Noname manuscript No.
(will be inserted by the editor)

Extracting Human Behavior Patterns from DNS Traffic

Martín Panza · Diego Madariaga ·
Javier Bustos-Jiménez

Received: date / Accepted: date

Abstract The Internet has evolved in the last decades as a fundamental part of human culture. Human patterns are present in network traffic due to users' activity regarding everyday tasks or other routines. Consequently, these patterns can be found in DNS (Domain Name System) traffic, as it is a critical element for the Internet's working. The present work shows a procedure to detect and extract some of those human patterns by applying Machine Learning techniques on real DNS data. Network traffic retrieved from an authoritative DNS server from the ccTLD (country-code Top Level Domain) from Chile *.cl*, was processed as multiple time series for pattern extraction. Particular and complex techniques have to be used in order to work with this data structure. The procedure consists of a first stage of Clustering analysis, to detect groups of domains based on their activity to analyze their behavior over time and determine persistent patterns; and a second stage of Association Rules extraction, to retrieve specific activity differences between the groups. Finding human patterns in the data could be of high interest to researchers that analyze human behavior regarding Internet usage. Through the application of the proposed procedure, trends and patterns present in DNS traffic were detected, which showed to be consistent over different time portions of the data.

Keywords DNS · Clustering · Time series · Machine Learning · Human Behavior · Association Rules

Nic Chile Research Labs
Santiago, Chile
+56-2-29407787

Martín Panza
E-mail: martin@niclabs.cl

Diego Madariaga
E-mail: diego@niclabs.cl

Javier Bustos-Jiménez
E-mail: jbustos@niclabs.cl

1 Introduction

As a critical component in Internet's infrastructure, the Domain Name System (DNS) plays a vital role in Internet's working. It is responsible for the translation between domain names and IP addresses, and therefore, almost every activity on the Internet starts with a DNS query (and often several) [1]. Additionally, the Internet has turned into an essential part of people's lives, undeniably affecting user's well-being [2]. Taking this into consideration, human behavior patterns can be understood by analyzing Internet traffic, and particularly, DNS traffic. In fact, top-level DNS servers are likely to be the entry of clients' DNS query behavior [3], and therefore, DNS traffic can clearly evidence users' Internet activity.

Indeed, some elementary periodic patterns can be identified by inspecting the temporal behavior of DNS queries. Figure 1 shows the temporal behavior of the number of DNS queries received by a real authoritative name server for the Chilean .cl ccTLD. The figure contains one week of real data, where the DNS queries are grouped into 10-min intervals. The periodicity showed in the figure and the periods of low activity during the night are clearly related to human circadian rhythms. Likewise, human activity is higher during weekdays rather than during the weekend, where Saturday and Sunday correspond to the two lower peaks at the end of the time series.

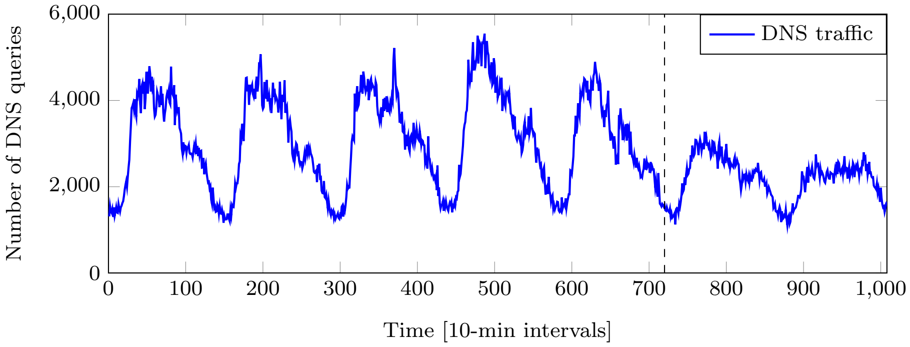


Fig. 1: DNS traffic time series. Number of DNS queries received during one week by an authoritative name server for the Chilean .cl ccTLD.

However, DNS traffic data contains much more information that can also be related to human behavior. The DNS translates human-readable domain names into IP addresses, and therefore, we can extract the queried domain names from each DNS packet. By looking at these domain names, it is possible to infer the content to which the users are attempting to access, and it is possible to extract patterns of human behavior regarding their Internet activity.

Recognizing and studying these patterns could be of high relevance for researchers interested in analyzing human behavior and its relationship with Internet usage. Besides, understanding DNS-related human behavior could help resource management and DNS operators to improve the service provided by their systems.

This work seeks to use Machine Learning techniques on real DNS traffic from an authoritative name server to discover and analyze human patterns. The following sections show a valuable process for this purpose based on state-of-the-art methods and evaluation measures.

2 Related Work

The study of human behavior has always been of great interest to researchers, mainly in the field of sociology [4]. However, understanding human behavior in computer science is an emergent research field that has significantly benefited from the rapid proliferation of wireless devices that frequently report status and location updates [5].

Most of state of the art works that address the study of human behavior through the analysis of networking-related data exploit the high periodicity present in network data. This high periodicity and, consequently, low entropy, is mainly attributed to the impact of the regularity of human patterns [6, 7] on the network state [8].

Recently, the temporal and spatial analysis of data traffic on the mobile network has shown that different human patterns have different effects over the network state, generating distinct data traffic patterns in diverse locations [8]. Additionally, previous research works have shown that human patterns may also impact DNS traffic, as it also exhibits a remarkable 1-day periodicity [9]. Therefore, important conclusions about user behavior can be deduced by analyzing this particular portion of the Internet's traffic, in order to optimize the performance of this critical component of the network.

Additionally, time series analysis has become an essential topic of research lately, mainly because it has been successfully applied over a great variety of research topics. Additionally, the use of time series provides flexibility to adapt to different scenarios, as concepts like similarity and summarization have different visions depending on the specific problem being solved [10]. On top of it, data mining studies have developed various adaptations of the standard techniques to be used with time series [11] since, in general, each problem is addressed with an original procedure that meets particular conditions.

3 Dataset Overview

In this paper, we used data collected directly by the official registry for the Chilean .cl ccTLD: NIC Chile (Network Information Center of Chile) [12]. NIC Chile maintains a network of name servers for the .cl ccTLD worldwide to provide a robust and stable service with excellent response times. The dataset used in this study consists of a month of normal operation traffic from one authoritative name server under the control of NIC Chile, belonging to an anycast configuration along with other servers [13]. This name server is located in Santiago, the capital city of Chile.

