Wireless Personal Communications (2021) 120:1791–1819
https://doi.org/10.1007/s11277-021-08535-8

# Large-Scale 802.11 Wireless Networks Data Analysis Based on Graph Clustering

**Germán Capdehourat[1,2] · Paola Bermolen[2] · Marcelo Fiori[2] · Nicolás Frevenza[2,3] · Federico Larroca[2] · Gastón Morales[2] · Claudina Rattaro[2] · Gianina Zunino[1]**

Accepted: 13 April 2021 / Published online: 21 April 2021


**Abstract**
This paper analyzes a large-scale dataset of real-world Wi-Fi operating networks, collected from more than 9,000 access points (APs) for 1 year. The APs are distributed among more than 1,200 educational centers in the context of a nation-wide one-to-one computing program, being most of them primary and secondary schools. The data corresponds to RSSI measurements between APs used to build the conflict graphs for each school Wi-Fi network. We propose a simple embedding for the Wi-Fi network conflict graphs based on classical graph features, which proves to be useful to analyze the behavior of the wireless networks, showing a high discrimination power among the different school networks. Moreover, we discuss some practical applications of the embedding. In particular, it enables to study the Wi-Fi network dynamics at each school, analyzing the conflict graphs temporal variations through clustering techniques. The presented methodology allows us to successfully separate the most stable scenarios from those with more significant variability, which therefore require more technical resources to optimize the network. Besides, we also compared the behaviour of the Wi-Fi networks of the different schools, which enable us to reuse the optimal configuration found for one school in all those sites that have similar conflict graph patterns.

**Keywords** IEEE 802.11 · Wi-Fi · RSSI · Conflict graph

## 1 Introduction

In recent years, the evolution of wireless communication technologies has substantially changed our lives. In this sense, the IEEE 802.11 standard's progress has had a huge impact on people's connectivity habits. Today, Wi-Fi carries more than half of the

✉ Germán Capdehourat
gcapdehourat@ceibal.edu.uy

[1] Plan Ceibal, Avda. Italia 6201, Edificio Los Ceibos, 11500 Montevideo, Uruguay

[2] Facultad de Ingeniería, Universidad de la República, Julio Herrera y Reissig 565, 11300 Montevideo, Uruguay

[3] Facultad de Ciencias Económicas y Administración, Universidad de la República, Gonzalo Ramírez 1926, 11200 Montevideo, Uruguay

Internet's traffic, according to the Wi-Fi Alliance [1]. No matter where we are, whether it is in the office, at home, in the shopping center, in the hospital or the bar, having this wireless access has become essential (and even more so with the pandemic due to the Covid-19). This increasing relevance also implies more significant challenges for any network infrastructure, such as availability and quality of experience.

Another aspect that has been increasing, linked to wireless networks' deployment, is the possibility of collecting massive amounts of data from the operational networks. Nowadays, every major Wi-Fi vendor offers solutions that enable a broad set of possible ways to collect a huge amount of data about the network operation. Many possibilities are available for each network layer, ranging from radio frequency (RF) and air utilization metrics at the physical layer, up to traffic analysis and user's device characteristics at the application layer. However, regardless of what marketing people may say [2], the available commercial solutions have not yet fully exploited this data. Although wireless networks (as many other areas) are also surfing the wave of the current hype of artificial intelligence and machine learning [3–5], much effort is still needed to convert the vast amounts of data available into useful information.

We believe this work is a step towards that goal, taking advantage of the abundance of data from Wi-Fi networks, and presenting novel ways to extract useful information from the data. We use a large dataset collected in an educational context, from the nation-wide Wi-Fi schools networks in Uruguay. A full school year with hourly RSSI measurements of more than 9,000 access points (APs) distributed in more than 1,200 educational centers [6], compose the raw data used to build the conflict graphs that model the interference between APs in the real-world operational Wi-Fi networks. Then, we base on network analysis tools and machine learning to study the graphs' temporal and spatial variations. In particular, we focus the study on the 2.4 GHz band, where less spectrum is available with only three 20 MHz non-overlapping channels. Thus, the interference between nodes is much more relevant in this frequency band, for which our dataset has more than 70,000 neighbouring AP pairs.

Solid mathematical models have been developed for graph analysis, which correspond to the area known as *network science* [7], recently popularized by social networks data analytics. In this work, we address whether it is possible to find a suitable embedding for the conflict graph of a Wi-Fi network in order to characterize the interference dynamics. We propose an efficient embedding based on standard *features*, such as centrality measures and other well-known graphs properties detailed in Sect. 4. We show that the proposed graph embedding is simple and useful to analyze the wireless network behaviour, obtaining a high level of discrimination power among the different school's Wi-Fi networks. Furthermore, the proposed graph embedding based on classical features enables us to study the graphs variations through clustering algorithms. Several clustering methods were compared to choose the most appropriate one to analyze the collected data from the different wireless networks.

Next, we analyze the graph time series' temporal variations for each school Wi-Fi network. In this case, the results were relatively stable for most of the schools, with more than 90% with only two or three *temporal modes*, associated with the number of clusters found on each graph time series. The number of temporal modes and the differences observed between them allows us to identify those schools that deserve more attention and require more technical resources. In addition, we look for common patterns between the graphs of the different schools, doing the clustering with the aggregated graph features of each time series. This methodology allows finding schools with similar Wi-Fi network conditions, making it possible to replicate the optimal solutions and configurations found in one site to

all other similar schools. This procedure avoids doing field surveys and RF analysis at each particular location, a costly task in terms of human resources.

The remainder of the article is structured as follows. The next section reviews previous works that have used conflict graph models in 802.11-based networks. In Sect. 3 the dataset collection process is described, while Sect. 4 introduces the selected graph features and presents an exploratory analysis to show how the data looks. In Sect. 5 we evaluate the discrimination power of the graph features selected, while Sect. 6 focuses on the selection of the clustering algorithm. Finally, the further analysis carried out with the complete dataset is presented in Sects. 7, and 8 concludes the paper with the main insights and the next steps to continue with this research line.

## 2 Related Work

Random graphs models have been extensively used in the past to analyze and design wireless networks [8]. Next, we highlight previous works that have been done using conflict graphs in 802.11-based networks. The conflict graph is a popular tool to model the interference among different wireless links. It is commonly used for contention-based access networks (i.e., like 802.11 networks), to indicate which wireless links interfere with each other, and hence, cannot be active simultaneously. It is also useful to model situations when links do operate simultaneously, in which case the conflict model (typically a weighted graph for this case) indicates how much one link affects the other when both operate simultaneously.

Since the advent of wireless multi-hop networks, such as Wireless Mesh Networks (WMNs) [9] and Vehicular Ad-Hoc Networks (VANETs) [10], one of the problems which received the most attention was channel assignment [11]. Thus, the different interfering models considered are typically expressed through the corresponding conflict graph [12]. For example, WMNs with nodes with multiple radios were considered [13, 14], while distributed conflict graphs at each network interface were proposed [15] for cognitive networks. More recently, [16] additionally included co-location interference in the model. Weighted conflict graphs were also introduced to take into account partially overlapping channels [17] or to add the link rate information to the graph [18]. We can also find in the recent literature machine learning algorithms to solve channel assignment [19], including methods based on deep reinforcement learning [20].

Many other problems have been addressed based on conflict graphs models, such as routing in multirate WMNs [21] and the evaluation of routing metrics [22], energy-efficient rate adaptation [23], VoIP performance [24] and association optimization [25] in WLANs, admission control, bandwidth sharing and QoS guarantees [26–29]. This interference model [30] also proved to be useful for performance analysis and capacity estimation [31, 32], enabling to calculate the throughput of each link as a function of the WLANs conflict graph [33].

Most of today high-end WLAN solutions are based on network controllers, which among many tasks, are responsible for Radio Resource Management (RRM). The central controllers typically collect data from all the APs, which are used to construct and periodically update a conflict graph, where the APs constitute the graph's nodes. Based on the conflict graph, the controller jointly generates optimal channel assignments and power control levels for the APs [34]. Besides, this feature allowed us to collect large amounts of data, enabling to analyze the graph evolution during the time and compare the graphs

between different networks. To the best of our knowledge, no previous works have dealt with large datasets' analysis, corresponding to the conflict graphs of real-world Wi-Fi networks and its evolution over time. In [35], the authors addressed problems related to the construction of the conflict graphs using measurement-calibrated propagation models in order to avoid the need for detailed signal measurements (see later in Sect. 3 the detail about what Cisco does [36]). On the other hand, a different approach was presented in [37], using a conflict graph embedding, which represents the wireless nodes with low-dimensional vectors while preserving their conflict relationships. Although these works are based on network measurements, they are focused on the conflict graph building process, but not in analyzing the resulting graphs.

