Discovering users with similar internet access performance through cluster analysis

Discovering users with similar Internet access performance through cluster analysis

Tania Cerquitelli*, Antonio Servetti, Enrico Masala

Control and Computer Engineering Department, Politecnico di Torino, Corso Duca degli Abruzzi, 24 – 10129 Torino, Italy.

Abstract

Users typically subscribe to an Internet access service on the basis of a specific download speed, but the actual service may differ. Several projects are active collecting Internet access performance measurements on a large scale at the end user location. However, less attention has been devoted to analyzing such data and to inform users on the received services. This paper presents MiND, a cluster-based methodology to analyze the characteristics of periodic Internet measurements collected at the end user location. MiND allows to discover (i) *groups of users* with a similar Internet access behavior and (ii) *the (few) users* with somehow anomalous service. User measurements over time have been modeled through histograms and then analyzed through a new two-level clustering strategy. MiND has been evaluated on real data collected by Neubot, an open source tool, voluntary installed by users, that periodically collects Internet measurements. Experimental results show that the majority of users can be grouped into homogeneous and cohesive clusters according to the Internet access service that they receive in practice, while a few users receiving anomalous services are correctly identified as outliers. Both users and ISPs can benefit from such information: users can constantly monitor the ISP offered service, whereas ISPs can quickly identify anomalous behaviors in their offered services and act accordingly.

Keywords: Cluster analysis, Internet access performance, Anomaly detection, Network monitoring

*Corresponding author. Phone: +39-011-0907178, Fax: + 39-011-0907099.
Email addresses: tania.cerquitelli@polito.it (T. Cerquitelli), antonio.servetti@polito.it (A. Servetti), enrico.masala@polito.it (E. Masala).

Submitted to Expert Systems with Applications

1. Introduction

Currently the vast majority of people use the Internet service for a wide range of everyday activities. Internet access is obtained by signing a contract between the subscriber (i.e., the final user) and an Internet service providers (ISP). Each subscription is linked to a maximum theoretical download speed, which sometimes cannot be achieved due to many factors (e.g., technical issues, service delivery optimization, business rules). Thus, the received service, in particular the download speed experienced in practice, may differ from the advertised value, and neither the users nor the ISP might easily detect such fact.

Different projects have been developed to monitor the Internet access performance on a large scale by frequently measuring the download speed at the end user location. Open source tools, such as NDT (NDT, 2016) and Neubot (Nexa Center, 2016), are voluntarily installed on user computers and they can provide basic information, e.g., the received download speed in the last few minutes, to the users. Furthermore, the collected data (partially anonymized) are also stored in publicly-available repositories for further inspection. An interesting but relatively unexplored research issue is how to analyze the large volume of collected measurements over time to verify whether the service received by the users is coherent with the one of other users with the same subscription or if there are anomalies. The latter information is, in general, useful for both users and ISPs. Users might be informed of the disservice which might be otherwise unnoticed or difficult to detect, and ISP might be alerted so that they can discover potentially unexpected network behavior.

In this paper we propose a novel data analytics methodology, named MiND (Mining Neubot Data), aiming at analyzing the statistical distribution of active measurements of Internet access download speed to address two research questions: (i) Statistical behaviors of the Internet access performance received at user locations are sufficiently similar to be clustered in groups? (ii) It is possible to detect some anomalous patterns in the Internet access performance that deserve to be investigated in-depth to understand their root causes?

To address the previous questions, we employed an exploratory analytics technique, i.e., cluster analysis. This analysis method identifies groups

of objects that share similar properties. Since it does not require previous knowledge of data (i.e., class labels, which in our case are anomalous services and services coherent with the one of other users with the same subscription), it has been widely exploited in many application domains, such as web page content (Chehreghani et al., 2009), social networks (van Dam & van de Velden, 2015), medical data (Combes & Azema, 2013; Cerquitelli et al., 2016), network data (Baralis et al., 2013).

In our context, MiND analyzes the statistical distribution of the download speed measurements over time (through a frequency histogram) collected at the user locations to group Internet users into homogeneous and cohesive groups according to the broadband access service that they really experience. In case of users with a regular access service, most of the download speed measurements are close to their maximum download speed and there are few or no occurrences of speed values below that threshold. Moreover, it is normal that the measured speed occasionally vary (i.e., few measurements are much lower than that the maximum download speed). However, when the distribution of the download speed measurements is anomalous over time, it may be a symptom of the fact that the ISP might not be able to provide the expected service with good reliability. From the point of view of the single user, if the user experiences a download speed similar to the one of a group of other users in a given considered collection we may assume that users receive a service coherent with the subscribed one. Otherwise, we assume that an anomalous behavior has been detected. In the latter case, both the user and the ISP should be informed: users might be interested to know that in practice they receive a service different from the subscribed one, whereas ISP might have the opportunity to investigate further the unexpected network behavior and eventually fix it.

The main novelties of MiND are fourfold. (i) *Data transformation.* To highlight the relevance of Internet access in terms of bandwidth, collected measurements (download speed measurements repeated over time) have been represented through frequency histograms. Specifically, Internet bandwidths are divided into intervals (or bins) defined by a domain expert. Each histogram reports, for each bin, the total number of measurements performed by a single user. Thus, the histogram allows to compactly model all the measurements performed by the same user over time. (ii) *Two-level clustering strategy.* To correctly identify groups of users according to the download speed that they really experienced and to correctly identify anomalous patterns, a two-level clustering strategy has been proposed, based on the DB-

SCAN (M. Ester et al., 1996) and K-means (J. A. Hartigan & M. A. Wong, 1979) algorithms. The proposed strategy allows dealing with Internet access measurements including both noise and outlier data, as well as to group users into well-separated clusters. (iii) *A novel distance measure* has been proposed to drive the DBSCAN algorithm into correctly identifying noise and outliers. (iv) *Performance of all users are analyzed together.* Differently from previous works, MiND analyzes the statistical distribution of Internet access performance experienced by all users together to correctly model a comprehensive view of the network.

The proposed methodology has been thoroughly evaluated on real and heterogeneous datasets including data belonging to a single ISP in different geographical areas and data collected in different time intervals. Data have been collected by means of Neubot (Nexa Center, 2016), an open source software research project supported by the Nexa Center for Internet and Society of the Politecnico di Torino in Italy. The datasets used in this paper and the source code for the cluster analysis are published online in a public repository on Github (Servetti, A., 2016) together with a short description of the work. Experimental results demonstrate that MiND correctly identifies homogeneous and cohesive groups of users receiving a similar download speed. The MiND findings allow enhancing user awareness of the Internet access service that they really receive and spotting anomalous network behavior that may require further analysis and investigation.

The paper is organized as follows. Section 2 summarizes the related work in the area concerning both Internet access measurement collection and their analysis. The proposed mining framework is described in Section 3 illustrating in details the algorithmic choices and how to optimally tune their parameters. A thoroughly experimental evaluation is presented in Section 4 showing the effectiveness and robustness of the proposed algorithms. Section 5 discusses the MiND findings and their possible exploitation from both the academic and managerial perspectives. Finally, Section 6 draws conclusions and discusses further developments.

2. Related work

Measurement of Internet access network speed is a popular field of investigation for multiple parties ranging from academia to governments (C. Duffy Marsan, 2013). On one hand, Internet regulators are actively supporting large scale network measurements to foster up to date and widespread mon-

itoring of Internet access services in order to be able to compare broadband providers and to frame better policies to regulate them. On the other hand, users are becoming eager and eager to know how their Internet connection behaves both with respect to other ISPs and, inside the same ISP, compared to other users. For instance, in the case of Ookla Speed Wave (Ookla, 2016), group of users can compare results against each other and compete for achievements such as highest download speed and lowest latency badges.

Most of the available platforms for broadband measurements are targeted on collecting and analyzing aggregate information for interested organizations. Such platforms are based on *spot measurements* of the different access networks that ISPs offer as broadband connection to Internet users. Thus, a relatively small number of probe points on each provider are used by these platforms to make assumption on the ISP quality of service (e.g., average speed, percentage of satisfied users, etc.). These implementations are generally based on highly reliable measurements that are performed by dedicated hardware that must be delivered to the user and installed on his network. This class of platforms include: the RIPE Atlas project (RIPE, 2016), that was started in late 2010 and that now counts 6,926 installed probes; the SamKnows project (SamKnows, 2016), that since 2008 is collaborating with governments and industries to benchmark broadband performance in several countries (e.g., the September 2013 campaign counted data from 6,398 subscribers (Federal Communication Commission, 2014)); the Bismark project, that at the end of 2014 counts 420 devices deployed, largely in developing countries (Project BISmark, 2016).