For this work's purpose, only the most essential domains on '.cl' were considered because of the vast number of domains that this ccTLD is responsible for, most of which present low activity. We based on Amazon Alexa's top sites from Chile [14] to determine the most relevant domains for our study. We further verified their importance based on the number of DNS queries received for those domains, resulting in 82 high-activity domains. By using the DNS traffic data corresponding to these 82 domains, we created 82 DNS-related time series. Each time series was built by aggregating into 10-min intervals all the successful DNS responses for each of the 82 selected domains. Therefore, each point of these time series corresponds to the number of DNS responses (related to one specific domain) with record types A, NS, AAAA, and MX [15] obtained in a 10-min interval. We selected these four DNS record types because of their importance, as they comprised more than 95% of the whole DNS queries in the dataset used. It is of great relevance to the significance of the present work that the data was captured during a normal operation of a name server, containing real users' usage from a vast population.

It is important to clarify that the DNS traffic dataset contained DNS queries for almost 2,000,000 different Chilean domain names. Therefore, the 82 selected domain names correspond to the 0.004% of all the queried domains. Additionally, 4.7% of all the DNS queries contain queries for one of these 82 domain names, which confirm the importance of these domains. Indeed, these 82 domain names are queried 1,000 times more than the value expected if the queries were evenly distributed among all the domain names. Finally, it has to be acknowledge that part of the traffic will not be perceived by the authoritative name server due to DNS caching over the resolution routine.

4 Methodology

Considering the domains taken into account based on the criteria described in Section 3, an experimental procedure was made consisting of two stages that are further described in the following sections.

The first stage corresponds to a clustering analysis over the 82 time series. This analysis allows us to find groups of domains according to their traffic activity, corresponding to the temporal behavior of the number of queries received from the users. Each domain's time series was pre-processed by applying a Simple Moving Average (SMA) method to reduce noise and capture the time series's regular shape. Additionally, all the time series were pre-processed by applying a Z-Score normalization, which is convenient to compare the distance among the shape of different time series. This pre-processing was helpful to provide the clustering algorithm a smoother and consistent input. The time series clustering algorithm used in the experiments was the Partitioning Around Medoids (PAM) [16], which has the benefit that it uses elements from the dataset as centroids. As explained in Section 5, the selected k value used for further analysis was determined by the internal clustering validation measure: Davies-Bouldin Index [17]. Concerning the time series distance metric used by the algorithm, the Shape-Based Distance proposed by Paparrizos and Gravano [18] was established in a sliding window of 12 hours, i.e., half a day. Before the experiment's execution, different tags were assigned to each domain as a way of both describe the domain's content type and evaluate the results using an external clustering validation measure: Rand Index. Lastly, after obtaining the results and selecting k, we display the groups given by the algorithm and discuss the nature of their domain members.

The second stage's objective was to establish a comparison between the groups obtained in the clustering analysis. To achieve this goal, an association rules analysis was made using a representative of each group, corresponding to the centroid from each cluster obtained in the previous stage. The algorithm used in this phase was the Apriori algorithm [19]. However, to properly feed this algorithm with the time series, a previous procedure to transform time series to a set of transactions was done. The most relevant rules are presented and discussed in Section 6.

Some important aspects of this process were implemented using the R packages *dtwclust* [20], containing time series clustering tools, and *apriori* [21] for association rules analysis. A diagram that summarizes all these procedures is presented in Figure 2.

Finally, some conclusions and future work are proposed in the final section.

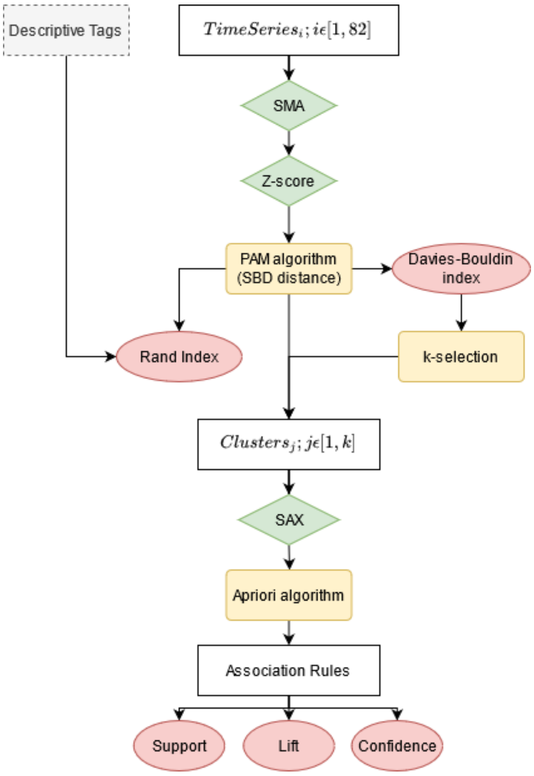


Fig. 2: Procedure summary diagram. Colorless rectangles indicate I/O data. Green rhombuses represent a method for data transformation. Yellow rounded rectangles denote an algorithm or a proceeding. Red ovals contain metrics. The blue plotted line rectangle corresponds to tags selection by ours criteria.

5 Clustering Analysis

5.1 Algorithm and Configuration

The clustering algorithm used for the experiments is the Partitioning Around Medoids, using the Shape-Based Distance (SBD) as a distance measure.

5.1.1 Partitioning Around Medoids

We selected the Partitioning Around Medoids (PAM), as it is different from k-means algorithm since it uses elements from the dataset as centroids. The advantage is that it is less sensitive to outliers as it minimizes dissimilarities between the cluster members, and not squared euclidean distances as k-means does. The algorithm proceeds as follows:

```
Select k domains as medoid-domain.
Link all the other domains to their closest medoid-domain.
Calculate the total cost (sum of dissimilarities).
while total cost decreases do:
    for all medoid-domain do
        for all non-medoid-domain do
            Use the non-medoid-domain as medoid-domain instead of the
            current medoid-domain.
            Link all the other domains to their closest medoid-domain.
            Recalculate the total cost.
            if total cost increased then
                Undo the substitution between the medoid-domain and the
                non-medoid-domains.
            end if
        end for
    end for
end while
```

A specific advantage of this algorithm to the benefit of this work is that, due to its nature, the final centroids are members from one of each cluster. In Section 6 we use this aspect to choose candidates for the Association Rules analysis directly.