With the increasing availability of data from real-world operational networks, we believe that more research efforts should be dedicated to finding proper ways to exploit this data for network optimization and management purposes. Wi-Fi solution providers have been looking for analytical tools to include in their products, but no significant improvements have been presented. Our work focuses directly on narrowing this gap, seeking to combine machine learning and network analysis tools to extract valuable information from a large dataset of interference measurements in Wi-Fi networks.

## 3 Dataset Description and Conflict Graph Construction

In order to build the dataset, RSSI measurements were gathered from the different school Wi-Fi networks managed by Plan Ceibal [38]. This organization is in charge of the implementation of a nation-wide one-to-one computing program in Uruguay. Thus, one of its most relevant responsibilities is to provide Wi-Fi Internet access at all educational centers throughout the country. This makes it one of the nation's largest Internet providers, with a total number of devices connecting to Plan Ceibal's networks comparable to the number of mobile network operators' subscribers.

It is important to note that most of the Plan Ceibal's Wi-Fi networks correspond to indoor scenarios at public primary and secondary schools. These educational centers are located in an enormous variety of buildings, ranging from centennial constructions with several stories and hundreds of students to small rural schools with just a few tens. This fact is illustrated in Fig. 1, which shows the histogram of the number of APs per building (i.e., per school Wi-Fi network). As we can see, each building is typically covered on average by 5 or 6 APs, although approximately 20% of the buildings required more than 10 APs.

Most of the Plan Ceibal's networks are based on high-end Wi-Fi solutions, which allows relatively complete and continuous monitoring of the network's state, which we leverage in this study. In particular, the vast majority of access points currently installed correspond to a Cisco solution, managed by two Cisco Flex 7500 Wireless LAN Controllers (WLCs), each of them supporting up to 6000 APs. The complete list of AP models and their most important parameters are detailed in Table 1. All of them are configured with 20 MHz channels in 2.4 GHz and 40 MHz in the 5 GHz band (using only non-overlapping channels in both).

One of the things managed by the WLCs manage are radio resources, by means of the Cisco RRM's proprietary algorithms. For this purpose, each AP in the network periodically sends a so-called NDP (Neighbor Discovery Protocol) packet on every channel and band possible. The NDP packets are broadcast messages transmitted at the maximum allowed

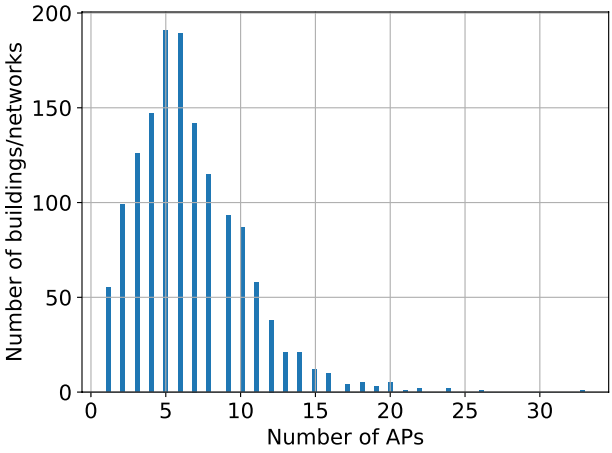



**Fig. 1** Number of buildings/networks with a given number of APs

**Table 1** AP models and how many are present in the complete network

| AP model (Cisco Aironet) | Number of APs | 802.11 Standard (2.4 GHz/5 GHz) |
|---|---|---|
| 2702I | 5098 | 802.11n/802.11ac-Wave 1 |
| 1702I | 2681 | 802.11n/802.11ac-Wave 1 |
| 1832I | 862 | 802.11n/802.11ac-Wave 2 |
| 2602I | 759 | 802.11n/802.11n |

power for the channel/band, at the lowest supported data rate and using a single radio chain (meaning no beamforming is applied in their transmission). By default, an NDP packet is sent over all channels every 180 s. The AP goes off-channel roughly every 16 seconds to send an NDP packet over the 11 channels in the 2.4 GHz band, and every 8 s for the 22 channels in the 5 GHz band. All received NDP packets, and the corresponding RSSI (expressed in dBm and with a resolution of 1 dBm) and channel are forwarded to the WLC. These values are averaged by the WLC over 15 minutes (the so-called pruning interval), corresponding to 5 measurements per neighbor.

A data collection system was set up, which sends SNMP queries to the WLCs so as to gather the information corresponding to all the APs. In particular, RSSI measurements were collected, which indicate how each AP *hears* all other APs in the network, for both frequency bands. The timescale was chosen in order to minimize the effect on the operational network. A typical sequence of RSSI measurements is shown in Fig. 2, where the different time-series correspond to how one particular AP is seen from all its neighbours in the 2.4 GHz band. The period is restricted to the school year (from March to December), and missing data (e.g. note the small gap in mid-April) is mostly due to holidays (when the equipment might be turned off at schools). Some missing measurements could also be due to problems in the connection between the data collection system and the WLCs. This data is not stored by either the APs or the WLCs, so they should be gathered live or lost.

The resulting dataset corresponds to RSSI measurements, which indicate how strongly each AP listens to its neighbors over the different schools' Wi-Fi networks. We will concentrate only on the 2.4 GHz band measurements, the more crowded one (not only because

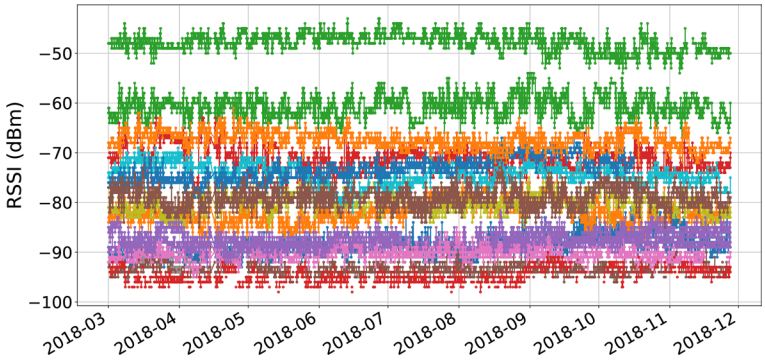



**Fig. 2** A typical sequence of RSSI measurements showing how one AP is heard from all its neighbours in the 2.4 GHz

of 802.11 but also by Bluetooth, Zigbee, etc.) and with less spectrum available, which makes the conflict graph much more relevant than for 5 GHz. So, in order to construct the corresponding conflict graphs, we consider each AP as a node. Then, each node will have an incoming edge for each AP that it hears, and the corresponding weight will be the power received (i.e., the RSSI values). This way, we end up with one directed graph for each school at each timestamp.

It is worth to mention that we are not taking into account the operation channel of each AP to build the conflict graphs. That is to say, the graphs considered this way indicate the potential interference they may produce to each other if they are on the same channel. As we will see in the next section, we will take into account an RSSI threshold value to prune the graphs, removing all those edges that are below a certain power level, assuming that these APs do not affect each other.

## 4 Graph Features and Exploratory Data Analysis

In order to study the graphs variations, a notion of distance (or similarity) between graphs is needed. Several approaches are available for this purpose, including distances based on global structures, which are strongly related to the notion of graph isomorphism [39], and relaxations of these ideas, such as graph-kernel based techniques, where the idea is that two vertices are considered similar if their neighborhoods are similar (e.g. *Weisfeiler-Lehman* algorithm [40]). A simpler alternative is to base the distance between graphs in a predefined set of *features*. Characterizing the graph by these features allows us to consider the similarity between graphs as the corresponding similarity between the feature sets. For instance, two graphs are considered similar if the euclidean distance in $\mathbb{R}^d$ for a vector of $d$ chosen features is small. In the next subsection, we introduce the different metrics extracted from the graph structure to analyze the Wi-Fi networks conflict graphs. Finally, we end this section presenting an exploratory analysis to describe the resulting graph features dataset.