Other platforms are oriented to informing users, as opposed to institutions and governments, about their specific Internet access service. Thus, to easily reach every potentially interested user, they are based on software applications that can be installed on different operating systems or used directly from the web browser. These implementations can characterize each single user connection with a very deep level of detail. In this scenario it is possible to distinguish between two schemes: user activated probes and periodic probes. The first scheme includes Ookla Speedtest.net and NDT where each test must be run directly by the user. Even if they are very popular (Ookla counts 5 million measurements each day and NDT 3 million measurements per month), both suffer from a relatively small number of *measurements per user* that clearly limits the ability to statistically characterize the behavior of the user's connection. For example, NDT completely lacks the concept of "user" because results are identified only by the client IP address which may

be reused by several users over time. The second scheme includes Neubot, that provides a smaller number of measurements, nearly 1 million per month, but that can periodically perform the measurements multiple times per day for the same user, thus allowing to sample and characterize each connection on a per user basis. For every installation, Neubot stores an unique user identifier that can be used to match each measurement with that user even if other parameters change, most notably the IP address that is dynamically assigned, and frequently modified, by the ISP.

Up to now, Neubot is the only active service that collects and publishes periodic measurements of users' Internet access services. Therefore, it is currently the only one that allows to characterize and compare the profile of the Internet connection of different users. However, an in-depth analysis is needed to transform such large volume of data into knowledge and ultimately, actions.

Many research efforts have been devoted to analyzing network traffic data through unsupervised data mining techniques, because they do not require previous knowledge of the application domain (e.g., a labeled traffic trace (Katris & Daskalaki, 2015)). Authors in (Apiletti et al., 2009) proposed to discover correlations at different abstraction levels among network data packet headers, while authors in (Apiletti et al., 2013) proposed a cloud-based service to extract frequent correlations on passive traffic measurement collections. Clustering algorithms represent a widely-used exploratory technique to identify groups of similar network flows. They have been exploited to address different and interesting network traffic issues such as deriving node topological information (Baralis et al., 2013), automatically identifying classes of traffic (Apiletti et al., 2016), unveiling YouTube CDN changes (Giordano et al., 2015), predicting the throughput on a network (Maia et al., 2010), characterizing P2P traffic (Chung et al., 2010), grouping network flows by application type (Carmo et al., 2008), identifying users' role based on their behaviors through the analysis of social features (Zhu et al., 2011), and supporting network management (Carvalho et al., 2016). This work instead proposes a two-level clustering strategy jointly with a new distance measure to analyze Internet access performance of different ISP users with the aim to discover groups of users according to the Internet access that they really received.

3. The MiND methodology

MiND aims at analyzing Internet access measurements to identify groups of users that receive a similar Internet access service. This system relies on innovative techniques to deal with data characterized by an inherent sparseness with the final aim to correctly identify cohesive and well-separated groups of users. Specifically, MiND identifies group of users by analyzing the statistical Internet access behavior of DSL subscribers as reported by Neubot on the basis of periodic measurements, not just on the basis of a single network measurement as done by similar projects such as NDT. The proposed methodology aims to answer to the following questions: (i) are there similar statistical behaviors of users that are sufficiently similar to be clustered in a single group? (ii) from the point of view of the single user, is the behavior of a given user similar to the one of a group of other users in the considered set of data?

The possibility to identify such clusters is interesting for both the final users and the network operator itself. In fact, it is reasonable to assume that users belonging to the same cluster have a similar experience to many others in the group, therefore they behave "normally". On the contrary, other users that cannot be easily classified into a cluster might experience issues with their Internet access, therefore this can be interesting to know for both the

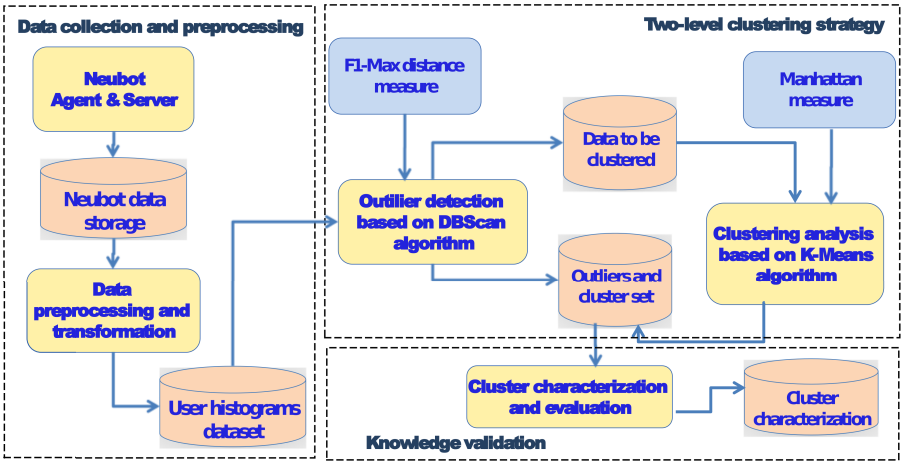


Figure 1: The MiND components

user and the operator. The users might be informed that their behavior is somehow anomalous (instead of assuming that, maybe, it is the same for all the others), and the operator can use such information to check if unexpected network behaviors are taking place for the users.

Figure 1 shows the main components of the MiND architecture as well as interactions between such components. The first activity of MiND is the *data collection* phase, which is performed through Neubot. Neubot data are typically characterized by an inherent sparseness and variable distribution over time because Neubot is installed as a background service but the user can decide to deactivate it at some times, for instance for privacy reasons. Moreover, depending on the situation, users may not be always connected to the Internet, so the periodicity of the measurements may strongly vary. The variability in data distribution increases with data volume, thus increasing the complexity of mining such data.

When dealing with inherently sparse distributions, it is recommended to apply a suitable *data transformation* prior to data analysis (T. Pang-Ning et al., 2006). Thus, an ad-hoc data transformation models the data on a different space, from which hidden and more interesting knowledge can be extracted. MiND exploits a frequency histogram technique to compactly model the Internet access service received by each user. Then, the actual service experienced is modeled through a histogram for each user. Given this new set of data, a *clustering analysis* can discover groups of users with similar Internet accesses over the time. To this aim we propose a *two-level clustering strategy* (as shown in Fig. 1) that first deals with noise and outlier data and then groups users into well-separated and homogeneous clusters. The proposed strategy is based on the DBSCAN (M. Ester et al., 1996) and K-means (J. A. Hartigan & M. A. Wong, 1979) algorithms. Furthermore, a *novel distance measure* has been proposed so that the DBSCAN algorithm can correctly identify noise and outliers in the set of user-histograms. Finally, MiND also includes a *knowledge validation* component (see Fig. 1) to evaluate the quality of the identified groups of users. This component is based on quality indexes (e.g., SSE (T. Pang-Ning et al., 2006) and Silhouette (Rousseeuw, 1987)) that can evaluate the goodness of the identified clusters. Algorithmic details of the MiND methodology are discussed in Appendix A.

uuid	subnet	asnum	speed (bytes/s)
2b37de0c-5f49-4446-8b8f-3b2dad14fb61	50.128.0.0/9	7922	3588959
72740cc4-b665-475c-acad-29e3f176af91	79.10.0.0/15	3269	489583

Table 1: An example of the data extracted from the Neubot speedtest database.

3.1. Data collection and preprocessing

Internet access measurements for the MiND framework are collected by the Neubot project, then they are retrieved from the Neubot Repository data storage and preprocessed to both extract only the data of interest for the analysis and add some additional field useful for the clustering process.

3.1.1. The Neubot Internet access measurements

Neubot is an open-source tool voluntarily installed by users on their computer to periodically monitor the characteristics of their Internet connection. More details of the Neubot collected information can be found in Appendix A.1. Neubot runs as a background service, periodically performing a set of transmission tests between the user's computer and a Neubot Server hosted in the M-LAB (M-lab, 2016) network. In this study we analyze the measurements of the *speedtest* test that measures the download bandwidth in terms of the application-level throughput (Kurose, 2013).

The *speedtest* test of the Neubot project collects a variety of features for each measurement performed by each final user. Among them, MiND exploits the Unique User Identifier (`uuid`) and the measured download speed (`speed`)[1]. We enrich these two features with the Autonomous System Number (`asnum`) and the IP address subnet (`subnet`) from which the measurement was performed to correctly group measurements performed at the same user location. Pairs of `uuid` and `subnet`, denoted as the *user* in the rest of this study, are used as the unique identifier of each set of measurements. Table 1 shows an example of the selected features.

3.1.2. Data transformation

The data transformation component of MiND aims at pre-processing the data to effectively support the subsequent data analysis by extracting interesting knowledge items.

[1]The measured download speed only inform on the quality of the Internet connection service experienced by the user at "that given moment" and it can not represent a measure of the user's Internet access service speed.

Since in the Neubot architecture each transmission test constitute a single record in the database, the monitored measurements for each user are spread over many records. As a consequence, it is unfeasible to direct apply clustering algorithms to such data because a user's Internet access characterization is spread over many records. Therefore, an ad-hoc data transformation process is needed to model the data in a different space to support more interesting analyses. Specifically, MiND tailors a given dataset storing collected measurements (e.g., download speed measurements for many users over the time) to a new space model based on user-histograms. To highlight the relevance of Internet access in terms of bandwidth, MiND represents all collected measurements (download speed measurements repeated over time) belonging to a single user through a frequency histogram. Thus, each user-histogram compactly represents the distribution of all measurements belonging to a single user. To create the histograms, first an expert of Internet access technology decides a suitable division of the typical available access bandwidth into intervals (bins), as detailed in Section 4.2. Then, each user-histogram is built to report, for each bin, the normalized number of times that a given download speed, measured for a given user, falls into the bin. Given this new set of data (one record for each user), a cluster analysis can be performed to discover groups of users with similar Internet access. Thus, homogeneous user groups will contain similar histograms, i.e., with similar shapes in terms of position of peaks and low values.