5.1.2 Shape-Based Distance

The Shape-Based Distance (SBD) is a similarity measure for time series. It is less costly than the popular Dynamic Time Warping (DTW). It is described by the following equation:

$$SBD(\vec{x},\vec{y}) = 1 - \underset{w}{max}(\frac{CC_w(\vec{x},\vec{y})}{\sqrt{\|x\| \cdot \|y\|}}) \qquad (1)$$

where $CC(x, y)$ is the cross-correlation and w is a value that maximizes $CC_w(x, y)$ based on the convenient shift of the time series with regard to the other one.

This measure reaches values between 0 to 2, and it is highly sensitive to scale. That is why normalization is required. We used Z-Score normalization as suggested by the distance's authors. In addition, we used a half a day window size for the calculations of the similarity.

5.2 Evaluation

Clustering validation measures are divided into two types regarding the information that they require: internal and external. Both have the objective of determining how good the clusters obtained by a clustering algorithm are.

While internal validation measures only require spatial information of the clusters themselves, external validation measures use information that instructs how the result is expected to be, such as what cluster members should or should not be together.

Since we are not interested in adjusting the algorithm to obtain a particular result, an internal validation measure was used to evaluate the clustering algorithm: Davies-Bouldin measure. More specifically, it was used to compare the clusters' quality for different values of k (number of clusters).

Nonetheless, tags were still given to each domain to describe what the domains are related to, allowing further discussion, and an additional external evaluation. The tags assigned to each domain are shown in Table 1.

Table 1: Descriptive tags assigned to the domains

Tag	Description
BA	Banking
BS	Big Stores
EC	E-Commerce
ED	Educational
GO	Governmental
JS	Job Sites
OS	Online shopping
NP	Newspaper
PD	Postal Delivery
RS	Radio Station
SE	Search Engine
SU	Supermarket
TC	Telecommunication
TO	Tourism
TV	Television

5.2.1 Davies-Bouldin Index

Davies-Bouldin Index (D-B) is given by the following equation:

$$DB = \frac{1}{N} \sum_{i=1}^{N} D_i \tag{2}$$

where N is the number of clusters, and:

$$D_i = \underset{i \neq j}{max}(R_{i,j}) \tag{3}$$

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \tag{4}$$

$$S_i = \frac{1}{T_i} \sum_{j=1}^{T_i} d(x, c_i) \tag{5}$$

$$M_{i,j} = d(c_i, c_j) \tag{6}$$

where c_i is the centroid of the cluster i, T_i is the size of the cluster i, and $d(c_i, c_j)$ is the distance between the two clusters.

This index measures the average distance between each cluster and its most similar one. Thus, a lower score means that the quality of the clusters is better.

5.2.2 Rand Index

The Rand Index (RI) is a similarity measure between two clustering solutions. It is given by the following equation:

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

where TP corresponds to the True Positives, i.e., the number of elements grouped together in both clustering results. TN are the True Negatives, elements that are separated in different clusters in both clustering results. FP and FN are False Positives and False Negatives. They represent the elements that belong to the same cluster only in one of the two clustering solutions, but do not belong to the same cluster in the other clustering solution. In which one of the clustering solutions this happens determines what would be a FP or a FN.

In this case, our tags compose a clustering solution that will be compared to the corresponding clustering solution after selecting the k value to obtain the Rand Index.

5.3 Visualization

Time series exhibit a high dimensionality, and the proposed Shaped-Based Distance, utilized to compare them by the Partitioning Around Medoids Algorithm, is not adequately intuitive. Therefore, a visual representation of the clustering results was generated to aid in interpreting the results. This aid is essential for the comparison of the results performed in Section 5.5.3.

The visualization consists of a Metric Multidimensional Scaling (mMDS) [22] applied to a distance matrix that results from the clustering algorithm. This procedure creates a projection to a space with lower dimensionality (in this case two-dimensional) that satisfies the corresponding distances. Even though the solution might not be unique, it greatly helps to analyze the difference between the time series' shapes according to the selected distance. The 2D visualizations were made using Matplotlib [23].

5.4 Data Pre-processing

To reduce the noise in the time series and capture the essence of their shape to facilitate the establishment of comparisons between the clustering algorithm, a smoothing and normalization process was made on every time series. First, a Simple Moving Average (SMA) was performed with five as the number of periods to reduce noise. Secondly, a Z-Score normalization was applied to modify the data scale because the distance measure to be used is sensitive to it.

5.5 Experimentation

The clustering algorithm was performed independently on each of the four weeks of DNS traffic that compose the complete dataset, considering all 82 domains as separate time series.

Following all the configurations described in previous chapters, the only missing parameter intrinsic to any clustering algorithm was selecting the number of clusters to be provided by the algorithm (k value). One can execute the clustering algorithm several times to get a suitable result, particularly by combining the initial medoids. However, all these iterations would follow a general trend regarding the evaluation metric that we established: the Davies-Bouldin index. Moreover, it is affected by the selection of the k value. This is showed in Table 2, where the mean index value for a hundred iterations of the algorithm for each of the four weeks and values of k between 3 and 9 is presented.

Table 2: Mean Davies-Bouldin Index over 100 iterations on each k value for a clustering algorithm perform on four different weeks.

k	**1st Week**	**2nd Week**	**3rd Week**	**4th Week**	**All Weeks**
3	0.523	0.496	0.460	0.565	0.511
4	0.489	**0.448**	**0.432**	0.505	0.469
5	0.498	0.449	0.455	0.476	0.469
6	**0.488**	0.476	0.453	**0.440**	**0.464**
7	0.499	0.486	0.473	0.477	0.484
8	0.526	0.486	0.487	0.484	0.496
9	0.500	0.483	0.517	0.480	0.495

This information is what we have taken into account for selecting a k value equal to 6, for all the results showed later in this section. Although the second week does not present the lowest mean according to Table 2, we will remain with the same k value to support a more straightforward comparison across the results.