### 4.1 Graphs Features Computation

While many graph features were initially inspired by social network analysis, nowadays they are widely used in several different areas (see chapter 7 of [41] for a detailed

presentation). In this study, we will rely on them to analyze the conflict graphs resulting from Wi-Fi networks data, since they allow a clear interpretation of the results based on their definitions. To do so, we have to compute for each graph the corresponding set of selected features, for which we used the NetworkX Python library [42]. Different preprocessing steps were carried out before computing the features with the corresponding NetworkX functions. First, we discard all the edges with an RSSI value below a minimum of -80 dBm, considered as the Clear Channel Assessment (CCA) threshold for APs [43]. In this way, edges between all the neighbouring APs that do not affect each other are removed. Then, we compute the following features:

*Number of edges* It corresponds to the number of links (i.e., ordered AP pairs) with RSSI values above $-$ 80 dBm. When symmetry holds (which is not always the case [6]), there would be two edges for each pair of APs that see each other above the threshold.

*In-degree centrality* It is the average over the incoming degree centrality values for each node, which is defined as the average over all the incoming edges of the node. These values correspond for each AP to the average RSSI of all the APs heard above -80 dBm. It is worth to note that this average over an entire graph is almost the same for the incoming degree than for the outgoing degree, so we select only the incoming values in this case. In addition to the mean value, we also include the standard deviation of the incoming degrees in each graph's feature vector.

Another preprocessing was needed for the rest of the features, as the edge weights, in that case, must resemble a distance measure. Thus, we have to convert the RSSI values (which are in dBm) into a distance metric between APs, which is done using Eq. (1), where $w_{ij}$ corresponds to the resulting distance between $\text{AP}_i$ and $\text{AP}_j$, i.e., the $i, j$ entry of the graph adjacency matrix.

$$w_{ij} = 10^{-\text{RSSI}_{ij}/10}. \tag{1}$$

Then, we compute the remaining selected features:

*Betweenness centrality* It is the average value of the shortest-path betweenness centrality for each node. The betweenness centrality of a node $i$ is defined to be the number of shortest paths that pass through $i$,

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}},$$

where $n_{st}^i$ takes the value 1 if the node $i$ belongs to the shortest path between $s$ and $t$ and $g_{st}$ is the number of such paths. If both $n_{st}^i$ and $g_{st}$ are zero we set $n_{st}^i / g_{st} = 0$. The shortest path in a weighted graph is defined as the path that minimizes the sum of the edge weights over the path.

*Page rank* It is the average value of the page rank algorithm values for each node, where each node has a score proportional to the sum of its neighbors' scores. The importance of its neighbors then increases the importance of a node in the graph. The measure is called *Page Rank* since Google uses it as a central part of its web ranking technology. In matrix terms, it is calculated as follows:

$$x = D(D - \alpha A)^{-1}\mathbf{1},$$

where $A$ is the adjacency matrix, $D$ is a diagonal matrix with $D_{ii} = \max\{d_i^{out}, 1\}$, being $d_i^{out}$ the number of outgoing edges of node $i$, and $\mathbf{1} = (1, 1, \dots)$. The equation has a free

parameter $\alpha$ that must be chosen a priori. In practical cases for directed graphs it is usually roughly of order 1 (the Google search engine uses $\alpha = 0.85$).

*Clustering coefficient* It is the average over the clustering coefficient values for each node, which quantifies the transitivity level of the graph. That is, if node $i$ is connected with node $j$ and node $j$ is connected with node $k$, how likely is that node $i$ is connected with node $k$. For undirected binary graphs, it is defined as the fraction of paths of length two in the graph that are closed (triangles). The *clustering coefficient* varies between 0 and 1, with $C = 1$ indicating perfect transitivity, and $C = 0$ for graphs with no closed paths such as trees. The previous definition can be generalized to weighted graphs by considering a function of the edge weights of the triangles instead of its number, i.e. that the value for node $i$ is defined as:

$$C_i = \frac{1}{d_i(d_i - 1)} \sum_{j,k} f(\hat{w}_{ji}\hat{w}_{ik}\hat{w}_{kj}),$$

where $d_i$ is the number of neighbours of node $i$ and $\hat{w}$ is the weight $w$ normalized over the maximum weight in the graph. In NetworkX the function $f$ corresponds to the geometric average of the subgraphs edge weights. Moreover, directed (and weighted) graphs can be also considered, dividing by $d_i^{tot}(d_i^{tot} - 1) - 2d_i^{\leftrightarrow}$ where $d_i^{tot} = d_i^{in} + d_i^{out}$ and $d_i^{\leftrightarrow} = A_{ii}^2$.

For undirected and binary graphs, we can also calculate the chromatic and independence numbers. A colouring of such a graph is defined as labeling the nodes with colours such that no two nodes sharing the same edge have the same colour. In our case, it requires to previously convert the graph into an undirected binary graph. The converted graph will have an edge between nodes $u$ and $v$ if at least one of the edges $(u, v)$ or $(v, u)$ is present in the original directed graph. Then, the two resulting features are detailed next:

*Maximum independent set* An *independent set* of a graph is a set of nodes that do not share any edge between them. The computation of an independent set of maximum size is an NP-hard problem and this maximum size is referred to as the *independence number*. We approximate the maximum independent set size taking the maximum value over 10 runs of the NetworkX algorithm [44] which finds a maximal independent set.

*Chromatic number* It is the smallest number of colors needed to colour a graph, and its computation is also an NP-hard problem. A classical colouring method is to define an order of the graph nodes and then assign to the node $i$ an available color not used by its smallest neighbours in the defined order, adding a new colour if needed. The quality of the colouring depends on the chosen strategy of nodes ordering. It may be used merely random ordering (usually known as greedy algorithms) or variants of degree dependent ordering (start first with the largest/smallest degree, for instance) or much-complicated strategies [45]. Another way to colour a graph is to assign the same colour to nodes that belong to an independent set of the graph. In our case we computed the chromatic number by means of the NetworkX greedy coloring algorithm using the largest first strategy.

Each of the selected features has a possible interpretation in the context of conflict graphs corresponding to the APs of a Wi-Fi network. For example, all the centrality measures, beyond their differences, seek to reflect how centric each node is and how much influence it has on other neighbouring nodes. In the context of a Wi-Fi network, this implies that an AP with large centrality values is likely to be related to areas of more significant interference with its neighboring APs. On the other hand, the maximal independent set and the chromatic number are related to critical issues, such as channel assignment and transmission power control, and in particular know how many channels are required to avoid interference between APs.

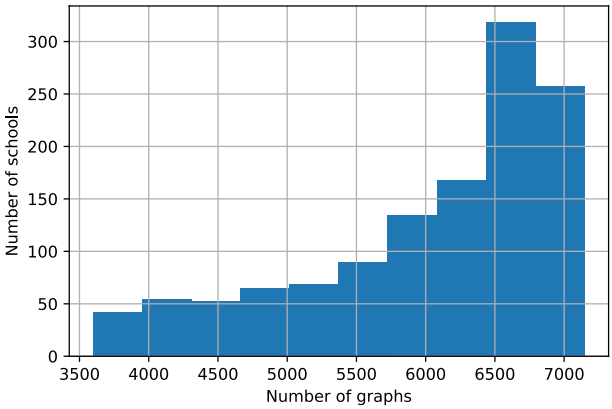



**Fig. 3** Number of graphs (i.e. different timestamps) for each school

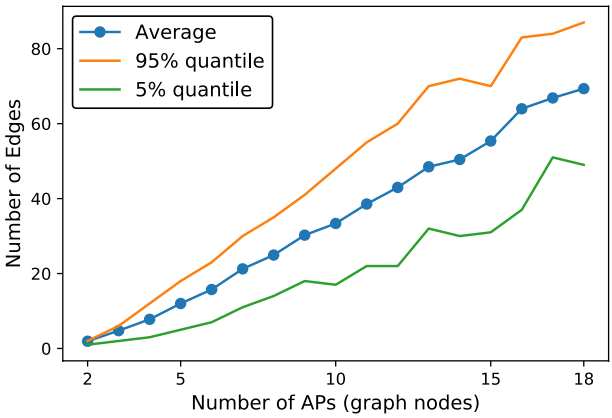



**Fig. 4** Relation between the number of APs (graph nodes) and the number of edges

## 4.2 Exploratory Analysis

Once we have computed all the features for the different graphs, we have also filtered some schools with few measurements (due to missing data during the collection process). Two conditions were imposed to get to the cured dataset for the analysis. The first one is to ensure that each school has the same number of APs for all the graphs. This may not occur, as some APs may have been removed or added during the year. In that case, what we do is keep for each school only the corresponding graphs with the number of APs with more measurements. Then, we set a minimum of 3600 graphs for each school (i.e., different timestamps), which corresponds to at least 150 full days during the year, with 24 measurements per day taken each hour. The resulting dataset has information from 1249 different educational centers, and the number of measurements for each of them is distributed according to Fig. 3.