Figure 2 shows the effect of the data transformation process for two users. The results of each download speed measurement over time are shown on the left plots while the corresponding two histograms are shown on the right plots. Measures are collected over a period of one year with an average of about three measurements per day. The histogram bin width is 0.5 Mb/s.

3.2. Two-level clustering strategy

MiND adopts a two-level clustering approach to analyze Internet access behavior of users over a long time span. First, noise and outliers are identified in the complete dataset to exclude users that received an anomalous Internet access service from the subsequent step. Then, a suitable clustering algorithm is applied to identify groups of cohesive users with homogeneous statistical behavior.

Figure 2 shows an example of the expected *normal behavior* for two users of different ISPs with different ADSL speed. The plots on the left represent the download speed measured by Neubot in each single test over a time span

of one year. Users are expected to experience a download speed close to the maximum DSL connection bandwidth they are paying for. As reported in Figure 2 the connection speed has an upper speed limit because all the measurements are below a threshold, that is close to 7 Mb/s for the user shown in the top part and close to 5 Mb/s for the user shown in the bottom part. The two plots in the right part of Figure 2 are the histograms that represent the distribution of the download speed measurements over bins of 0.5 Mb/s. The upper limit is also visible in the histograms, but here we also notice that the distribution peak is very close to that limit and that as we move away from that value the probability of measuring that speed decreases. Given this typical download speed distributions, which is in accordance with (Paxson, 1994), homogeneous users that belong to the same ISP should aggregate their download speed measurements around few values according to the available broadband plans of the ISP (e.g., 5, 10, 25, 50, Mb/s). On the contrary, users with an anomalous Internet access would experience a different distribution of the download speed measurements, i.e. a higher variability of their connection speed over a wider range of speed values (below the upper limit).

MiND adopts the DBSCAN algorithm (Ester et al., 1996) for the first level of analysis and the K-means algorithm (Juang & Rabiner, 1990) for the second one. More details about both DBSCAN and K-means algorithms

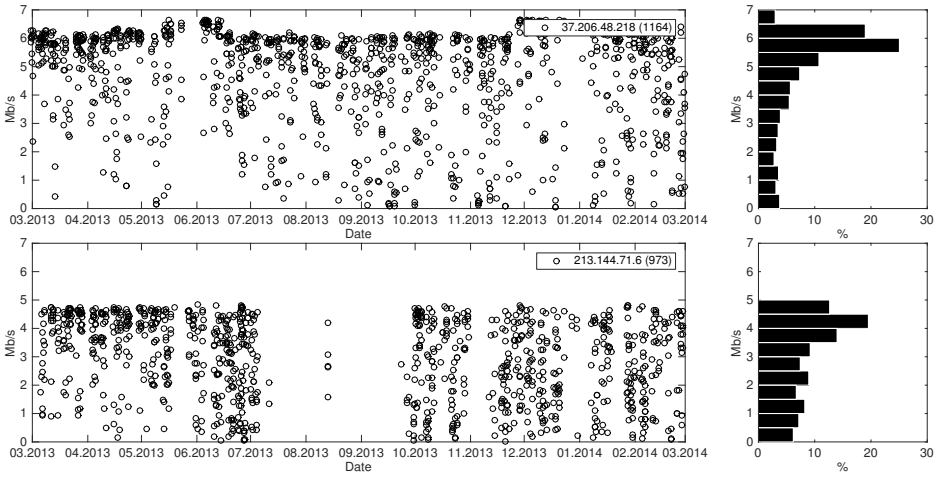


Figure 2: Download speed measured with Neubot speedtest over a year (left) and speed histogram (right) for two users of different ISPs.

can be found in Appendix A.2. A key operation to perform a good analysis is to effectively measure the similarity among data objects. Similarity is usually measured according to a notion of distance in a measurement space describing the object features, as detailed in the next section.

3.2.1. Distance measure

MiND integrates (i) a new distance measure, named F1-Max, able to identify outlier and noisy user profiles, and (ii) the Manhattan distance to correctly discover groups of homogeneous user profiles based on their histograms. Traditional distance measures, such as Euclidean, Overlap and Jaccard distances (Ackermann et al., 2010), are not suited to compute the distance between two user-histograms due to the following two issues. (i) User-histogram bins (dimensions) are not orthogonal, (ii) peak values in the user-histogram introduce a distortion in the calculation of the distances. The relevance of the above issues increases when dealing with noisy data (i.e., datasets including some anomalous user-histograms) as real datasets. These issues have been addressed by our newly defined F1-Max aimed at measuring the distance between two user-histograms.

The F1-Max distance measure is a cross-bin distance measure that contains additional terms that also compare non-corresponding bins within a given "bin distance". The main idea is to reduce the sensitivity of the algorithm to the position of bin boundaries so that users with small shifts of the measured connection speed, e.g. one bin shift, may still be considered homogeneous. On the contrary, users with larger shifts will still appear as distant points. Thus, F1-Max overcomes both the non-orthogonality issue and the distortion introduced by peak values.

For the non-orthogonality issue, let us consider an n-dimensional hyper-space where dimensions are ordered and not all independent of each other. In this hyper-space all dimensions will be orthogonal to all dimensions except to the closer ones. For example, if the ordered dimensions are: x_1, x_2, x_3, x_4, x_5, x_6, x_7, then dimension x_3 will be non-orthogonal to dimension x_2 and x_4 (case (i)) and to dimension x_1 and x_5 (case (ii)) while x_3 will be orthogonal (independent) to all the other dimensions (case (iii)), thus the latter are not considered. To remove the non-orthogonal relationship between two dimensions (cases (i) and (ii)) the corresponding distance can be properly weighted. Specifically, we use w_1 and $w_2 < w_1$ to weight the distances related to case (i) and case (ii) respectively.

To minimize the distortion in the calculation of the distances due to the

peak values, histograms have been preprocessed before distance computation. The top six values of each histogram have been normalized with the following criterion: the highest value has been replaced by the average of the highest value of each histogram, the second highest value by the average of the second highest value of each histogram, and so on. Since the contribution to the distance tends to zero by considering lower top values of each histogram, we neglect such contributions.

The F1-Max measure between two histograms $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ is computed as follows:

$$\begin{aligned} \text{F1-Max}(x, y) = & \sum_{i=1}^{n} dist(x_i, y_i) \\ & + w_1 \sum_{i=1}^{n} \frac{1}{2|k_1|} \sum_{j \in k_1} (dist(x_i, y_j) + dist(x_j, y_i)) \\ & + w_2 \sum_{i=1}^{n} \frac{1}{2|k_2|} \sum_{j \in k_2} (dist(x_i, y_j) + dist(x_j, y_i)) \end{aligned} \quad (1)$$

where $k_1 = \{i-1, i+1\}$, $k_2 = \{i-2, i+2\}$ and $dist(x_a, y_b)$ is defined as:

$$dist(x_a, y_b) = |x_a - y_b| \cdot \max(x_a, y_b) \quad (2)$$

where the distance between two user histogram bins (x_a, y_b) is based on the Manhattan distance emphasizing the differences between the bins with a weight equal to the maximum between the two.

Unlike the Euclidean distance, the Manhattan distance considers as equal all the diagonals of all the rectangles with the same perimeter. Thus, it computes the distance between two objects measured along axes at right angles, which is equal to the distance that would be traveled to get from one data point to the other if a grid-like path is followed (T. Pang-Ning et al., 2006). The traditional Manhattan distance between two user-histograms is the sum of the differences of their corresponding bin values (i.e., normalized number of times that a given download speed range is measured by the user).

The Manhattan distance formula between two user-histograms $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$ is:

$$Manhattan(X, Y) = \sum_{i=1}^{n} |x_i - y_j| \quad (3)$$

where n is the number of histogram bins, and x_i and y_i are the values of the i-th bin, at user-histograms X and Y respectively.

The first level clustering of MiND exploits the DBSCAN algorithm (Ester et al., 1996) jointly with the F1-Max distance to correctly identify outlier data. For the second level clustering the K-means (Juang & Rabiner, 1990) is exploited using the Manhattan distance. As discussed in Section 4.3 this configuration is able to correctly identify anomalous user-histograms as well as cohesive and well-separated groups of user-histograms.

3.3. Knowledge validation

MiND integrated two objective measures (i.e., Silhouette and SSE) to evaluate the quality of the clustering results and to perform a sensitivity analysis on the parameters used as input for the clustering algorithms. Specifically the *Silhouette* index (Rousseeuw, 1987) measures both intra-cluster cohesion and inter-cluster separation by evaluating the appropriateness of the assignment of a data object to a cluster rather than to another. The higher the index, the better the clustering. The *Sum of Squared Error (SSE)* (T. Pang-Ning et al., 2006), instead, evaluates the cluster cohesion for center-based clustering techniques, i.e., K-means. The smaller the index, the better the quality of discovered clusters. More details about the equations of both the Silhouette and the SSE are reported in Appendix A.3.