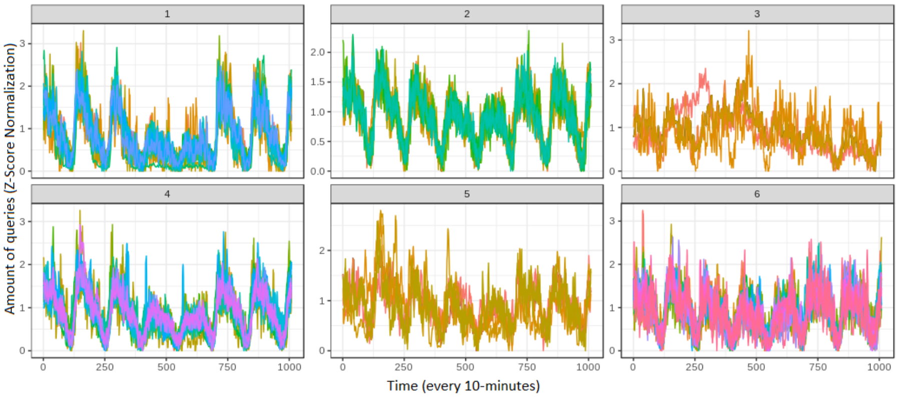
Also, note that the same analysis could be made to obtain a better external evaluation index according to our tagging. However, we do not seek to influence the clustering results; we are interested in what well-conformed clusters show. Thus, we did not consider the Rand Index at this point, even though both indices are highly correlated, as a Pearson's correlation test reveals (0.521) when comparing them from the iterations mentioned above.

The results and discussion for the clustering performed on the first week of data are presented in the following sections for individual analysis. The results obtained for the next three weeks are shown later in section 5.5.3 to discuss the algorithm's consistency over time.

5.5.1 Results

According to Table 2, the time series corresponding to each domain were grouped in 6 different clusters using the PAM algorithm. Figure 3 shows the six resulting clusters, where the 82 domain-related time series are distributed in these clusters. In the sub-figures, each time series is shown with a different categorical color. This slightly shows the shape in which the elements from different clusters differ. However, for a more straightforward interpretation of the distances and grouping of the clusters' members, the procedure mentioned in Section 5.3 was used to generate another visualization shown in Figure 4. Table 3 displays, in detail, the groups obtained by the clustering algorithm, listing all their members by their domain names and their assigned tag. It also shows its Davies-Bouldin Index and Rand Index, as described in sections 5.2.1 and 5.2.2 respectively.

Fig. 3: All domains time series for each cluster.



5.5.2 Discussion

As observed in Figure 3, the process successfully made groups of domains depending on each time series's attributes. Even in such a straightforward visualization, the differences between the time series' arrangements can be seen between distinct clusters. One clear aspect is on the weekend, which can be easily identified as the lower peaks in the middle zone of the time series in Cluster 1. These peaks correspond to Saturday and Sunday in the data, which means that users of those domains reduce their activity on weekends. On the other hand, members from other clusters, such as Cluster 6, do not demonstrate these distinctions between weekdays and weekends, as users of those domains maintain uniform usage throughout the whole week. Moreover, members from Cluster 3 show an opposite behavior, with peaks on weekends.

Nevertheless, all the domains seem to share in common a decrease in activity during nighttime.

The clusters listed in Table 3 also demonstrate a valuable outcome as patterns can be observed when considering the domains' content type, especially when considering our initial descriptive tags. For instance, every domain originally tagged as Educational [ED] was grouped together in Cluster 6, just for *aiep* who was assigned to Cluster 1. This tag considers many of the most important universities and institutes in Chile. Such as Universidad de Chile (*uchile*), Universidad Católica de Chile (*uc*), Universidad de Concepción (*udec*), and Departamento Universitario Obrero y Campesino (*duoc*). As well as some government educational-related domains, such as *conycit* (National Commission for Scientific and Technological Research), and also *mineduc* (Ministry of Education) and *curriculumnacional* (National Curriculum) that were originally tagged as Governmental [GO]. Logically, this kind of domain should present similar traffic, which is successfully recognized by the algorithm. However, some other not-related domains are also included in the cluster, such as *chilevision* [TV] or *chileautos* [EC].

Table 3: Domains and tags by cluster. Results from the first week.

Cluster	Domain	Tag
1	aiep	ED
	bancoedwards	BA
	bancosantiago	BA
	bci	BA
	bluex	PD
	chileatiende.gob	GO
	chilexpress	PD
	correos	PD
	dt.gob	GO
	entel	TC
	mercadopublico	GO
	officebanking	BA
	scotiabankazul	BA
	scotiabankchile	BA
	scotiabank	BA
	sii	GO
	sistemadeadmision	GO
2	13	TV
	24horas	TV
	adnradio	RS
	biobiochile	RS
	elmostrador	NP
	mega	TV
	paris	BS
	pcfactory	BS
	soychile	NP
	t13	TV
	tvn	TV
	wom	TC
4	bancochile	BA
	bancoestado	BA
	bancofalabella	BA
	bancoripley	BA
	claveunica.gob	GO
	cmr	BS
	dafiti	OS
	despegar	TO
	emisora	RS
	lider	SU
	mercadolibre	EC
	pjud	GO
	publimetro	NP
	registrocivil	GO
	ripley	BS
	santander	BA
	sodimac	BS
	trabajando	JS
	transbank	BA
	yapo	EC
3	airbnb	TO
	google	SE
	redgol	NP
	tripadvisor	TO
5	clarochile	TC
	df	NP
	groupon	TO
	itau	BA
	linio	OS
6	abcdin	BS
	bsale	OS
	buscalibre	OS
	chileautos	EC
	chiletrabajos	JS
	chilevision	TV
	comunidadescolar	ED
	conicyt	ED
	cooperativa	RS
	curriculumnacional	GO
	duoc	ED
	easy	BS
	extranjeria.gob	GO
	inacap	ED
	laborum	JS
	mercadopago	BA
	mineduc	GO
	mitarjetacencosud	BA
	movistar	TC
	santotomas	ED
	uc	ED
	uchile	ED
	udec	ED
	webescuela	ED

Davies-Bouldin Index	**Rand Index**
0.292	0.772

Another estimable result corresponds to the group formed on Cluster 2 as it contains all the domains tagged as Television [TV], except for one. It also incorporates two Radio-Station [RS], and two Newspaper [NP] tagged domains. If we consider that all these tags fit as part of mass media, then we distinguish an interesting pattern captured by our procedure.