Next, we present an exploratory analysis of the dataset. To begin with, Fig. 4 shows the relation between the number of APs, i.e. the graph nodes, and the number of edges. We can see that the relationship between them seems linear, which we verify with Fig. 5, looking at the convergence of the ratio between edges and nodes (reaching a value close to four). Please recall that the graph is directed, so 4 additional edges per

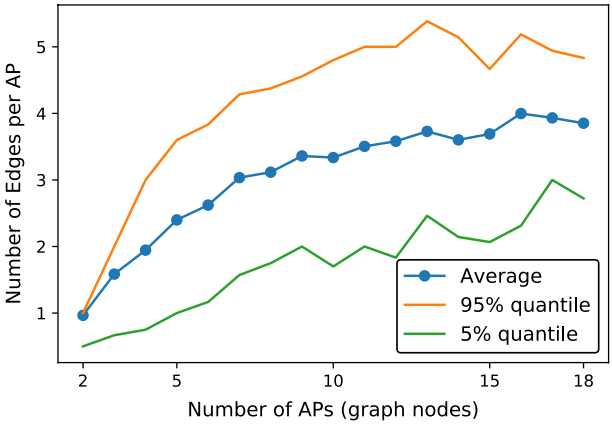



**Fig. 5** Ratio between the number of edges and the number of APs (graph nodes)

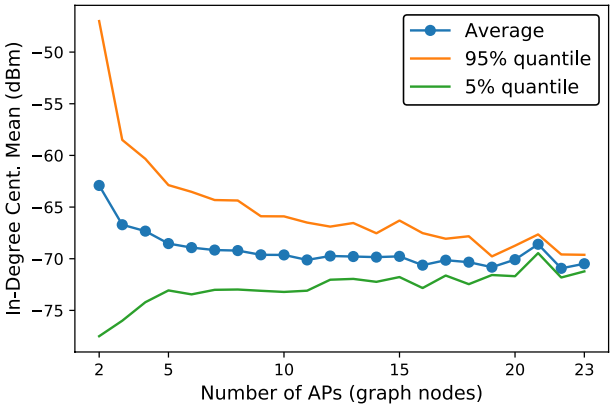



**Fig. 6** Relation between the number of APs (graph nodes) and the mean in-degree centrality

AP means that typically an AP has at least two neighbours on average. The curves of the 5% and 95% quantiles show that the dispersion is large, and variations of more than 20% with respect to the average are not rare. Such variability is a relevant property for our purposes, as we seek to find significant differences or common patterns between the graphs of the different schools. Furthermore, Fig. 6 shows the average in-degree centrality for the different number of APs per school. As we can see, there is no clear correlation between them, with an average value that remains stable when the number of nodes reaches five, but again with large variations around it. This property enables to fairly compare graphs from different schools with different number of APs, as their influence in the selected features is less relevant.

We extend the correlation analysis between the selected features, which is summarized in the correlation matrix shown in Fig. 7. As we can see, most of the features do not have a high correlation between each other, which avoids to have too much redundant information in the features vector. The highest correlation values correspond to the influence of the number of edges on other graph features, such as page rank, the maximum independent set and the chromatic number. Fig. 8 shows with more detail the correlation between the in-degree and betweenness centrality, where it becomes clear that they are not much related with each other.

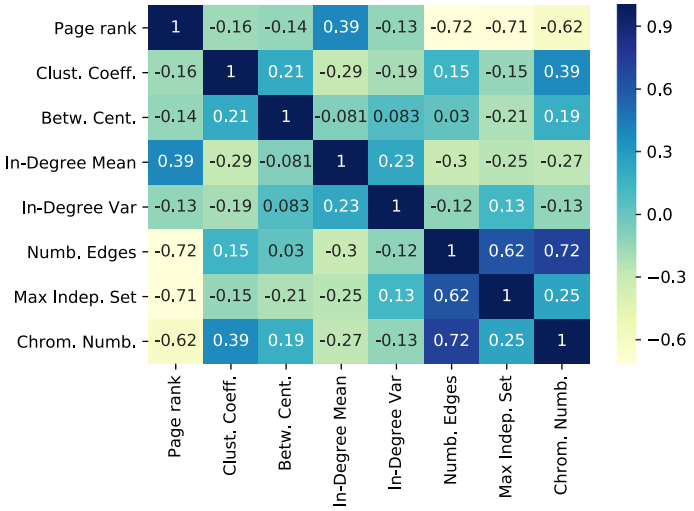



**Fig. 7** Correlation matrix for the different graph features

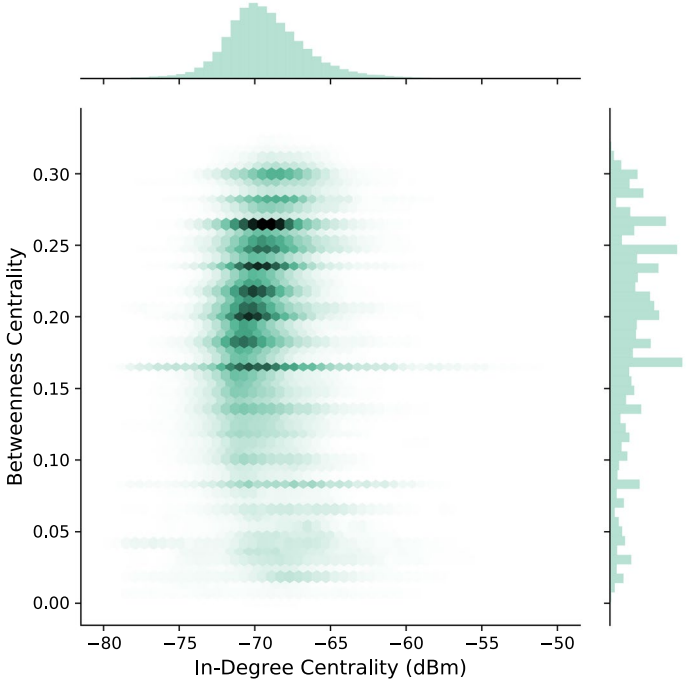



**Fig. 8** Relation between the mean in-degree and betweenness centrality

Finally, to end up with the data exploration and feature analysis, we present a PCA decomposition [46] to look further into each characteristic's relevance. For this purpose, first, we applied standard normalization to the features (i.e., centering and scaling each of

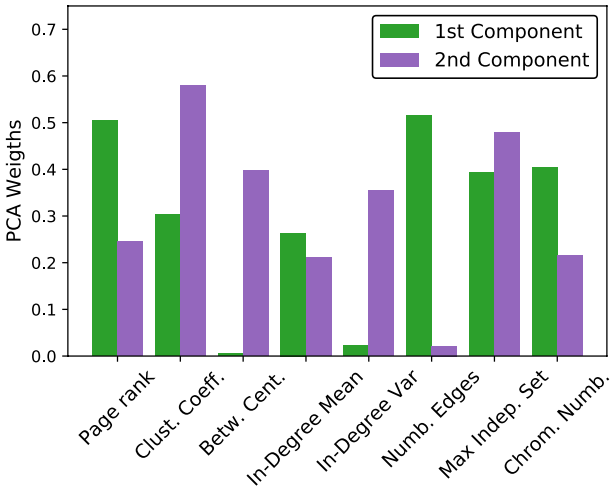



**Fig. 9** Features weigths for the 1st and 2nd PCA components

them to zero mean and standard deviation one). Then, we did the PCA decomposition of the normalized features, resulting in the weights presented in Fig. 9 for the first two PCA components. It is worth noting that all the selected features have significant weights in at least one of the two main PCA components, which indicates that all of them are relevant to discriminate the graphs (we will go further into this in the next section). This observation is also consistent with the low correlation between the features previously observed. The PCA results have also shown that only with the first three components an 88% of the variance explained is reached, and almost 96% with five, which means that the school's Wi-Fi graph space dimension is probably lower than the number of features considered.