4. Experimental results

To validate the effectiveness of the MiND framework, we addressed four issues: (i) MiND performance (Section 4.3), (ii) time stability analysis (Section 4.4), (iii) MiND sensitivity and robustness (Section 4.5) to parameter setting, (iv) MiND robustness to distance measure selection (Section 4.5.3).

A large set of experiments have been performed on two real datasets (Section 4.1) collected by Neubot. Before the application of the proposed two-level clustering strategy, MiND employs a data transformation as discussed in Section 4.2.

The open source RapidMiner toolkit (Rapid Miner, 2016) has been used for the cluster analysis. The new distance measure has been developed in Java and it is used by the clustering algorithms available in RapidMiner.

Both the datasets and the RapidMiner code used in this section are available on Github (Servetti, A., 2016).

Table 2: Datasets collected by Neubot from July 2012 to June 2014. Statistics include lower quartile (lq), median (med), and upper quartile (uq) measured in Mb/s.

ID	Provider	Users	Measurements	Statistics (lq, med, uq)
D1	Telecom Italia	3659	206884	2.72 5.64 8.00
D2	Comcast	1568	778052	5.91 15.47 23.75

4.1. Datasets

We considered two real datasets collected by means of Neubot. We recall that, among the network measurement platforms, Neubot is the only tool that allows to aggregate the collected measurements by user and then build a histogram of its Internet access speed. This section describes the main characteristics of the considered datasets and the corresponding data transformation applied on them before performing the two-level clustering strategy.

Table 2 describes the two Neubot datasets used to evaluate MiND in terms of time-span, different number of users and measurements. Each dataset includes a subset of the Neubot users in the same Internet Service Provider (ISP), as identified by the Autonomous System Number (ASNUM) to which the user IP address belongs. *D1* is the dataset including measurements performed by users of the largest Italian ISP, Telecom Italia S.p.a. (AS3269). The *D2* dataset includes measurements performed by users of Comcast Cable Communications Inc. in the United States (AS7922). The latter is the ISP with the largest number of measurements collected by Neubot.

4.2. Data transformation

The data transformation component of MiND represents, by means of a frequency histogram, all collected measurements (download speed measurements repeated over time) related to a single user. Thus, each user-histogram compactly describes the statistical behavior of the *download speed* measurements recorded by the same user in a given subnet.

The data transformation component discards histograms with less than 50 measurements, because they are not deemed statistically significant. For the *D1* dataset, the download speed values are included in a very short range (i.e., 0–20 Mb/s), thus we set a uniform bin widths of 1 Mbit/s.

For the Comcast dataset, instead, the variability of the download speed values is wider (i.e., 0–120 Mb/s), thus using a uniform distribution for the

Table 3: Non uniform bin widths for download speed histograms. Download speed upper boundary of each bin (ds) is measured in Mb/s.

bin #	1	2	3	4	5	6	7	8	9	10	11
ds	1	2	3	4	5	6	7.1	8.5	10.3	12.6	15.6
bin #	12	13	14	15	16	17	18	19	20		
ds	19.4	24.0	30.0	37.6	47.3	59.6	75.2	95.0	120.1		

histogram bin widths is not appropriate. Therefore, we use bin widths that follows a logarithmic scale so that the higher the measured speed the larger the bin width. The logarithmic function is defined in Eq. (4), where the download speed (ds) is expressed in Mb/s.

$$bin(ds) = \begin{cases} \lceil ds \rceil & ds \leq 6Mb/s \\ \lceil ln\left[(ds - 1.81)^{4.19}\right] \rceil & ds > 6Mb/s \end{cases} \quad (4)$$

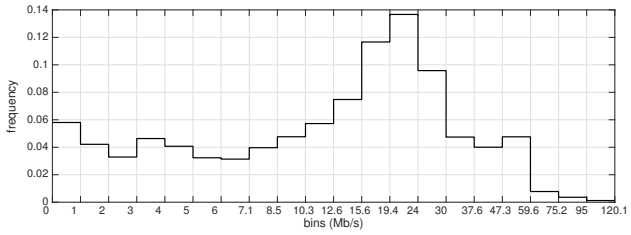
The corresponding bin boundaries are those listed in Table 3. Figure 3(a) shows the histogram representing the statistical distribution of all the download speed measurements in *D2*. Most of the probes report speeds between 5 and 60 Mb/s that represent the vast majority of the Comcast users.

4.3. M*i*ND *performance*

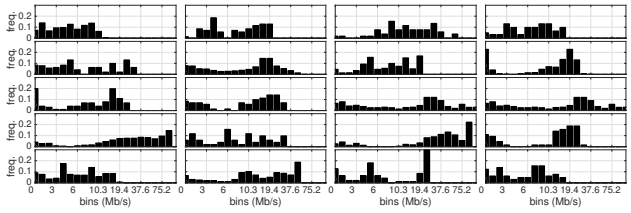
In this section we evaluated the MiND performance to show the effectiveness of the proposed framework in (i) discovering a set of clusters that correctly represent users whose connection is homogeneous in terms of statistical behavior and (ii) identifying also users that do not fit well in those clusters because their connection behaves differently from the others. To this aim, a two-level clustering strategy has been proposed. The first-level clustering addresses the issue (ii) (Section 4.3.1), while the second-level the issue (i) (Section 4.3.2). The Comcast trace (*D2* in Table 2) is discussed as a representative dataset.

4.3.1. First-level clustering

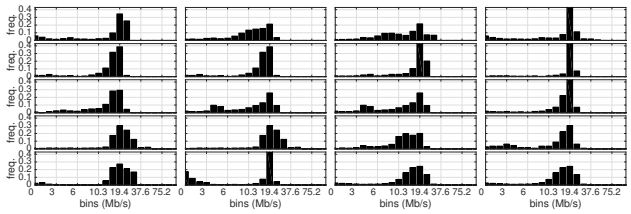
In MiND the DBSCAN clustering algorithm, coupled with the new distance measure (F1-Max) defined in Section 3.2.1, is first used to identify noise and user-histogram outliers. The DBSCAN parameters Eps and $MinPts$ are set to 0.25 and 4 respectively, as the result of the sensitivity analysis detailed in Section 4.5.



(a) Global histogram of the measured download speeds



(b) User-histograms identified as noise



(c) User-histograms identified as non-noise

Figure 3: Histograms from the Comcast dataset.

The DBSCAN algorithm identifies as outliers/noise a set of 37 user-histograms (out of the 796 histograms of the users with more than 50 measurements) characterized by an anomalous download speed pattern. Figure 3(b) shows some user-histograms in the outlier cluster and Fig. 3(c) shows some user-histograms in a non-noise cluster.

We observe that user-histograms considered as outliers in Fig. 3(b) have multiple peaks (bi/tri-modal distribution) or present a "plateau" with many small peaks very close together that resemble a quasi-uniform distribution. These are two characteristics that may identify anomalous Internet access services or the presence of a source of noise in the Neubot measurements.

On the contrary, users with a regular access service have most of the

download speed measurements close to their maximum download speed and few or no occurrences of speeds above that threshold. In fact, it is not possible that all the probes result in the maximum speed value, but hopefully they should report a speed not too lower than that value. The more the distance from that value, i.e., the provider advertised speed, the less the quality of the service offered. At the same time, the measured speed should not vary too much otherwise it may be a symptom of an anomalous connection that is not able to provide the expected service with the required reliability.

Figure 4 shows a 3D representation of all user-histogram in the outlier cluster and all user-histogram in three homogeneous clusters (the ones with the highest number of user-histograms) identified by DBSCAN. Figure 4 visualizes the dispersion of the user-histograms inside a cluster using a special representation. Specifically, each user-histogram is shown as a row of the image where the frequency value in each bin is represented by a grayscale, white corresponds to 0 and black to 1. A visual analysis of the cluster representations shows that the dark regions (i.e., the user-histogram peaks) of the top left noise cluster are, as expected, widely dispersed among the bins. On the contrary, the representations of the other three clusters indicate a concentration of the dark regions in few bins. Thus, as documented in the following sections, the MiND framework appears to be able to correctly identify anomalous Internet access services.

A similar methodology has also been applied to the Telecom Italia dataset (D1). The DBSCAN identified as outliers noise a set of 79 user histograms characterized by an anomalous download speed pattern (out of the 909 histograms of the users with more than 30 measurements). The DBSCAN parameters Eps and $MinPts$ were almost identical to the ones used in the Comcast dataset analysis.

