with the 24 hours of the day as rows and the categories previously mentioned as columns. For each hour, the sum of the queries of each category was considered. Also in this case, the matrices-columns' values were centered and standardized.

The school-matrix (24×14) and the high-school matrix (24×14) were the input of the PCA and k-means algorithm. The silhouette method was used to obtain the best number of clusters in both cases (Fig. 8 (a) and (b)).

Also in this case, the dispersion of silhouette scores is very low. The silhouette average corresponding to k=2 and k=3 is very similar and is the highest value among all the k-options. We finally decided to consider k=3, to have a better discretization of the day hours.

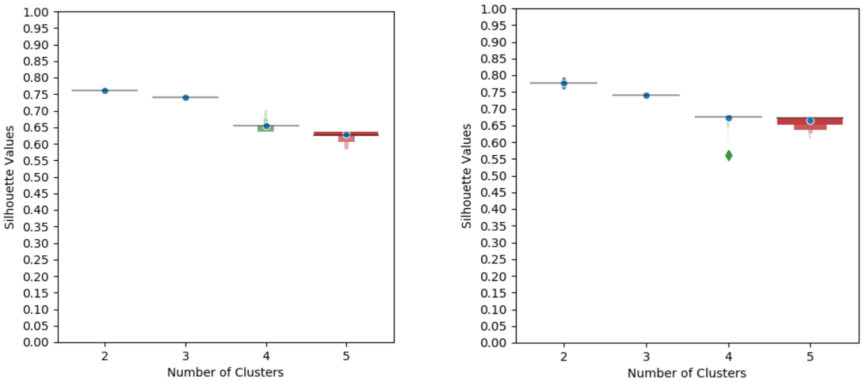

Fig. 8. Boxenplot of silhouette values for **(a)** school-matrix analysis and **(b)** high-school-matrix analysis.

The first two PCs were selected in both cases since they represented more than 98% (Fig. 9) and 99% (Fig. 10) of the variance.

Fig. 9 and Fig. 10 show the outcomes obtained for the school and high-school dataset, respectively. The scores plot summarizes the behavior of the data points in the two components and highlights their similarities. In both cases, three net clusters that represent school-hours (morning and afternoon), evening, and late-night were identified. The loadings plot analyzes the role of all the variables (categories) in the two PCs chosen, their correlations, and their relationships with the day hours. It is worth noting that, in both cases, all the vectors are oriented towards the cluster that represents the class hours, showing the extensive Internet use exclusively for school activities. Furthermore, *EducationalInstitutions* and *SocialNetworking* vectors are almost orthogonal, showing their independence. In particular, *SocialNetworking* mostly occurs around 12:00, when the shift between morning and afternoon classes happens, while *EducationalInstitutions* occurs in the morning, from 8:00 until 12:00, and in the first hours of the afternoon, 13:00 and 14:00.

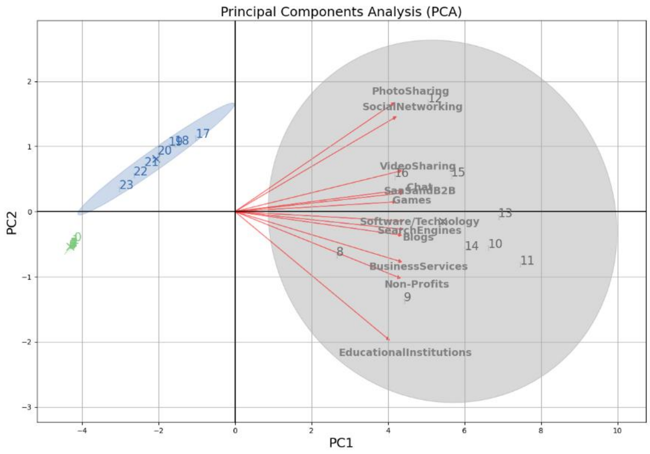

Fig. 9. PCA biplot to identify the school students' behavior during the day hours.