## 5 Evaluation of the Discrimination Power of the Features

In this section, we evaluate the discrimination power of the selected features. Assessing this aspect is crucial to ensure the clustering methods' proper performance later used in the analysis. For this purpose, a supervised learning problem was posed, considering the corresponding school as the label for each graph. The underlying assumption is that the schools should have several similar graphs in the different timestamps. So, if the selected features are good enough to discriminate between the graphs, we should identify the schools based on their corresponding contention graphs. That is to say; it should be possible to solve the supervised classification problem with high accuracy. To determine if this is the case, we analyzed the resulting performance solving this problem for several standard classification algorithms.

Two different datasets were used, considering only the subset of schools with 10 Access Points (APs). The purpose of this is to tackle a more complex classification problem than considering all the schools in the dataset since the number of nodes for all the graphs in the selected subset is the same. Each dataset corresponds to the different periods taken into account. On the one hand, a one-month dataset (considering only October data), and on the other hand, the whole ten-month school year (from March to December). To avoid the schools with too much missing data (due to the data collection process described in Sect. 3), we discarded those with information for less than 100 days during the school year.

**Table 2** Accuracy for the different supervised classification algorithms

| Classification algorithm | Accuracy (%) | |
|---|---|---|
| | 1-month data (Oct. 2018) | 10-month data (Mar.–Dec. 2018) |
| Random forest | 82.7 ± 3.1 | 79.9 ± 0.1 |
| k-nearest neighbors | 82.3 ± 3.4 | 79.0 ± 0.1 |
| SVC (rbf) | 80.9 ± 5.5 | 77.4 ± 0.1 |
| Gradient boosting | 81.3 ± 3.1 | 76.7 ± 0.1 |
| Decision tree | 75.0 ± 3.2 | 72.8 ± 0.2 |
| SVC (linear) | 80.1 ± 4.0 | 71.8 ± 0.1 |
| Logistic regression | 64.4 ± 6.7 | 59.6 ± 0.1 |

We also kept only the data for school days (i.e., from Monday to Friday). This way, we ended up with a subset of 69 schools with 10 APs (i.e., 69 different categories).

We followed a classical machine learning approach, randomly dividing the data into two sets: 80% for training and 20% for test. The data split was done in a stratified fashion to ensure a suitable data proportion for each school in the training and test sets. We applied a standard scaler for the normalization, which converts each feature to have a mean value of 0 and a standard deviation of 1. We compared the results for six different algorithms from the standard machine learning Python package scikit-learn [47]. We used $k = 3$ for k-NN and $C = 1.2$ for support vector classifier (SVC) with RBF[1] kernel (found via grid search for the October data). The standard parameters were used for the rest of the algorithms.

In Table 2, we present the results for the different classification methods with both datasets. They are ordered according to the performance with the larger dataset. As we can see, all the methods have better performance for the 1-month dataset, which was to be expected, since the variations for each school should be smaller for a shorter period. The best algorithm reaches an accuracy of 80%, which is high, considering that the classification is between 69 different schools. Thus, the results verify that the selected features have a high discrimination power among graphs. This fact allows us to move forward to the clustering analysis, ensuring that the clusters found should be meaningful, at least according to the graphs' similarities.

Finally, we further analyze two more questions about the results obtained. Concerning the accuracy, it is computed as the average for all the samples on each test dataset. That is to say; we have an average accuracy that mixes the results for the different schools. So the first question that arises is: *Is the classification performance similar for all the schools?* The answer is provided by Fig. 10, where we can see the accuracy for each school. As we can see, there are significant differences between schools, ranging from below 50% up to near 100%. The performance is above 65% for most of them, and just a few have a worse accuracy, close to 50%. This result is a preview of what we will see in the next section, with significant differences in the different school networks behaviour.

The other question we addressed is for the cases where the classification was incorrect: *Are the misclassified graphs for a particular school always confused with the same other school?* The intuition behind this question is that if two schools are *similar*, so should be

[1] RBF: Radial basis function.

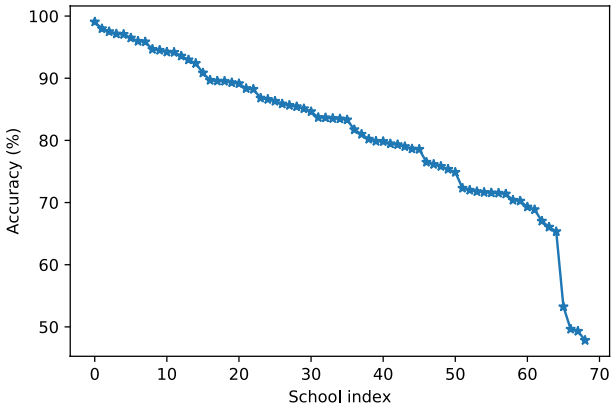



**Fig. 10** Classification accuracy for each school (Mar.–Dec. 2018 data)

their graphs. In this context, *similar* refers not only to the building architecture and the APs placement but also to the behavior and mobility of people inside each school. We have found that only two schools have more than 50% of the errors with the same other school, which is less than 3% of the schools in the dataset. If we consider at least 40% with the same other school, the number of cases climbs to 12 (which is still low, reaching just 17% of all the schools). Also, it is interesting to note that the effect observed is not reciprocal. That is to say, a large number of errors assigning school A graphs to school B do not imply that many graphs from school B are assigned to school A.

## 6 Clustering Algorithm Selection

After evaluating the discrimination power of the selected features, now we need to choose a clustering method from all the different options available to continue with the proposed dataset analysis. For this purpose, we defined a classification problem considering again as ground truth the corresponding school as the label for each graph. The difference this time is that we use a clustering algorithm, which is an unsupervised learning method, in order to solve the classification problem.

Next we explain the procedure followed to compare the different clustering algorithms. First, it is worth to mention that the two datasets used for this algorithm comparison were exactly the same as in the previous section. That is to say, the 1-month dataset (with October data) and the 10-month school year dataset (from March to December), both of them corresponding to the subset of 69 schools with 10 APs. For each dataset, we have run the different clustering algorithms, varying the number of clusters parameter (from 100 to 1000), and using the standard values for the rest of the different parameters of each method. Thus, we ended up with different clustering results for each algorithm, and each of them corresponds to a certain label assignment for each graph of the datasets.

In order to be able to compare the algorithms classification performance with respect to a *ground truth*, we have used the cluster labels assigned by each method to map this resulting clusters with the different schools. To do so, we have assigned for each cluster label, the corresponding school with the largest number of graphs within each group (in case of ties we just select one school randomly). In this way, we have for each algorithm

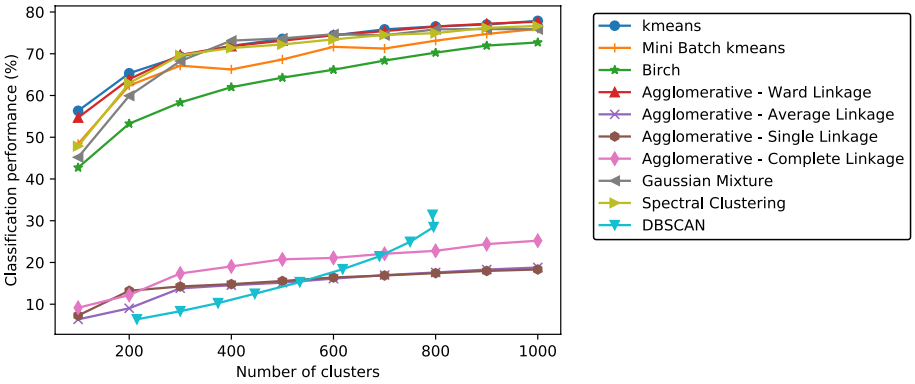



**Fig. 11** Performance comparison of the different clustering algorithms for the 1-month dataset (October 2018)

execution, the corresponding classification of the graphs indicating which school each one corresponds to.