4.3.2. Second-level clustering

The second-level clustering in MiND is performed in the dataset after outliers have been identified and removed. This step exploits the K-means algorithm and the Manhattan distance measure (see Section 4.5.2 for parameter settings). Figure 5 (left) shows the three clusters identified by the algorithm on the Comcast dataset. Specifically, both the average histogram and the corresponding size are reported for each cluster. The histogram of the first cluster presents three peaks in the range between 0 and 20 Mb/s, but gradually decreases with the increase of the download speed value. The other two clusters are very concentrated around bin 13 (19.4–24 Mb/s) and

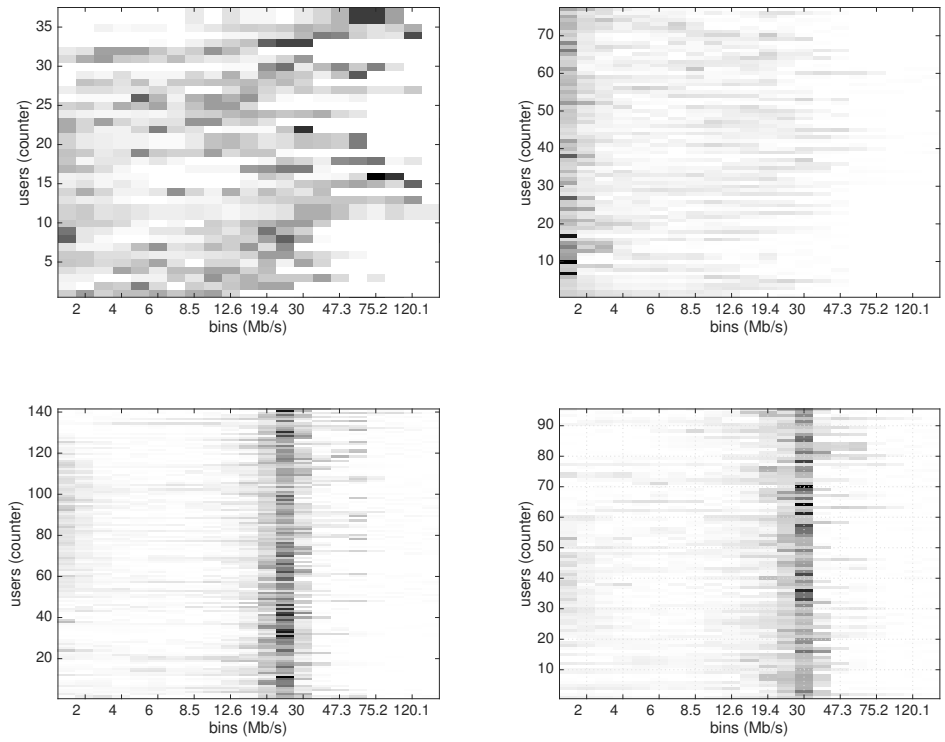


Figure 4: Download speed histogram colormaps for the clusters identified by DBSCAN in the Comcast dataset. Each row of the colormap shows a user download speed histogram where the bin frequency value is represented with a grayscale. The top left plot represents the users' histograms assigned to the noise cluster.

bin 17 (47.3–59.6 Mb/s) that may correspond to DSL services of 25 and 50 Mb/s. Note, however, that the first cluster is the one that, alone, includes almost half of the analyzed records (311 out of 759). These results are in line with Internet access services provided by Comcast and on average subscribed by customers.

Figure 5 (right) shows the download speed distribution for each cluster in Fig. 5 (left). All box plots are compact showing that the speed distribution variance of each cluster is limited, and so the compactness of the cluster.

Figure 6 (left) shows both the average histogram and the corresponding size for each cluster identified on the Italian dataset[2] (*D1* in Table 2).

[2]Telecom Italia is the former monopolist that build the network physical infrastructure in Italy.

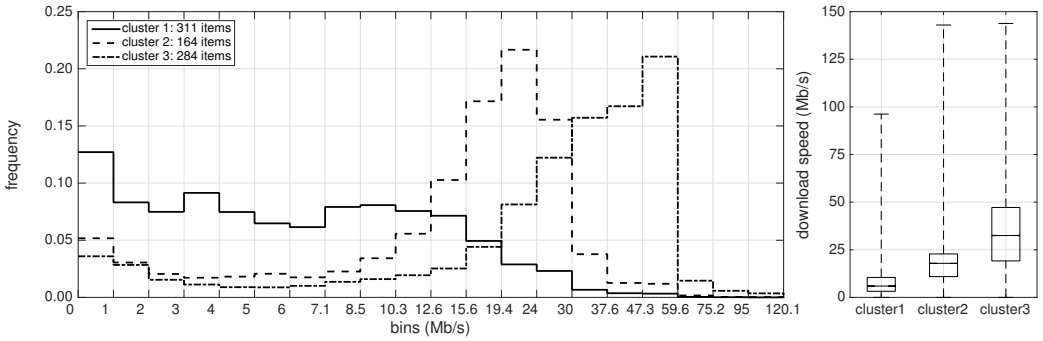


Figure 5: Second-level clustering on Comcast (US) dataset. Average of user-histograms (left) and boxplot (right) per cluster.

These results are also interesting because they exhibit a strong relation with the services offered by the Telecom Italia ISP. The typical speeds for this provider are in fact 7 Mbit/s and 4 Mbit/s (though this is not advertised, but appears to be limited by the ISP when the SNR of the physical channel is not very good). For the higher speed cluster (peak around 10 Mbit/s), there is currently no offer around 10 Mbit/s, but there is one around 20 Mbit/s. Therefore 10 Mbit/s might stem from the impossibility to take full advantage of the network physical speed for a different reason (e.g., network congestion, other concurrent download activity performed by the clients).

Figure 6 (right) shows the speed distribution for each cluster in Figure 6 (left). All the three box plots have the last quartile values close to typical speeds for the Italian provider (e.g., 4 Mbit/s, 7 Mbit/s). Furthermore, all box plots are very compact, as the inter-quartile range (IQR) values are very close to the median ones, proving that the speed distribution variance of each cluster is very limited, thus the compactness of the cluster. These results support the effectiveness of MiND in discovering compact and interesting groups of users based on the Internet access services that they really received.

4.4. Time stability analysis

We performed a time stability analysis in order to further assess the usefulness of MiND and its effectiveness in discovering interesting clusters of users. Specifically, we compared the results obtained using datasets covering different time periods related to the same provider. The Telecom Italia dataset (*D1*) is discussed as representative example. To this aim, we compared results from the first time frame (from Jul 1, 2012 to Jun 30, 2013,

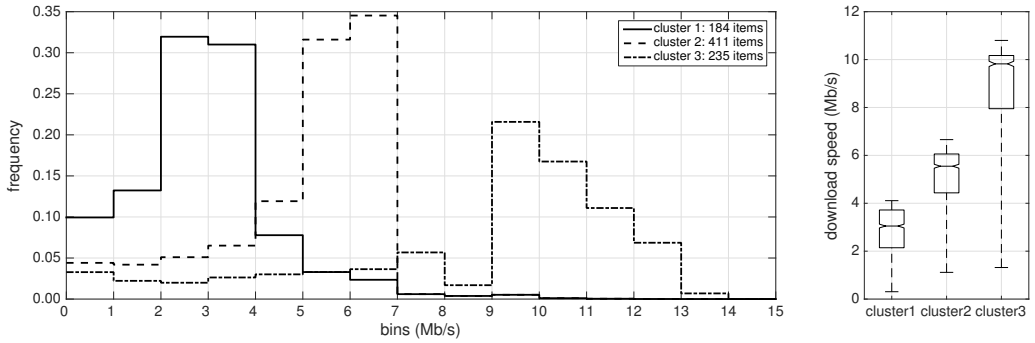


Figure 6: Second-level clustering on Telecom Italia dataset. Average of user-histograms (left) and boxplot (right) per cluster.

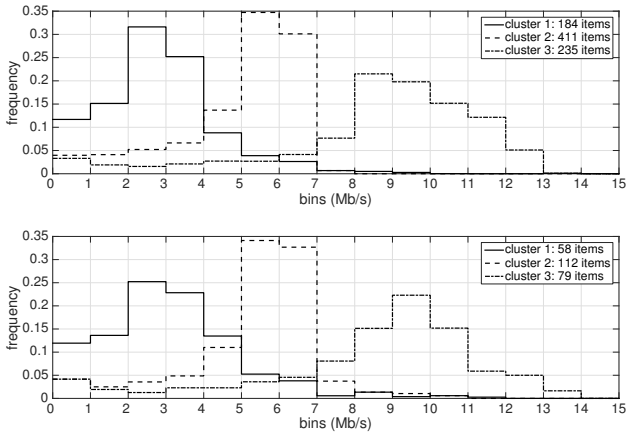
Table 4: Improvement for Telecom Italia download speed from 2012–13 to 2014, measured at significant values of the three clusters.