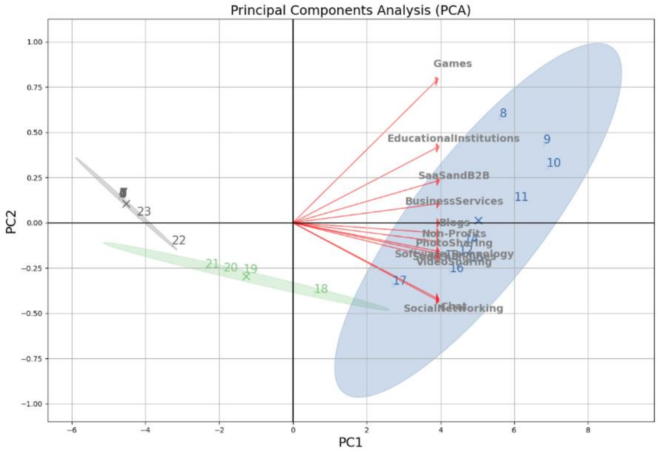

Fig. 10. PCA biplot to identify the high-school students' behavior during the day hours.

With the aim of thoroughly investigating the similarity among all the data points and confirm or add more information to the outcomes of the PCA biplots (Fig. 9 and Fig. 10), a heatmap analysis was used as a complementary analytical tool. Based on previous results, we decided to consider a data matrix in which each day of the week (Monday to Friday) was divided into three hour-groups: *morning* (from 6:00 a.m. to 12 p.m.), *afternoon* (from 12:00 p.m. to 6:00 p.m.), and *evening* (from 6:00 p.m. to 6:00 a.m.). Also in this case, for each of these strips, we grouped (sum) the numbers of queries per category (columns), and centered and standardized the columns' values.

The two hierarchical heatmaps were run using Ward linkage and Euclidean distance. The outcomes obtained for school and high-school datasets are represented in Fig. 11, and Fig. 12 respectively.

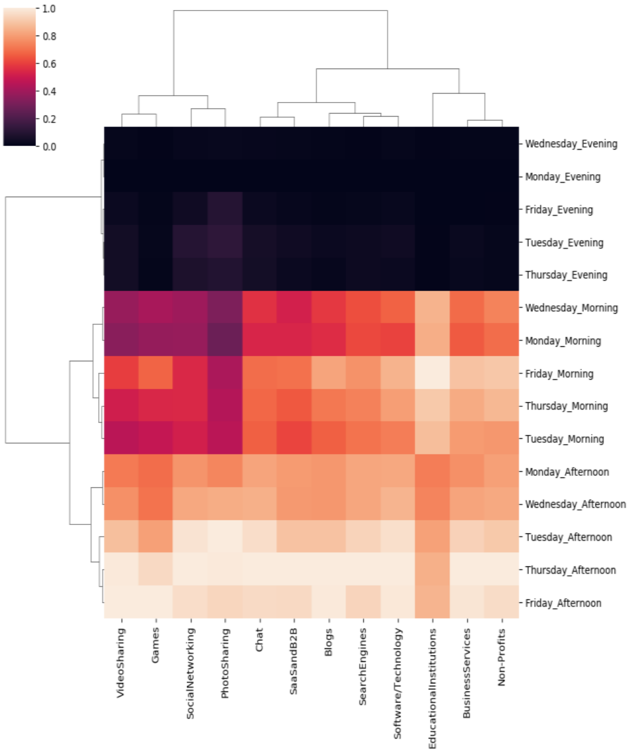

Fig. 11. Heatmap-analysis outcome for school dataset.

From the two heatmaps, considering the left dendrograms, it is noteworthy that the three hour-groups are perfectly identified. Furthermore, the Internet use is very low during the evening/late-night hours. In particular, it seems that high-school students use it more than school students in this time window. This is justified by the fact that teenagers have a more extensive study schedule that may make them study until evening/night. These results confirm and complete the PCA outcomes.