Clustering methods can be classified according to what they are based on (see [48, 49] for references on classic and modern methods). We considered several different algorithms, such as hierarchical clustering or methods based on centroids (e.g. *k*-means), among others. All of them are included in the scikit-learn Python library [47]. We compared the different algorithms, according to the classification performance obtained by each of them. Fig. 11 shows the results for the different algorithms for the one-month dataset corresponding to October 2018. As we previously mentioned, we have varied the number of clusters parameter from 100 to 1000, obtaining different results for each clustering algorithm. For the case of the hierarchical agglomerative clustering, all the different linkage methods were tested (Ward, single, average and complete). Mini batch *k*-means and Birch are computationally efficient variants of *k*-means, while Spectral clustering enables the detection of non-convex clusters. The Gaussian mixture model is the most popular method based on a prior probability distribution. Finally, DBSCAN is a density-based method, which does not require the number of clusters as input. Thus, in order to cover a similar range than for the rest of the algorithms (i.e. from 100 to 1000), we have varied the DBSCAN $\epsilon$ parameter accordingly. The default values were used for all the other parameters of each algorithm.

As we can see, the performance is significantly better for four algorithms: *k*-means, agglomerative hierarchical clustering using Ward linkage, Gaussian mixture model and Spectral clustering. It is worth to mention that the first two are computationally much less expensive than the last two. Moreover, Fig. 12 shows another comparison between the clustering algorithms results. In this case, we analyze which percentage of the schools in the dataset has at least one corresponding cluster (using the assignment method described previously, which is to assign to each cluster the school with more graphs within the cluster). We can see that the methods that achieve larger school coverage with fewer number of clusters are the same that have a better classification performance.

Next, we repeated the same analysis with the ten-month dataset, but this time only for the most computationally efficient algorithms that had better results with the other dataset. As we can see in Fig. 13, the best results were obtained with the *k*-means clustering algorithm, which systematically performs better than the other algorithms. We can also notice that the results for the whole school year dataset are slightly worse than for the 1-month

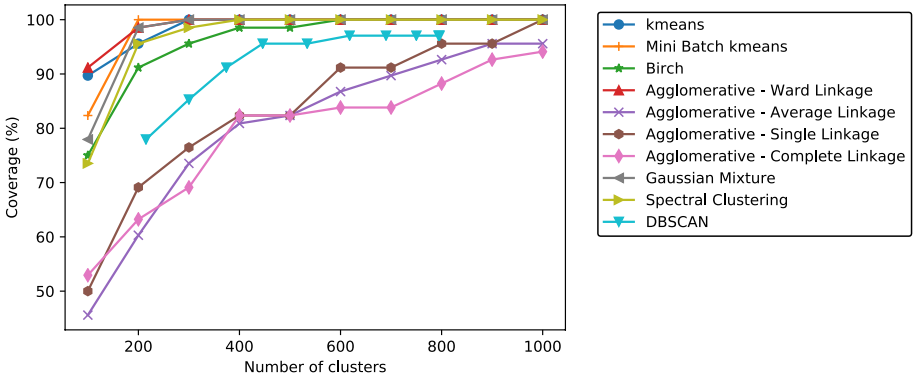



**Fig. 12** School coverage of the different clustering algorithms for the 1-month dataset (October 2018)

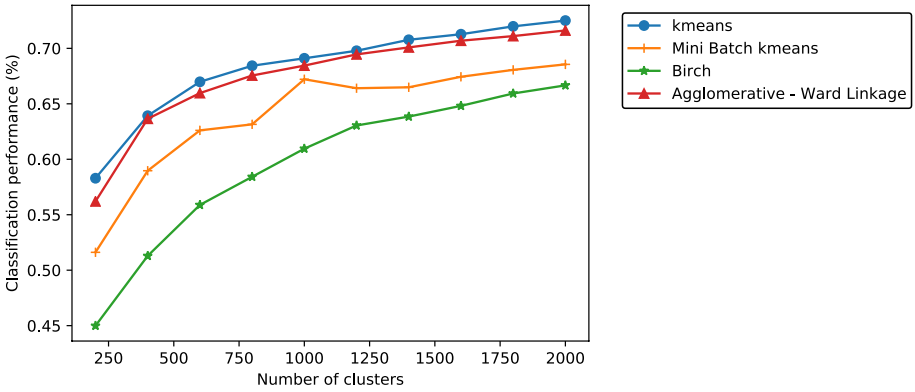



**Fig. 13** Performance comparison of the different clustering algorithms for the 10-month dataset (March-December 2018)

dataset. This result seems reasonable, since over a longer period of time it is expected that there will be more variations in the graphs, and therefore it will be more difficult to classify them according to the corresponding school. Thus, we require a greater number of clusters to achieve the same classification performance (please notice that in this case the number of clusters goes up to 2000 and not up to 1000 as before).

One of the drawbacks that k-means may have is its sensitivity to outliers. Thus, we also compared the performance of *k*-means with two robust clustering algorithms that tackle this issue: *tk*-means [50] and *k*-taucenters [51]. For both cases we used the available packages for the statistics software R [52]. The results for the one-month dataset are presented in Fig. 14. As we can see, *k*-means has almost the same results as *k*-taucenters, and both of them outperform *tk*-means when the number of clusters raises. Since robust algorithms are more computationally expensive than *k*-means, the results obtained do not justify their use in this case.

Based on these results we reach to the conclusion that the *k*-means algorithm is the best one for the analysis of the Wi-Fi conflict graphs generated from the RSSI measurements. We believe that the better performance in these *supervised* tests implies a suitable modeling of the graph features' space geometry. In addition, the selected method has low

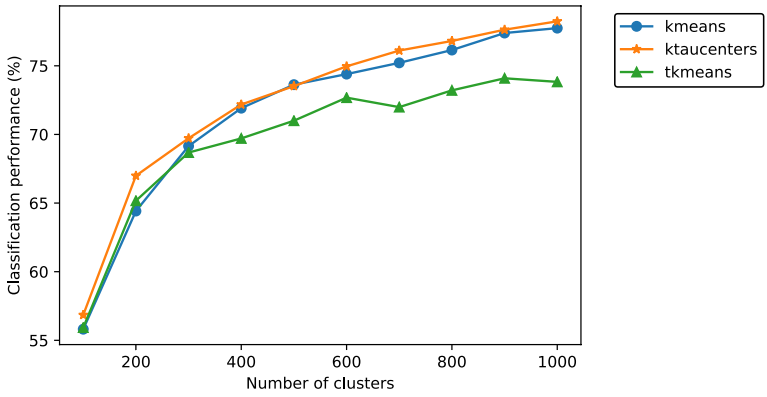



**Fig. 14** Performance comparison of *k*-means vs robust clustering algorithms for the 1-month dataset (October 2018)

computational cost, making it more appropriate for analyzing large datasets, as in this case. In the next section, we will use the *k*-means clustering algorithm to analyze the temporal and spatial variations of the graphs corresponding to the schools Wi-Fi networks.

## 7 Clustering Analysis and Resulting Insights

Having chosen the most suitable clustering algorithm, we now go back to the complete dataset introduced in Sect. 3 for the analysis. Let us recall that the cured dataset is composed by the different graphs, described by its features, of the 1249 schools that have data for at least 3600 different timestamps (i.e. 150 full days of 24 h). Next, we analyze this dataset in two different and complementary ways. First, we focus on the graphs' temporal variations, in this case separately studying the particular dynamics of the Wi-Fi network at each location. Then, we look for the similarities between the networks of the different schools, searching for common patterns among them.

### 7.1 Analysis of Temporal Variations: When and How Often Does the Wi-Fi Contention Graph of Each School Change?

First, we focus on the analysis of the temporal variations of each school graph. For this purpose, we use the selected clustering algorithm, *k*-means, to find out how many graphs clusters does each school have during the year. To do so, we run *k*-means independently for each school data and look for the most suitable number of clusters using two standard techniques, such as the Elbow method [53], combined with an automatic elbow detection method [54], and the Silhouette score [55], implemented in the scikit-learn Python library [47]. We integrate the results of both methods taking the minimum number of clusters between them. We base this selection in the criteria of choosing for each school the most simple model that suits the data.