We can also observe that the three domains that we manually tagged as Postal-Delivery [PD] were grouped together in Cluster 1. In this cluster, there are also five Governmental [GO] domains and seven Banking [BA] domains.

Additionally, two of the four domains previously tagged as Tourism [TO] were grouped in the smallest cluster along with two other domains.

5.5.3 Consistency Over Time

The same procedure was applied to each of the following three weeks of DNS traffic to compare how the results vary over time and inspect if the newly formed clusters are consistent with the initial ones from the first week. The following pages show, as in Section 5.5.1, a detailed table of all the domains with tags and their respective cluster number, and a figure of a two-dimensional space projection of their Shape-Based distances (SBD).

To ensure a better comparison, the number that represents each of the clusters was selected in each week based on the domains it has, so that similar clusters could be easily identified throughout the weeks. The same applies to the colors of the clusters in each visual representation.

From the tables listing the grouped domains (Tables 4-6), we can find consistency within some of the groups. For instance, in the four weeks, Cluster 2 is mostly composed with tags TV (Television), RS (Radio Station), and NP (Newspaper), which are related to mass media domains. Examples of this are *24horas* [TV], *mega* [TV], *tvn* [TV], *elmostrador* [NP], and *biobiochile* [RS] which are frequently grouped together. This cluster corresponds to the color red in the 2D-projections.

The second cluster in which a pattern can be recognized is Cluster 1, frequently composed almost entirely by BA (banking), GO (governmental), and PD (Postal Delivery) domains. Particularly, it always contains two of the three domains tagged as PD in the entire dataset. *Bancosantiago* [BA], *scotiabank* [BA], *chilexpress* [PD], and *sistemadeadmision* [GO] are domains that are always grouped in this cluster. Its color in visualizations is blue.

A cluster seems to relate commerce and shopping domains. Number 4 often contains *lider* [SU], *cmr* [BS], *ripley* [BS], *sodimac*[BS], and *yapo* [EC]. Moreover, *bancofalabella* [BA] and *bancoripley* [BA] are banks that belong to retail stores, as well as *transbank* [BA], that manages Chilean debit and credit card transactions. Represented by green.

One last cluster that is worth discussing is number 6, that groups together Educational (ED) and Job Sites (JS) tags. For example *santotomas* [ED], *udec* [ED], *uc* [ED], *webescuela* [ED], *chiletrabajos* [JS] and *laborum* [JS]. In addition, as stated also before, some government domains related to the two previous subjects are also frequently contained. *Mineduc* [GO] and *curriculumnacional* [GO]. The color of this cluster is light blue.

The remaining clusters are usually smaller, and they do not manifest an evident pattern. However, they do tend to collect what would be outliers in the projections.

The comparison of the two-dimensional visualizations in Figure 4 also shows consistency over the four weeks as clusters tend to be in a similar position with regard to the others. For example, red is consistently closer to light blue than it is to green or blue. In addition, as shown in Table 7, another consistent behavior is seen at the values obtained for both internal evaluation (D-B index) and external evaluation (Rand index).

Given all the above, it is possible to assure that human behavior patterns influence the DNS traffic of the domains, establishing important differences between them, that can be detected by the used time series distance measure. Moreover, these patterns can be detected by the clustering algorithm to successfully create groups whose members show similar behavior and are very likely to share content meaningful to humans. Thus, detecting human patterns in DNS is feasible by employing clustering techniques.

Table 4: Domains and tags by cluster. Results from the second week.

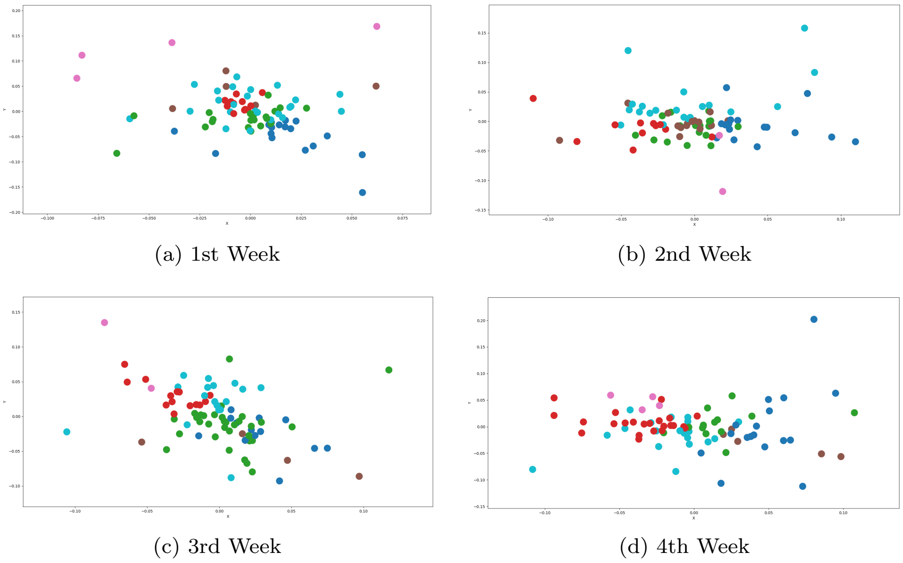
Cluster	Domain	Tag
1	bancochile	BA
	bancoedwards	BA
	bancosantiago	BA
	bci	BA
	bluex	PD
	chileatiende.gob	GO
	chilexpress	PD
	emisora	RS
	entel	TC
	mercadopublico	GO
	mitarjetacencosud	BA
	officebanking	BA
	scotiabankazul	BA
	scotiabankchile	BA
	scotiabank	BA
	sii	GO
	sistemadeadmision	GO
3	correos	PD
	itau	BA
2	24horas	TV
	biobiochile	RS
	chilevision	TV
	clarochile	TC
	cooperativa	RS
	elmostrador	NP
	groupon	TO
	linio	OS
	redgol	NP
	tripadvisor	TO
	tvn	TV
5	13	TV
	bsale	OS
	df	NP
	extranjeria.gob	GO
	laborum	JS
	lider	SU
	mega	TV
	mercadopago	BA
	paris	BS
	registrocivil	GO
	soychile	NP
	t13	TV
	trabajando	JS
	transbank	BA
	wom	TC
4	abcdin	BS
	adnradio	RS
	aiep	ED
	bancoestado	BA
	bancofalabella	BA
	bancoripley	BA
	dafiti	OS
	despegar	TO
	dt.gob	GO
	google	SE
	mercadolibre	EC
	pjud	GO
	publimetro	NP
	ripley	BS
	santander	BA
	sodimac	BS
	yapo	EC
6	airbnb	TO
	buscalibre	OS
	chileautos	EC
	chiletrabajos	JS
	claveunica.gob	GO
	cmr	BS
	comunidadescolar	ED
	conicyt	ED
	curriculumnacional	GO
	duoc	ED
	easy	BS
	inacap	ED
	mineduc	GO
	movistar	TC
	pcfactory	BS
	santotomas	ED
	uc	ED
	uchile	ED
	udec	ED
	webescuela	ED

Table 5: Domains and tags by cluster. Results from the third week.