	Cluster 1	Cluster 2	Cluster 3
First quartile	+8%	+9%	+10%
Median	+13%	+8%	+5%
Third quartile	-3%	+8%	+4%

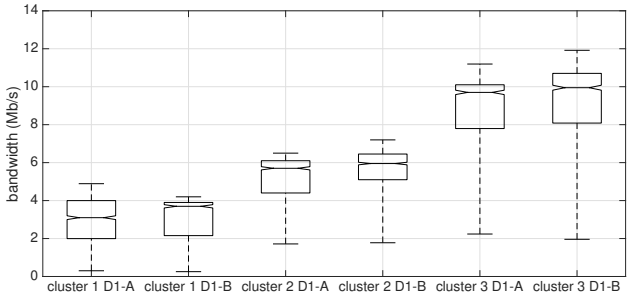
named *D1-A*) of the *D1* dataset with the one collected from Jan 1, 2014 to Jun 30, 2014, named *D1-B*.

We may assume that any difference is due to some variations in the services offered to the users. Figure 7(a) shows both the average histogram and the corresponding size for each cluster identified on *D1-A* and *D1-B* respectively. Reported results show that the behaviors are similar but it is possible to notice a slight improvement in the download speed for the year 2014 (*D1-B*). This is also visible in Figure 7(b) and quantified in Table 4. Specifically, Figure 7(b) reports the speed distribution for each cluster in Figure 7(a) to compare the two sets of discovered clusters in *D1-A* and *D1-B*. Table 4 shows the percentage improvement, in term of received bandwidth, for each cluster discovered in *D1-A* with respect to the ones in *D1-B*.

On average, all three user groups feature a connection better in 2014 (*D1-B*) than the one obtained in 2012-2013 (*D1-A*). As shown in Table 4 all the index positions of the box plot (the first quartile, the median, and the third quartile) undergo a substantial increase (percentage of increased bandwidth) with the exception of the third quartile in cluster 1. The increase of the median value ranges from 5% for users in the higher speed cluster



(a) Average user-histograms



(b) Boxplot of the clustering

Figure 7: Clustering of Telecom Italia data collected in the time period Jul 1, 2012 – Jun 30, 2013 (a)(top), and Jan 1, 2014 – Jun 30, 2014 (a)(bottom), and boxplot comparison (b) showing the improvement for all groups.

to 13% for those in the lower speed cluster. The largest group of users (cluster 2) has a nearly constant increase for all index positions. Therefore, it seems reasonable to conclude that, over time, the download speed service, as measured by Neubot has, in general, improved from 2012-2013 (*D1-A*) to 2014 (*D1-B*).

4.5. Algorithm sensitivity and robustness

We analyzed the robustness of the clustering quality to parameter settings. MiND parameter setting addressed the following issues. (i) *Reduce data fragmentation.* Since clusters should summarize Internet access behav-

ior, we avoid the generation of a large number of clusters including few users. (ii) Exhibit good silhouette values, showing that they include *subsets of correlated users*. (iii) *Avoid many unclustered users*, by limiting the number of users labeled as outliers.

To address the above issues, a large set of experiments have been run to find the optimal input parameter settings, using, when available, tools to optimize algorithm performance (e.g., K-dist graph (Ankerst et al., 1999) for the DBSCAN algorithm, as shown in Appendix A.2) or objective measures to evaluate the discovered clustering structures as discussed in Section 3.2.1. The latter has been exploited to find the best value for the K parameter of the K-means algorithm (see Section 4.5.2). The Comcast trace (*D2* in Table 2) is discussed as representative dataset since it includes a large variety of services and users.

4.5.1. Setting DBSCAN parameters: K-dist graph

The DBSCAN algorithm exploits two input parameters: $MinPts$ and Eps. For DBSCAN parameter setting, we rely on the k-dist graph (T. Pang-Ning et al., 2006) plotting. It shows, for each data object, the distance to its k^{th} nearest neighbor. The F1-Max measure is used for distance computation. On the x-axis data objects are sorted by the distance to the k^{th} nearest neighbor, while on the y-axis distances to the k^{th} nearest neighbor are reported.

When the distance with the k^{th} nearest neighbor is small, the object will be labeled as core or border point and included in a cluster. Instead, when the distance is high the object will be labeled as outlier and noise point and not included in a cluster.

Figure 8 shows the k-dist graph for the Comcast dataset. k corresponds to the $MinPts$ parameter, while the y-axis contains possible values of the Eps parameter. Since $MinPts$ indicates the minimum number of points in a cluster, we set it at 4 (and 8) and we analyzed the impact of Eps values on the clustering result.

By intercepting the curve in Fig. 8 at a given Eps value on the y-axis, the corresponding p_x value on the x-axis partitions data objects into the following two subsets. Points placed on the left side of p_x are labeled as core points, and those on the right side of p_x as noise/outlier or border points.

Usually, the Eps value is selected where a rather sharp change (T. Pang-Ning et al., 2006) appears in the curve. For our cluster analysis, we intercept the curve at the sharp slope change, i.e., Eps in the range [0.225 – 0.325].

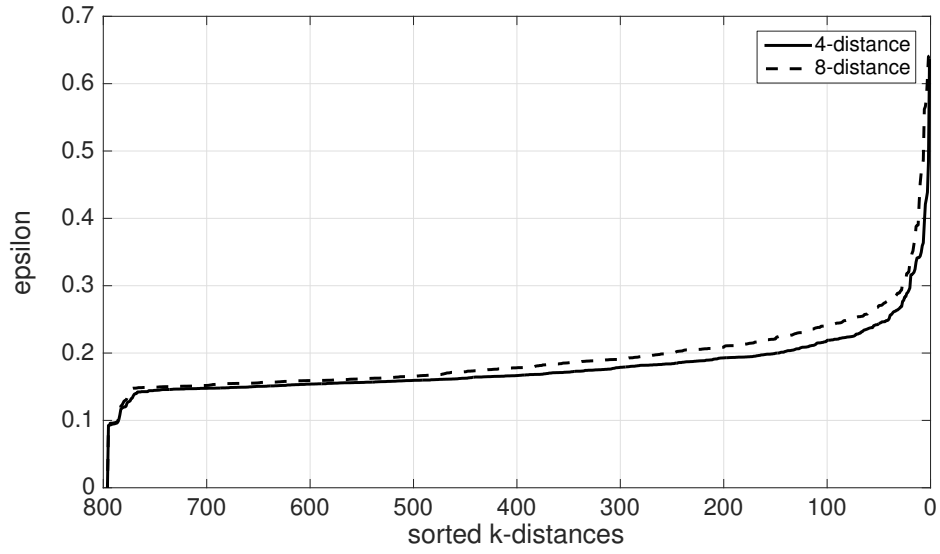


Figure 8: k-dist graph for the Comcast dataset using F1-Max measure with weighting parameter $w_1 = 0.3$.

4.5.2. Setting the K-means parameter

The K-means algorithm requires as input parameter the number of clusters (K), which is in general very difficult to define, given the wide range in which it may vary.

To address this issue we analyzed two traditional quality indexes (i.e., Sum of Squared Error and Silhouette). The smaller the SSE, the better the quality of discovered clusters. However, as the number of cluster increases, the SSE decreases because smaller and more cohesive clusters are identified. In contrast, in many real applications the actual number of interesting clusters is usually small. Thus, we need to identify a good trade-off between the number of clusters and their significance.

To measure both intra-cluster cohesion and inter-cluster separation we exploited the Silhouette index to evaluate the appropriateness of the assignment of a user histogram to a cluster rather than to another. Negative Silhouette values represent wrong user histogram assignments, while positive values good user assignments. Given a clustering result, its Silhouette value is the average weighted Silhouette value on all user histograms assigned to each cluster. The higher the Silhouette, the better the quality of discovered clusters.

Many runs of the K-means algorithm have been carried out with varying values of K, and for each run, the cluster set is evaluated by computing both the SSE and the Silhouette. Figures 10(a) and 10(b) show the SSE values and the average Silhouette values, respectively, computed on different clusters sets by varying the K parameter. By analyzing the SSE index, good values for K are in the range from 3 to 4, by considering the average Silhouette, the best value for K is 3. Thus, in MiND we set $K = 3$ for the second level clustering algorithm based on K-means.

4.5.3. The distance measure selection

In MiND two distance measures have been exploited to correctly identify interesting groups of user histograms. Here, we analyzed the robustness of the clustering quality achieved by MiND to select the distance measure. Since MiND exploits a two-level clustering strategy, we analyzed the impact of the distance measure on each level separately. MiND uses the DBSCAN algorithm as a first level clustering. Thus, we first analyzed the robustness of the clustering quality yielded by DBSCAN by varying the distance measure (F1-Max, Manhattan). To evaluate the cluster quality we computed the average silhouette by considering all user histograms (without noise) clustered by DBSCAN (group #1), and the corresponding average silhouette by considering all user histograms labeled as outliers (group #2). The better clustering quality corresponds to a high silhouette value for group #1 and low silhouette value for group #2. Table 5 reports both the average silhouette for groups #1 and #2 by also varying the weight w_1 in the F1-Max measure. Different values for Eps parameter in the range $[0.225 - 0.325]$ (identified through the K-dist plot, see Section 4.5.1) have been considered. F1-Max yielded a better cluster quality than the Manhattan distance measure. Among the considered values for w_1 and Eps, the best trade-off between the maximization of the average silhouette (group #1) and the minimization of the average silhouette (group #2) is yielded for $w_1 = 0.3$ and $Eps = 0.25$. Thus, for the first level clustering MiND exploits the DBSCAN algorithm with $Eps = 0.25$, $MinPts = 4$, and the F1-Max as the distance measure. A visual comparison of average silhouette for group #1 is also shown in Fig. 9.