In Fig. 15 we can see the resulting number of clusters for each school. Each cluster can be interpreted as a different *temporal operation mode* of each school Wi-Fi network, represented by its conflict graph. It is worth noting that most of the schools have a low number of *temporal modes*, which suggests a fairly stable Wi-Fi operation. Moreover, we

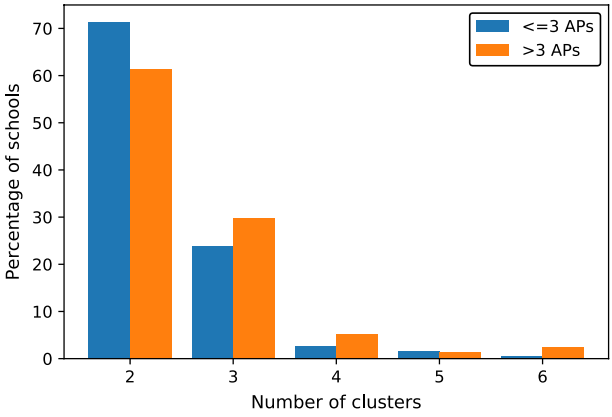



**Fig. 15** Number of clusters (*temporal modes*) for each school

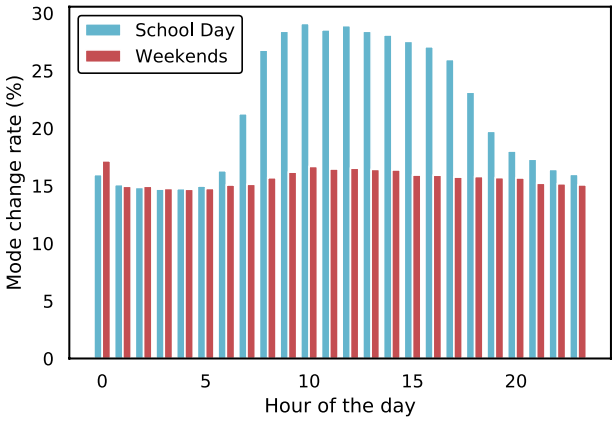



**Fig. 16** Temporal mode change rate according to the hour of the day

can see that schools with more than 3 APs, tend to have a larger number of clusters. Since there are only 3 non-overlapping 20 MHz channels in the 2.4 GHz band, schools with more than 3 APs have at least one overlapping channel in this band. This fact implies that at this point the radio resource management relevance increases, since the frequency channels and transmission powers must be adjusted in order to mitigate the effect of the unavoidable interference between APs.

Concerning the question about when do the operation modes changes happen, different time windows were studied, i.e. month, week, day and hour. As we can see in Fig. 16, most of the changes occur during school hours (8 am–5 pm), which are the busiest times in the buildings (i.e. teachers and students moving to and from classrooms). The change rate during the weekends (see Fig. 17) is similar to that observed during the night and early morning hours the rest of the week. The peak hours in terms of changes in the network, almost double the change rate observed during the most stable operation moments.

Another relevant question that arises is how often do the operation mode changes occur. For this purpose, we analyze the distribution of the time between changes, which is shown in Fig. 18. It is worth to note that most of the changes (almost half of them) occur within

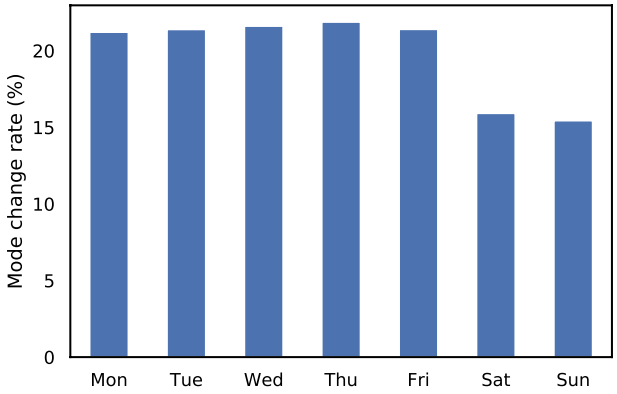



**Fig. 17** Temporal mode change rate according to the weekday

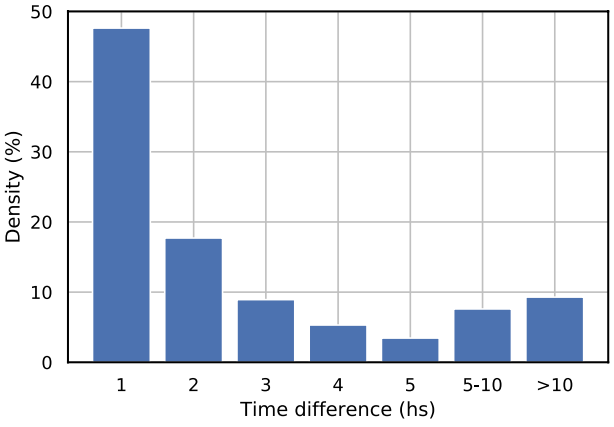



**Fig. 18** Histogram of the time difference between mode changes

a time difference of one hour, which corresponds to consecutive measurements, as it is the data collection frequency. The graph clearly shows a heavy tail in the distribution, which means that a significant number of changes occur after many hours or even days of elapsed time since the last operation mode change. This fact is also reflected in Fig. 19, where significant differences are found looking at the average elapsed time between changes for the different schools. This fact indicates that there are Wi-Fi networks of certain schools with long periods of very stable operation.

Finally, we concentrate on the larger group of schools with only two temporal modes, in order to dig into the average feature differences observed between the two modes. In Fig. 20 the boxplot for the different features is shown, where the relative differences are computed as $100 \times (a - b)/(a + b)$, being $a$ and $b$ the average feature values for each mode, where the order is randomly chosen (so possible values are either positive or negative). The first thing we can notice is that page rank shows almost no differences between the two modes, so it is not useful to differentiate the temporal variations for the same school network. Other centrality measures, such as betweenness and clustering coefficient, showed significant differences between the modes. The average in-degree was another feature without much difference between the modes, in contrast with its

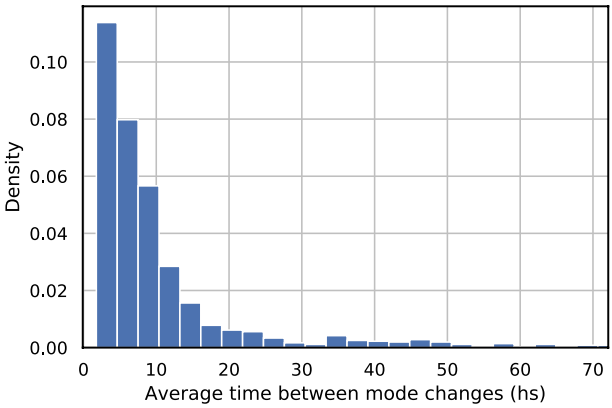



**Fig. 19** Histogram of the average time between changes for each school

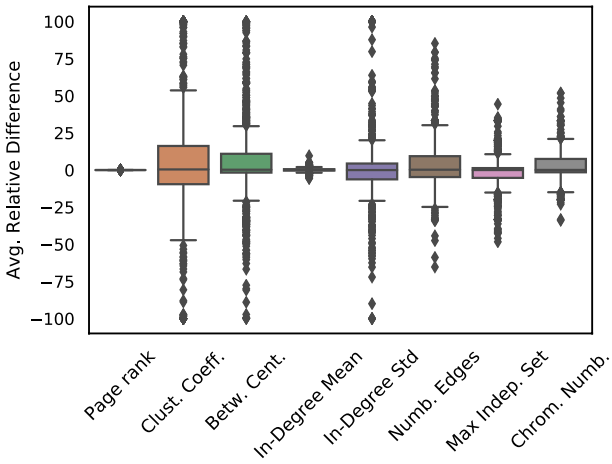



**Fig. 20** Relative differences between average features for schools with two *temporal modes* (i.e. differences for each feature ranges from − 100 to 100%)

variance which showed much larger differences. This is also reflected in the difference observed in the average number of edges for each mode, and also to a lesser extent in the maximum independent set and the chromatic number.