Cluster	Domain	Tag
1	aiep	ED
	bancoripley	BA
	bancosantiago	BA
	chileatiende.gob	GO
	chilexpress	PD
	correos	PD
	extranjeria.gob	GO
	mercadopublico	GO
	mitarjetacencosud	BA
	registrocivil	GO
	scotiabank	BA
	sistemadeadmision	GO
	trabajando	JS
2	24horas	TV
	biobiochile	RS
	bsale	OS
	chilevision	TV
	comunidadescolar	ED
	cooperativa	RS
	elmostrador	NP
	groupon	TO
	linio	OS
	mega	TV
	mercadolibre	EC
	movistar	TC
	tvn	TV
	uchile	ED
5	bci	BA
	entel	TC
	ripley	BS
	sodimac	BS
4	13	TV
	abcdin	BS
	adnradio	RS
	airbnb	TO
	bancochile	BA
	bancoedwards	BA
	bancoestado	BA
	bancofalabella	BA
	bluex	PD
	clarochile	TC
	claveunica.gob	GO
	cmr	BS
	conicyt	ED
	despegar	TO
	dt.gob	GO
	easy	BS
	emisora	RS
	google	SE
	lider	SU
	officebanking	BA
	paris	BS
	pjud	GO
	publimetro	NP
	santander	BA
	scotiabankazul	BA
	scotiabankchile	BA
	sii	GO
	soychile	NP
	t13	TV
	transbank	BA
	wom	TC
	yapo	EC
3	inacap	ED
	redgol	NP
6	buscalibre	OS
	chileautos	EC
	chiletrabajos	JS
	curriculumnacional	GO
	dafiti	OS
	df	NP
	duoc	ED
	itau	BA
	laborum	JS
	mercadopago	BA
	mineduc	GO
	pcfactory	BS
	santotomas	ED
	tripadvisor	TO
	uc	ED
	udec	ED
	webescuela	ED

Table 6: Domains and tags by cluster. Results from the fourth week.

Cluster	Domain	Tag
2	13	TV
	24horas	TV
	abcdin	BS
	adnradio	RS
	buscalibre	OS
	chileautos	EC
	chilevision	TV
	clarochile	TC
	comunidadescolar	ED
	despegar	TO
	elmostrador	NP
	groupon	TO
	linio	OS
	mega	TV
	mercadopago	BA
	pcfactory	BS
	redgol	NP
	soychile	NP
	t13	TV
	trabajando	JS
	tvn	TV
	uchile	ED
	wom	TC
1	aiep	ED
	bancochile	BA
	bancoedwards	BA
	bancosantiago	BA
	bluex	PD
	chilexpress	PD
	claveunica.gob	GO
	duoc	ED
	google	SE
	itau	BA
	mitarjetacencosud	BA
	officebanking	BA
	scotiabankazul	BA
	scotiabankchile	BA
	scotiabank	BA
	sii	GO
	sistemadeadmisionr	GO
4	bancoestado	BA
	bancofalabella	BA
	bancoripley	BA
	cmr	BS
	easy	BS
	emisora	RS
	lider	SU
	publimetro	NP
	registrocivil	GO
	ripley	BS
	santander	BA
	sodimac	BS
	transbank	BA
	yapo	EC
3	airbnb	TO
	biobiochile	RS
	cooperativa	RS
	mercadolibre	EC
5	bci	BA
	correos	PD
	dt.gob	GO
	entel	TC
	mercadopublico	GO
6	bsale	OS
	chileatiende.gob	GO
	chiletrabajos	JS
	conicyt	ED
	curriculumnacional	GO
	dafiti	OS
	df	NP
	extranjeria.gob	GO
	inacap	ED
	laborum	JS
	mineduc	GO
	movistar	TC
	paris	BS
	pjud	GO
	santotomas	ED
	tripadvisor	TO
	uc	ED
	udec	ED
	webescuela	ED



(a) 1st Week

(b) 2nd Week

(c) 3rd Week

(d) 4th Week

Fig. 4: Comparison between two-dimensional representations of the cluster results for each of the four weeks studied. Each representation is obtained by the application of Metric Multidimensional Scaling (mMDS) algorithm, which maps the points to a reduced dimensional space where the distances (SBD) remain as similar as possible for visualization means. Each color represents a cluster. 1: blue, 2: red, 3: pink, 4: green, 5: brown, 6: light-blue.

Table 7: Cluster evaluation measures results for each of the four studied weeks.

Week	Davies-Bouldin Index	Rand Index
1	0.292	0.772
2	0.254	0.788
3	0.281	0.723
4	0.278	0.775

6 Association Rules

To establish comparisons between the groups obtained from the clustering algorithm, association rules are expected to highlight the trends and patterns within the time series. For this purpose, we will use the clusters obtained in Section 5.5.1. The resulting centroid from every cluster, which corresponds to a domain's time series, was considered as the representative for the experiments and analysis performed in this section. In this way, the association rules procedure was applied on six time series representing each cluster's members.

The association rules algorithm used is the popular Apriori algorithm. In order to feed it with our data, some transformations were required as a preprocessing stage. That is why an SAX was used to convert the time series to symbols, in addition to a rule for feature extraction.

6.1 Apriori Algorithm

Apriori algorithm was designed to generate association rules that indicate patterns and trends inside a dataset composed of multiple collections of items commonly associated with transactions. It focuses on the frequency with which the items appear in the transactions, and with what other items they are usually present.