We also analyzed the robustness of the clustering quality yielded by K-means as a second-level algorithm in MiND by varying the distance measure (F1-Max, Manhattan). Figure 10(a) shows the SSE by varying the distance measure. Different values for the K parameter of K-means have been considered. The Manhattan measure here yielded a better clustering quality than

Table 5: Average silhouette values for DBSCAN clustering with different distance measures and varying *Eps* range, *MinPts* is fixed to 4. Distance F1-Maxis evaluated with three different values of the weighting factor (w_1).

		Epsilon					
	w_1	0.225	0.250	0.275	0.300	0.325	
F1Max	0.2	0.497	0.425	0.276	0.260	0.180	w/o noise (group #1)
		-0.554	-0.552	-0.543	-0.594	-0.381	noise only (group #2)
F1Max	0.3	0.505	0.521	0.520	0.519	0.355	w/o noise (group #1)
		-0.562	-0.531	-0.524	-0.542	-0.518	noise only (group #2)
F1Max	0.4	0.167	0.451	0.431	0.426	0.433	w/o noise (group #1)
		-0.497	-0.506	-0.540	-0.521	-0.501	noise only (group #2)
Manhattan		-0.203	-0.285	-0.335	-0.246	-0.246	w/o noise (group #1)
		-0.606	-0.588	-0.509	-0.437	-0.437	noise only (group #2)

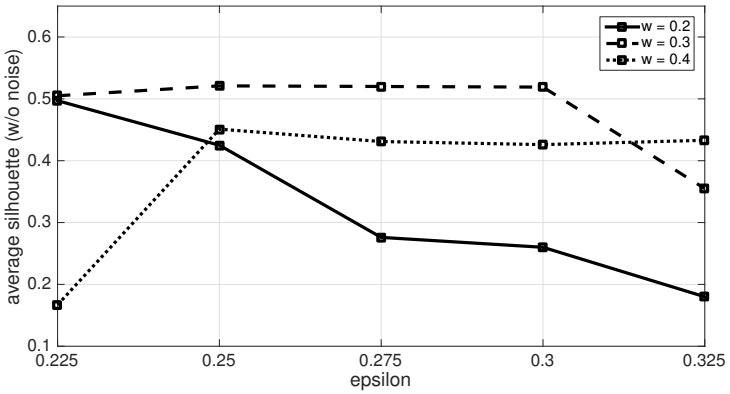
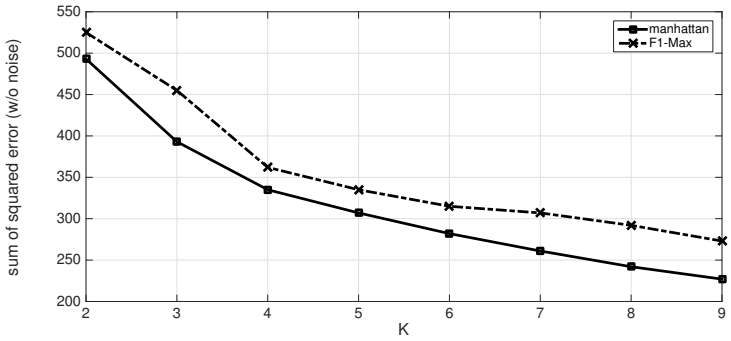


Figure 9: Average silhouette for DBSCAN clustering with different distance measures and varying *Eps* range. *MinPts* is fixed to 4. Distance F1-Maxis evaluated with three different values of the weighting factor (w_1).

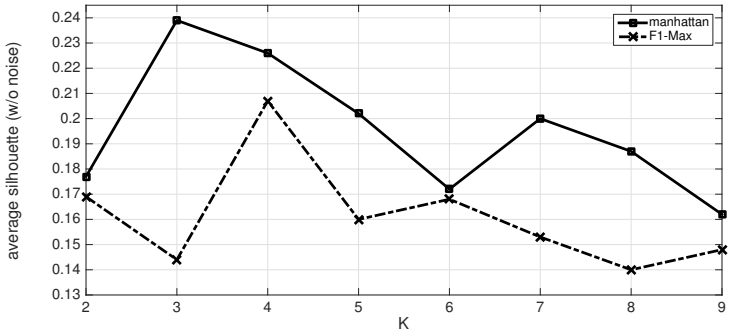
F1-Max, thus in MiND we exploited it to drive the second-level clustering. Figure 10(b) shows the average silhouette.

4.6. Additional case studies

To further validate the MiND methodology, we report the results obtained on two new datasets, collected in a more recent time period (June 2014 – May 2016), for other major ISPs: MCI/Verizon in the US (AS701), and Wind in Italy (AS1267). Table 6 shows the key metrics of the two additional datasets, similarly to the ones already shown for D1 and D2. MiND identifies (i) few users receiving anomalous services, i.e., 50 (12%) for D3 and 26 (6.6%) for D4 and (ii) three groups of users receiving a usual service,



(a) Sum of squared error as a function of the number of clusters



(b) Average silhouette as a function of the number of clusters

Figure 10: Comparison of SSE (a) and average Silhouette (b) for the Manhattan or the F1-Max distance measure by varying the K parameter for K-means clustering.

which are shown in Figure 11. Note that, despite the change of the time period and the ISPs, good clustering performance can be achieved similarly to the case of D1 and D2, with well separated download speed peaks for the different clusters.

5. Discussion

This section aims to discuss the previous MiND findings and how they can be exploited from both the academic and the managerial perspective. MiND analyzes the download speed measurements over time of all users for a given ISP. MiND discovers (i) *groups of users* with a similar and usual Internet access behavior and (ii) *a few users* with somehow anomalous service.

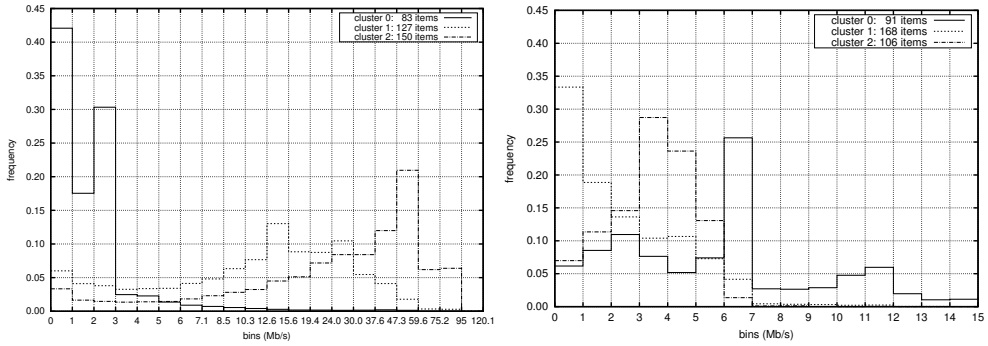


Figure 11: Average of user-histograms on additional datasets: MCI/Verizon, US (left) and Wind, Italy (right), for the period June 2014 – May 2016.

Table 6: Datasets collected by Neubot from June 2014 to May 2016. Statistics include lower quartile (lq), median (med), and upper quartile (uq) measured in Mb/s.

ID	Provider	Users	Measurements	Statistics (lq, med, uq)
D3	MCI/Verizon	775	670839	6.68 19.35 41.45
D4	Wind	1119	198817	1.44 3.29 5.06

To this aim, the statistical distribution of the download speed measurements (i.e., user-histogram) is analyzed for each user.

MiND has been thoroughly validated on two main case studies (Comcast and Telecom Italia) and summary results have been presented for other two (MCI/Verizon and Wind). A set of 37 user-histograms out of the 796 histograms (i.e., 4.65%) and a set of 79 user histograms out of the 909 histograms (8.7%) have been identified as anomalous behaviors on the Comcast and Telecom Italia datasets respectively. These user-histograms (see Fig. 3) have different peaks that sometimes resemble a quasi-uniform distribution. Instead, the set of users whose connection is homogeneous in terms of statistical behavior identified by MiND on both datasets are in line with Internet access services offered by both providers and on average subscribed by customers. Thus, we can conclude that these users receive an Internet access service in line with the subscribed one.

Differently from other widespread projects for Internet access performance monitoring such as NDT (NDT, 2016), MiND analyzes the statistical distribution of Internet access performance experienced by tracking unique users *over time*. Moreover, the whole set of measurement is fed into MiND so that it can have a comprehensive view of the network. This new analytics

perspective allows to get different types of insights with respect to the other works. In fact, the MiND findings can provide feedback to both users and ISPs. The large majority of other projects, instead, typically aim at either providing a direct, immediate, but limited feedback to the user on the basis of a single measurement (e.g., Ookla Speedtest.net), or at collecting large sets of data but without information that can match the data with each single user (e.g., NDT only collects IP addresses). Therefore, such large sets can only be useful to ISPs for a general network performance overview but they cannot provide feedback to the ISP about the experience of single users.