Regarding the hierarchical school grouping of the columns (top dendrograms), the cluster formed by the categories *EducationalInstitutions*, *Non-Profits*, and *BusinessServices* can also be identified in the PCA biplot, highlighting in both cases the category *EducationalInstitutions*. Furthermore, it is interesting to see that *VideoSharing*, *Games*, *SocialNetworking*, and *PhotoSharing* occurs more during the afternoon hours, in particular, on Thursday and Friday afternoon (end of the week). While *EducationalInstitutions* is queried more during the morning, all week long.

Considering the left dendrogram of the high-school heatmap, it is possible to see that the class schedule is highly variable depending on the educational center. Clearly, from the top dendrogram, Tuesday and Thursday afternoon are the time windows with more activities, and teenagers, unlike kids, use *VideoSharing*, *SocialNetworking*, and *PhotoSharing* during classes' hours too. These outcomes enhance PCA results.

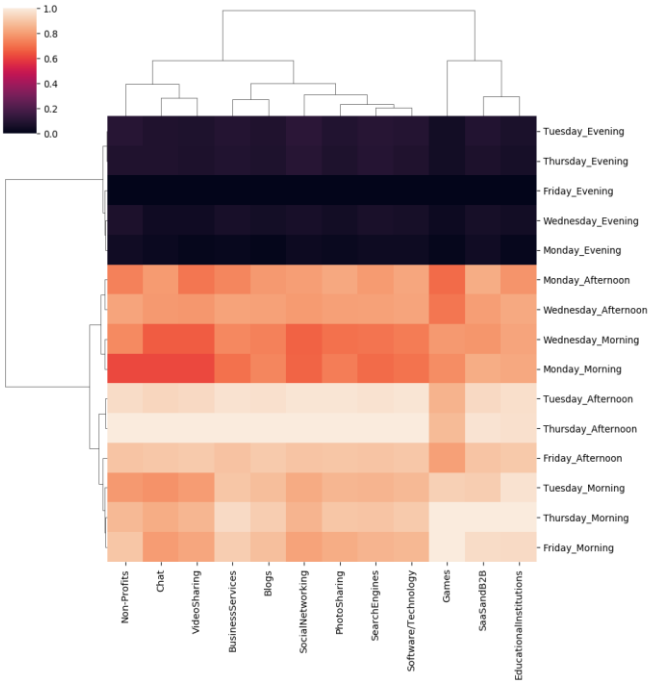

Fig. 12. Heatmap-analysis outcome for the high-school dataset.

5 Concluding Remarks

In this paper, we analyzed the behavior of users in educational networks. For this purpose, we studied a large volume of Domain Name System (DNS) data collected during three months from a major nationwide ESP.

We performed two different analyses with two different aims. The first one, confirmed that kids and teenagers have different Internet-use behavior. The second one described users' Internet behavior during the day hours. It showed that schools students use Internet access, mainly for school activities during class hours. While high-school students also connect to social networks during class time. The differences in the requirements for the applications used at each educational center could be taken into account in the network design process, for example using different design criteria for schools and high-schools.

An in-depth analysis presented that students' Internet-use behavior depends on the classes' schedule, which is highly variable depending on the educational center. The results show that Internet-use behavior is influenced by age-group and time of the day. However, it is independent of the geographical location of the users. The analysis of the user behavior in Internet access is important for any service provider, and the education case is not an exception. The contribution in the particular educational context is not only relevant from the network-operator perspective, but also for other studies combining different sources of information for learning analytics purposes.

Acknowledgement

The research leading to these results received funding from the ANII-FSDA program under proposal number FSDA_1_2018_1_154853.