The algorithm receives a minimum support as input, as well as the transactions, and generates candidate itemsets whose appearances in the transactions are filtered by the minimum support given. Finally, it outputs all the association rules that remain. Selecting the relevant rules after this process falls entirely to the user criteria, depending on some common evaluation indicators for these rules:

1. Support: Percentage of the total number of transactions in which a group of elements X is included.
$$Supp(X) = \frac{|\{t \epsilon T; X \subseteq t\}|}{T} \tag{8}$$
2. Confidence: Proportion of transactions that include elements in group X in which elements from group Y are included as well.
$$Conf(X \rightarrow Y) = \frac{Supp(X \cup Y)}{X} \tag{9}$$
3. Lift: Measures how likely are elements from group Y to be included in the transactions if X is included.
$$Lift(X \rightarrow Y) = \frac{Supp(X \cup Y)}{Supp(X) \times Supp(Y)} \tag{10}$$

where T is the total number of transactions and t is a single transaction.

6.2 Data Pre-Processing

Given that the Apriori algorithm receives a list of transactions as input, a transformation must be previously made to the time series. A direct solution is transforming the time series into symbols and pass collections of symbols to the algorithm. This is taken care of by the Symbolic Aggregate approXimation (SAX) [24].

However, SAX uses Piecewise Aggregate Approximation (PAA) to obtain the symbolic values. This procedure reduced the time series length from 1008 points to 168. Five symbols were used in the transformation, resulting in the following time series:

$$S = \{\, s_t : t \,\epsilon\, T,\; s \,\epsilon\, \{a, b, c, d, e\} \,\} \tag{11}$$

where e corresponds to the previous time series's highest values, and a to lowest ones. Also, $|T| = 168$.

Additionally, one last feature was added to the time series to maintain some relevant information: using the resulting time series describes in equation (11), each symbol was assigned an integer in the following way: a=0; b=1; c=2; d=3; e=5.

This was made to obtain the difference every two points in the time series to establish a measure of flow change in the traffic to not only know its position at a given time, but also its direction.

For example, if a time series has the symbol b at a given point, and in the next point it changes to d, we will note this change as $d - b = 4 - 2 = 2$, and we will say that it increased by 2.

Adding this feature and grouping by every two points leaves our final DNS traffic time series as:

$$S = \{\, (s, n)_t : t \,\epsilon\, T;\; s \,\epsilon\, \{a, b, c, d, e\}\,;\; n \,\epsilon\, \mathbb{N};\; n \,\epsilon\, [-4, 4]\,\} \tag{12}$$

With $|T| = 84$.

This is the final form of the time series that the Apriori algorithm received as a transactions array.

6.3 Results

Table 8 shows what we consider as the most relevant rules after mining the association rules resulting from the Apriori algorithm. The table is subdivided by rules that contain only numeric values, only alphabetic values, and both of them.

6.4 Discussion

The rules showed in Table 6 indicate some patterns in the comparison between the members of each cluster obtained in Section 5.

For example, rule number 3 tells us that every time there was a significant increase (magnitude 2) experienced in Clusters 2 and 4, there was also the same increase in Cluster 1 with a tremendously high lift value; however, not with a considerable value of support.

Rule number 2 states, with a high lift value, that if Clusters 4, 5, and 6 experience a decrease, we can safely expect that Cluster 2 will decrease too. With higher support but lower lift, rule number 8 states that if only Clusters 5 and 6 decrease, Cluster 2 will decrease likewise. Rules number 11 and 12 indicate that this behavior will also occur in the other way. That is, if Cluster 2 experiences a decrease along with 4 or 5, the remaining one will be very likely to decrease as well.

As for the rules containing symbols, some rules like 16, 17, 21, and 22 tell us what clusters tend to stay in their peaks or valleys when other clusters experience the same. However, other rules such as number 18 tell us that when

Table 8: Relevant Association Rules obtained from the Apriori algorithm.

Number	Body	Head	Support	Confidence	Lift
1	C3=0, C4=-1	C5=-1	0.11	0.818	2.864
2	C4=-1, C5=-1, C6=-1	C2=-1	0.06	1	5.600
3	C2=2, C4=2	C1=2	0.04	1	28
4	C2=1	C3=0	0.13	0.917	1.510
5	C3=0, C5=0, C6=0	C4=0	0.18	1	1.615
6	C2=0, C3=-1	C5=0	0.13	0.917	1.878
7	C1=0, C2=0, C3=0, C5=0, C6=0	C4=0	0.10	1	1.615
8	C5=-1, C6=-1	C2=-1	0.07	0.857	4.800
9	C2=-1, C3=0, C5=-1	C6=-1	0.06	0.833	4.118
10	C1=0, C2=-1	C4=-1	0.07	1	4
11	C2=-1, C5=-1	C4=-1	0.10	0.889	3.556
12	C2=-1, C4=-1	C5=-1	0.010	0.800	2.800
13	C2=0, C3=-1, C6=0	C5=0	0.11	1	2.049
14	C1=1, C2=0	C6=0	0.08	1	1.474
15	C1=0, C2=0, C4=1	C6=1	0.036	1	12
16	C1=a, C2=a	C6=a	0.18	1	5.600
17	C1=a, C4=a, C5=a, C6=a	C2=a	0.12	1	5.250
18	C2=c, C3=e	C6=c	0.08	1	4.941
19	C2=b, C3=a, C6=b	C4=b	0.07	1	4.667
20	C1=e, C4=e	C6=e	0.14	0.800	4.200
21	C5=e, C6=e	C1=e	0.13	0.917	4.053
22	C6=e	C4=e	0.18	0.938	3.938
23	C1=e, C4=e, C6=e	C5=e	0.12	0.833	3.684
24	C4=b, C6=b	C2=b	0.13	0.917	3.667
25	C1=c, C6=b	C2=b	0.07	0.857	3.429
26	C1=b, C6=c	C5=b	0.07	0.857	3.130
27	C1=c, C6=b	C2=b	0.07	0.857	3.429
28	C3=c, C6=c	C4=b	0.06	0.833	3.889
29	C1=d, C4=c	C1=-1	0.06	1	3.652
30	C2=c, C5=b	C5=1	0.06	0.833	5
31	C2=0, C5=0, C1=e, C5=e	C3=-1	0.07	0.857	4.500
32	C4=0, C3=a	C3=0	0.08	1	1.647

some clusters are currently in their top or bottom values, others can be found in their intermediate values; in this case, Cluster 6 always obtained *c* value when Cluster 2 was in *c*, but Cluster 3 was in his peak *e*.