From the managerial perspective, MiND findings could be exploited to inform both users and ISPs about the correspondence between the subscribed services and the received/provided ones. In practice, the MiND analysis can be run periodically by both the ISP or the users (e.g., using the publicly available Neubot data), so that both parties can be informed about the presence of anomalous behaviors or, conversely, reassured about the absence of any anomaly. From the user side, users receiving a disservice have a tool that can help them to objectively demonstrate the issues they are experiencing. From the ISP side, the tool can be used to isolate unexpected network behaviors for further analysis and investigation, as well as potentially preventing user complaints. In fact, in presence of repeated anomalous behaviors over time, an ISP could schedule ad-hoc maintenance sessions to improve the reliability of the provided services. The ISP could also use the tool to show, with an objective third-party instrument, that a large share (if not all) of their users are receiving a service in line with the one they subscribed for.

From the academic perspective, MiND findings demonstrate the ability of the proposed methodology to correctly analyze large collection of measurements distributed over time and automatically discover similar statistical behavior together with anomalous ones. There is a large variety of events that can be monitored over time, with a large set of admissible values (as in the case of the domain of the download speed values) thus resulting in datasets with inherent sparseness and variable distribution which are typically difficult to handle. We believe that the MiND methodology can be easily ported also to different application domains (e.g., smart city applications, medical applications) where the collected data have properties similar to the ones of the datasets considered in this study. For instance, consider a smart urban environment where sensor networks are deployed to continuously monitor environmental parameters. In general, each sensor measures a single phenomenon (e.g., humidity, temperature, traffic) over time and per-

forms a measure every roughly few minutes. The collected measurements may have large domains. A possible relation between this work and the example of the smart urban environment could be to map each sensor onto a Neubot probe, then analyze the collected data as done in this work, i.e., modeling the statistical distribution of collected measurements as histograms and applying the same techniques. In this application scenario MiND could be exploited to identify groups of sensors with similar statistical behavior together with a few sensor with anomalous behaviors which can potentially indicate anomalous situations in a given part of the urban environment.

Finally, there is still room for improvement of the MiND methodology. In fact, one of its main drawbacks is that it requires a minimum number of measurements to model the statistical distribution of the received Internet access service through user-histogram. We are currently investigating novel strategies to model users with a limited number of measurements.

6. Conclusions

This work presented MiND, an innovative cluster-based system aimed at automatic and efficient characterization of groups of users with a similar Internet access behavior. To characterize Internet access parameters, publicly available download speed measurements provided by the Neubot platform have been exploited and analyzed in-depth. The rationale behind the MiND framework is presented and discussed in details investigating which data transformation, clustering algorithms, and distance measure provide the best performance for the specific characteristics of the collected data. We believe that the promising results open a set of new possibilities for Internet users to enhance their awareness of the Internet access service they really receive. Using MiND, for instance, it would be possible to automatically perform activities such as alerting users about unusual behaviors or automatically spot behaviors that may be interesting for further analysis and investigation. Future extensions of this work include the development of cloud-based services for the analysis of Internet access parameters and the exploitation of different frequency methods (e.g., TF-IDF method (T. Pang-Ning et al., 2006)) to model user-histograms. Furthermore, the exploitation of the MiND methodology in different application domains can also be investigated as exemplified in the discussion section.

Acknowledgments

The authors are grateful to Alessio Bongiorno that run several preliminary experiments leading to the results presented in this work.

References

Ackermann, M. R., Blömer, J., & Sohler, C. (2010). Clustering for metric and nonmetric distance measures. *ACM Transactions on Algorithms*, *6*, 59.

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). Optics: Ordering points to identify the clustering structure. In *Proceedings of ACM SIGMOD International Conference on Management of Data* (pp. 49–60). New York, NY, USA: ACM.

Apiletti, D., Baralis, E., Cerquitelli, T., Chiusano, S., & Grimaudo, L. (2013). SeaRum: A cloud-based service for association rule mining. In *11th IEEE International Symposium on Parallel and Distributed Processing with Applications, ISPA-13* (pp. 1283–1290).

Apiletti, D., Baralis, E., Cerquitelli, T., & D'Elia, V. (2009). Characterizing network traffic by means of the NetMine framework. *Computer Networks*, *53*, 774–789.

Apiletti, D., Baralis, E., Cerquitelli, T., Garza, P., & Venturini, L. (2016). SaFe-NeC: a Scalable and Flexible system for Network data Characterization. In *IEEE/IFIP Network Operations and Management Symposium, Istanbul, Turkey, 25-29 APRIL 2016*.

Baralis, E., Bianco, A., Cerquitelli, T., Chiaraviglio, L., & Mellia, M. (2013). NetCluster: A clustering-based framework to analyze internet passive measurements data. *Computer Networks*, *57*, 3300–3315.

C. Duffy Marsan (2013). IAB plenary explores challenges of network performance measurements. *IETF Journal*, *8*, 7–8.

Carmo, M. F. F. d., Maia, J. E. B., Siqueira, G. et al. (2008). An internet traffic classification methodology based on statistical discriminators. In *Network Operations and Management Symposium, 2008. NOMS 2008. IEEE* (pp. 907–910). IEEE.

Carvalho, L. F., Jr., S. B., de Souza Mendes, L., & Jr., M. L. P. (2016). Unsupervised learning clustering and self-organized agents applied to help network management. *Expert Systems with Applications*, *54*, 29 – 47.

Cerquitelli, T., Chiusano, S., & Xiao, X. (2016). Exploiting clustering algorithms in a multiple-level fashion: A comparative study in the medical care scenario. *Expert Syst. Appl.*, *55*, 297–312.

Chehreghani, M. H., Abolhassani, H., & Chehreghani, M. H. (2009). Density link-based methods for clustering web pages. *Decision Support Systems*, *47*, 374–382.

Chung, J. Y., Park, B., Won, Y. J., Strassner, J., & Hong, J. W. (2010). An effective similarity metric for application traffic classification. In *Network Operations and Management Symposium (NOMS), 2010 IEEE* (pp. 286–292). IEEE.

Combes, C., & Azema, J. (2013). Clustering using principal component analysis applied to autonomy-disability of elderly people. *Decision Support Systems*, *55*, 578–586.

van Dam, J.-W., & van de Velden, M. (2015). Online profiling and clustering of Facebook users. *Decision Support Systems*, *70*, 60–72.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Knowledge Discovery and Data Mining (KDD)* (pp. 226–231).

Federal Communication Commission (2014). Measuring Broadband America Report — Technical Appendix. http://data.fcc.gov/download/measuring-broadband-america/2014/Technical-Appendix-fixed-2014.pdf Last access March 2016.

Giordano, D., Traverso, S., Grimaudo, L., Mellia, M., Baralis, E., Tongaonkar, A., & Saha, S. (2015). YouLighter: An unsupervised methodology to unveil Youtube CDN changes. *CoRR*, *abs/1503.05426*.

J. A. Hartigan, & M. A. Wong (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*, 100–108.

Juang, B.-H., & Rabiner, L. (1990). The segmental k-means algorithm for estimating parameters of hidden markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *38*, 1639–1641.

Katris, C., & Daskalaki, S. (2015). Comparing forecasting approaches for internet traffic. *Expert Systems with Applications*, *42*, 8172 – 8183.

Kurose, J. F. (2013). *Computer networking: a top-down approach*. Pearson Education.

M. Ester, H.-P. Kriegel, J. Sander, & X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 226–231).

M-lab (2016). M-lab. https://console.developers.google.com/storage/m-lab/ Last access March 2016.

Maia, J. E. B. et al. (2010). Network traffic prediction using pca and k-means. In *Network Operations and Management Symposium (NOMS), 2010 IEEE* (pp. 938–941). IEEE.

NDT (2016). Network Diagnostic Test. http://www.measurementlab.net/tools/ndt Last access March 2016.

Nexa Center (2016). Neubot Project. http://neubot.org Last access March 2016.

Ookla (2016). Ookla Speedtest. http://www.speedtest.net/ Last access March 2016.

Paxson, V. (1994). Empirically derived analytic models of wide-area TCP connections. *IEEE/ACM Transactions on Networking*, *2*, 316–336.

Project BISmark (2016). Network Dashboard. http://networkdashboard.org Last access March 2016.

Rapid Miner (2016). The Rapid Miner project for machine learning. http://rapid-i.com/ Last Access March 2016.

RIPE (2016). RIPE network coordination centre. https://atlas.ripe.net Last access March 2016.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, (pp. 53–65).

SamKnows (2016). SamKnows. https://www.samknows.com/ Last access March 2016.

Servetti, A. (2016). MIND Repository. https://github.com/servetti-polito/mind-compnet-15 Last access March 2016.

T. Pang-Ning, M. Steinbach, & V. Kumar (2006). *Introduction to Data Mining*. Addison-Wesley.

Zhu, T., Wang, B., Wu, B., & Zhu, C. (2011). Role defining using behavior-based clustering in telecommunication network. *Expert Systems with Applications*, *38*, 3902 – 3908.