References

1. One Laptop Per Child, http://one.laptop.org/, last accessed
2. http://www.impo.com.uy/bases/decretos/144-2007/1, last accessed
3. Plan Ceibal, https://www.ceibal.edu.uy/en/institucional, last accessed
4. Cisco – Umbrella, https://umbrella.cisco.com/products/our-cloud, last accessed 2020/03/10.
5. Zhauniarovich, Y., Khalil, I.M., Yu, T., Dacier, M.C.: A Survey on Malicious Domains Detection through DNS Data Analysis. Cryptography and Security 1(1), 1-35 (2018).
6. S. Torabi, A. Boukhtouta, C. Assi and M. Debbabi: Detecting Internet Abuse by Analyzing Passive DNS Traffic: A Survey of Implemented Systems. IEEE Communications Surveys & Tutorials, vol. 20, no. 4, pp. 3389-3415, Fourthquarter (2018).
7. Plonka, D., Barford, P.: Context-aware Clustering of DNS Query Traffic. IMC '08: Proceedings of the 8th ACM SIGCOMM, p. 217–230, Vouliagmeni, Greece (2008).
8. Gao, H., Yegneswaran, V., Chen, Y., Porras, P., Ghosh, S., Haixin Duan, J.J.: An empirical reexamination of global DNS behavior. SIGCOMM '13: Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM, p. 267–278, Hong Kong, China (2013).
9. Ruana, W., Liub, Y., Zhaob, R.: Pattern Discovery in DNS Query Traffic. Procedia Computer Science 17, 80-87 (2013).
10. Su, J., Li, Z., Grumbach, S., Salamatian, K., Han, C., Xie, G.: Toward Accurate Inference of Web Activities from Passive DNS Data. 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), pp. 1-6, Banff, AB, Canada (2018).
11. Schomp, K., Rabinovich, M., Allman, M.: Towards a Model of DNS Client Behavior. In: Karagiannis T., Dimitropoulos X. (eds) Passive and Active Measurement. PAM 2016. Lecture Notes in Computer Science, vol. 9631, Springer, Cham (2016).
12. Li, J., Ma, X., Guodong, L., Luo, X., Zhang, J., Li, W., Guan, X.: Can We Learn what People are Doing from Raw DNS Queries?. IEEE INFOCOM 2018 - IEEE Conference on Computer Communications, pp. 2240-2248, Honolulu, HI (2018).
13. Jia, Z., Han, Z.: Research and Analysis of User Behavior Fingerprint on Security Situational Awareness Based on DNS Log. Research and Analysis of User Behavior Fingerprint on Security Situational Awareness Based on DNS Log. 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC), pp. 1-4, Beijing, China (2019).
14. Cisco – Umbrella Log Formats and Versioning, https://docs.umbrella.com/deployment-umbrella/docs/log-formats-and-versioning, last accessed 2020/03/10.
15. Cisco – Umbrella Manage Content Categories, https://docs.umbrella.com/deployment-umbrella/docs/content-categories#section-content-categories-definitions, last accessed 2020/03/10.

16. Gorgoglione, A., Gioia, A., Iacobellis, V.: A Framework for assessing modeling performance and effects of rainfall-catchment-drainage characteristics on nutrient urban runoff in poorly gauged watersheds. Sustainability, 11, 4933 (2019).
17. Massart D L, Vandeginste B G M, Deming S M, Michotte Y, Kaufman L 1988 Chemometrics-A Text Book (Elsevier: Amsterdam, The Netherlands) chapters 1–4 pp 14–21.
18. Adams, M.J. 1995 Chemometrics in Analytical Chemistry. The Royal Society of Chemistry, Cambridge.
19. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Comput. Surv. 31 264-323(1999).
20. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65(1987).
21. Baker, F.B., Lawrence, J.H. Measuring the Power of Hierarchical Cluster Analysis. Journal of the American Statistical Association, 70, 349, (1975).
22. Friendly, M.: The history of the cluster heat map. The American Statistician, (2009).
23. Hortonworks Data Platform, https://www.cloudera.com/products/hdp.html, last accessed 2020/03/06.
24. Apache Hadoop, https://hadoop.apache.org/, last accessed 2020/03/06.
25. Apache Spark, https://spark.apache.org/, last accessed 2020/03/06.
26. Apache Hive, https://hive.apache.org/, last accessed 2020/03/06.
27. scikit-learn, https://scikit-learn.org/, last accessed 2020/03/06.
28. seaborn, https://seaborn.pydata.org/, last accessed 2020/03/06.