Considering the analysis from the previous sections, we can associate Cluster 2 with mass media domains and Cluster 3 with shopping domains. Rule number 4 tells with high support that an increase in Cluster 2 will be likely to be accompanied by no change in Cluster number 3. In other words, when the number of queries for mass media domains is increasing, shopping domains will not experience that increase.

Similarly, Cluster 1 corresponds to banking and government institutions domains, and Cluster 6 to educational domains. From rule number 14, we can observe that when banking/government domains' traffic increases and mass media does not change, then educational domains will not experience a significant increase or decrease. Comparably, rule number 16 also involves these clusters, saying that when banking/government domains are on their lowest values, as well as mass media domains, then educational domains are very likely to be on their lowest traffic.

Following that analysis, rule number 25 says that when banking/government domains are in their middle amount of traffic and educational domains closer to their valley, then mass media domains will be closer to their valley as well.

Finally, some more complex rules regarding both symbols and numeric changes were obtained in the last rows. For example, they tell us that when Cluster 2 has value *c* and Cluster 5 has value *b*, Cluster 5 tends to increase with a very high lift index. (Rule number 30).

Another case is in rule number 32, saying that when Cluster 4 is not changing its activity and Cluster 3 is at its lowest activity, Cluster 3 tends to maintain its behavior. This corresponds to information that is tremendously hard to obtain by other means.

7 Conclusions and Future Work

The procedure proposed in this work was able to identify some patterns in the used time series data. The first stage of our experimentation was able to group domains with similar content for humans, obtaining an acceptable external evaluation index to further comparison, but most importantly demonstrating semantic coherence in the domains that were grouped. Results also showed consistency in the clustering grouping over the four different weeks of the dataset, which also contributes to conclude on the presence of strong patterns in DNS traffic. As for the second stage, association rules showed interesting trends when comparing the centroids from each cluster that could be useful for performing further analysis and pattern mining.

The semantic meaning of the found clusters, in addition to the association rules, allows us to compare their activity, showing trends that could be of interest to analyze how humans behave on a daily basis. For example, mass media domains do not show a traffic increase when banking/government domains do. On the other hand, studying when peaks and valleys are observed on different groupings of human activity could be important to service providers to optimize the resources in favor of the amount of traffic they expect to experience, or based on the type of users they attend.

Taking these results into account, we conclude that human patterns are present in the DNS data and that these techniques were able to find some of them. This demonstrates that they could be mined and recognized using the appropriate methods and data processing. Additionally, the described procedure applied to DNS data allows inferring trends over a wide population without carrying out surveys on people.

Every step from our procedure was associated with an evaluation index as a way of comparison. We suggest as future work the use of other methods that could both find different patterns in the data, and improve the quality of their extraction. Moreover, we claim an achievement of our goal of finding human patterns present in DNS data, however, we encourage a more in-depth analysis of the patterns singularly, to recognize more detailed information about them. We strongly believe that these patterns could be of interest for researchers that analyze human behavior, in this case over activity on the Internet.

References

1. Stéphane Bortzmeyer. DNS Privacy Considerations. RFC 7626, August 2015.
2. John A Bargh and Katelyn YA McKenna. The internet and social life. *Annu. Rev. Psychol.*, 55:573–590, 2004.

3. Z Whang and Shian-Shyong Tseng. Anomaly detection of domain name system (dns) query traffic at top level domain servers. *Scientific Research and Essays*, 6(18):3858–3872, 2011.
4. Bernard Berelson and Gary A Steiner. Human behavior: An inventory of scientific findings. 1964.
5. Nicola Bui, Matteo Cesana, S Amir Hosseini, Qi Liao, Ilaria Malanchini, and Joerg Widmer. A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques. *IEEE Communications Surveys & Tutorials*, 19(3):1790–1821, 2017.
6. Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779, 2008.
7. Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, Carlos Sarraute, Jorge Brea, and Ignacio Alvarez-Hamelin. On the regularity of human mobility. *Pervasive and Mobile Computing*, 33:73–90, 2016.
8. Huandong Wang, Fengli Xu, Yong Li, Pengyu Zhang, and Depeng Jin. Understanding mobile traffic patterns of large scale cellular towers in urban environment. In *Proceedings of the 2015 Internet Measurement Conference*, pages 225–238. ACM, 2015.
9. Diego Madariaga, Martín Panza, and Javier Bustos-Jiménez. Dns traffic forecasting using deep neural networks. In *International Conference on Machine Learning for Networking*, pages 181–192. Springer, 2018.
10. Carmelo Cassisi, Placido Montalto, Marco Aliotta, Andrea Cannata, Alfredo Pulvirenti, et al. Similarity measures and dimensionality reduction techniques for time series data mining. *Advances in data mining knowledge discovery and applications*, pages 71–96, 2012.
11. Tak-chung Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, 2011.
12. NIC Chile. Official registry for the .cl cctld.
13. NIC Chile. .cl nameservers map.
14. Amazon. Amazon alexa topsites, 2019.
15. Paul V Mockapetris. Rfc1035: Domain names-implementation and specification, 1987.
16. Leonard Kaufman and Peter J Rousseeuw. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 344:68–125, 1990.
17. David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.
18. John Paparrizos and Luis Gravano. k-shape: Efficient and accurate clustering of time series. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1855–1870. ACM, 2015.
19. Han Jiawei, Micheline Kamber, and Morgan Kaufmann. Data mining: Concepts and techniques. 2001. *University of Simon Fraser*, 2001.
20. Alexis Sarda-Espinosa. *dtwclust: Time Series Clustering Along with Optimizations for the Dynamic Time Warping Distance*, 2019. R package version 5.5.4.
21. Michael Hahsler, Sudheer Chelluboina, Kurt Hornik, and Christian Buchta. The arules r-package ecosystem: Analyzing interesting patterns from large transaction datasets. *Journal of Machine Learning Research*, 12:1977–1981, 2011.
22. I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, 2005.
23. J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.
24. Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11. ACM, 2003.