One question that arises looking at the differences between the two modes, is whether all cases really have two temporal modes or if some of them should be reduced to only one network operation mode. The results indicate that 28.4% of the schools have an average difference over all features below 5%, and the percentage raises to 58.9% if the threshold is 10%. If we consider the maximum relative difference between all the features, 30.2% are below 20% and 44.2% below 30%. Thus, a significant number of schools networks presented small variations of the conflict graph during the year, which could be interpreted as a single operation mode. Unfortunately, neither the elbow method nor the silhouette score are useful to evaluate the case of $k = 1$. Thus, we used another metric called Gap Statistic [56] which allows us to evaluate the case of a single cluster. We analyzed the dataset of schools with 10 APs and for 51.5% of the cases the Gap Statistic indicated that $k = 1$ was the optimal number of clusters. This result is consistent with the cases observed with small differences in the features between the two modes.

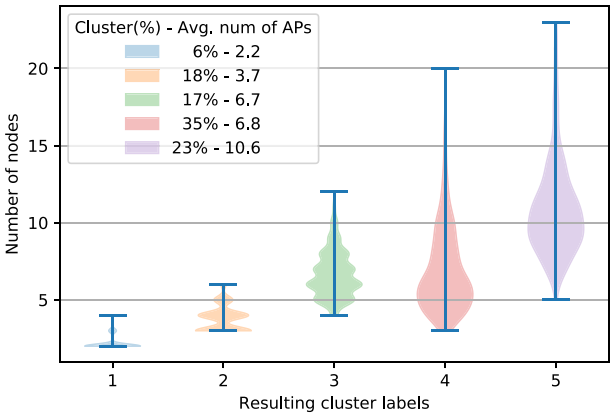



**Fig. 21** Clustering results for the different {school, temporal mode} pairs

One last observation to highlight, is the correlation between the number of temporal mode changes observed and the relative difference between the average features for each mode. The results showed no relation between both, which implies that knowing how changeable the RF environment of the Wi-Fi network is does not provide any information on whether such variability is large or small. This fact is relevant when analyzing the network operation, because in many cases design and optimization decisions are based on field surveys that correspond to a single picture of the state of the network.

### 7.2 Analysis of the Spatial Variations: How Different are the Graphs of the Different Schools?

After the temporal analysis, now we would like to find out if there are common conflict graph patterns between the different schools. For this purpose, we take advantage of the previous clustering results and consider all the {school, temporal mode} pairs. For each of them we compute the aggregated features, taking the average over all the graphs that corresponds to each pair. Finally, we apply the *k*-means clustering algorithm to this novel dataset, composed by all the different temporal modes for each school. Another interesting thing to observe is whether the different temporal modes of the same school are grouped together by their similarity or not.

The *k*-means clustering of the 3139 {school, temporal mode} pairs resulted in a five-group classification, according to the Elbow method, used to find out the proper *k*. Fig. 21 summarizes the resulting clusters, according to the distribution of the number of APs within each class. As we can see, the number of APs has an influence on the grouping, mainly due to the average number of edges, included as one of the graph features. However, this parameter is not decisive in the result, since we also find many schools with the same or similar number of APs that fall into different classes.

Concerning the different temporal modes of the different schools, we have found that for schools with two or three temporal modes, most of them fall into the same cluster, while for schools with more temporal modes, at least one temporal mode falls in a different cluster in the majority of cases. All the results are summarized in Table 3. It is worth to note that for schools with three or more temporal modes, in all cases they are divided into at most three of the clusters found. Being able to group the different schools (actually more than 1200) only into a few classes has significant value for

**Table 3** Distribution among clusters according to the number of temporal modes (all values correspond to the % of schools)

| | | Number of temporal modes | | | | |
|---|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 5 | 6 |
| Number of classes in which they are distributed | 1 | 74 | 64 | 48 | 35 | 43 |
| | 2 | 26 | 34 | 49 | 59 | 50 |
| | 3 | – | 2 | 3 | 6 | 7 |

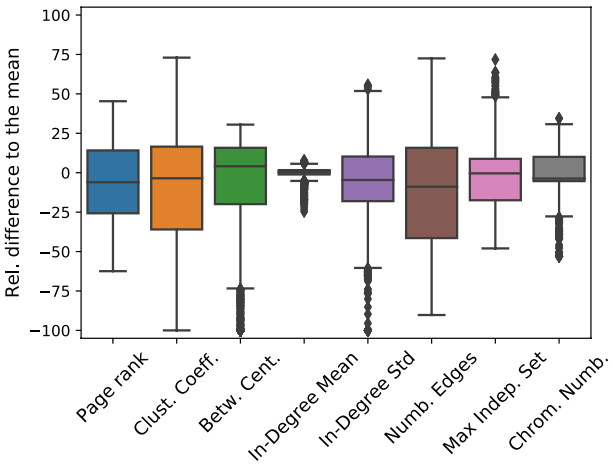



**Fig. 22** Relative differences to the mean average feature values among the different schools (i.e. differences for each feature ranges from − 100 to 100%)

network management. For example, it is possible to take advantage of these similarities and reuse configuration parameters for sites with similar conflict graph patterns.

Next, we look again at the differences observed for each of the selected graph features. In this case, the variations correspond to different schools, and are shown in Fig. 22 with respect to the mean. That is to say, we compute the average of each feature over each school graph time series, and compare the value with the mean average for all schools. As we can see, now the variations are larger than the ones previously shown in Fig. 20, which is reasonable because it indicates that the differences between different schools are larger than the ones observed between the graph time series of each school. Moreover, we can notice that now the page rank feature also shows significant variability, unlike what was observed for the time series of the same school.

Finally, we try to see if there is any correlation between the clusters found and the building characteristics of the schools. For this purpose we focused on the subset of schools with 10 APs, for which we had information regarding the number of floors in each building and the existence or not of prefabricated container classrooms (in some schools, additional classrooms have been added to increase the capacity, which are made with standard ship containers). Both attributes have shown a possible influence in the number of temporal modes of each school. While 72% of the schools with only one floor have two temporal modes, this percentage falls down to 65% for the schools with two or more floors. The presence of container classrooms seems to have even more

influence, as 78% of the schools without them have only two temporal modes, while it is 55% for the schools which do have container classrooms.

It seems that having more floors or container classrooms has an upward influence in the number of temporal modes, possibly explained by a more variable RF environment of those Wi-Fi networks. Although this is not conclusive, they are indications that some building characteristics could have influence in the variability observed in the temporal dynamics of the Wi-Fi networks in the different schools. Nevertheless, for practical purposes it is enough to know which schools have similar Wi-Fi network conditions. This makes it possible to replicate the optimal solutions and configurations found in one site, to all those that are similar, avoiding the efforts involved in the field survey and analysis of each particular location.

### 7.3 Discussion on the Practical Implications

One of the most time consuming tasks involved in the design, deployment, optimization and maintenance of Wi-Fi networks, has to do with the field surveys and RF measurements analysis. It typically requires many hours working on site, in addition to the travel required and the subsequent data analysis. An in-site validation is essential for the initial deployment, when the location of the different APs is planned and the solution is evaluated once it is installed. Then, the workload for operation and maintenance depends mainly on how stable is the network operation.

The presented methodology proved to be useful for the analysis of Wi-Fi networks, which in addition allows to plan the most appropriate resource allocation for preventive maintenance. The results obtained based on the conflict graph features embedding, enable to detect the most problematic sites which correspond to a more dynamic RF environment. This is of great help for the network operation, providing key information to decide how to prioritize the allocation of technical resources to ensure the optimal network performance.

Another key point for the deployment of Wi-Fi networks is to adjust the most suitable configuration for each site. Typically, high-end solutions include several parameters related to the RF management, for example for tuning their proprietary algorithms for channel assignment or transmission power control. The graph clustering procedure presented makes it possible to group together different sites with similar conflict graph patterns. This resulting grouping of similar Wi-Fi networks allows to optimize the operation reusing the same parameters in those sites that are similar, avoiding the important workload required to fine tune this parameters for each site individually.

## 8 Conclusions and Future Work

In this paper, we analyze a large-scale dataset of real-world Wi-Fi operating networks, corresponding to all the primary and secondary schools throughout a country. For this purpose, we use very well know mathematical tools, such as graph analysis and clustering algorithms. While these tools have been widely used to address a variety of wireless network problems, we believe that they still have much to contribute when it comes to analyzing the large volumes of data that can be collected and processed today from real-world operating networks. The biggest challenge is how to incorporate all the available data and turn it into useful information for the deployment, optimization and maintenance of wireless networks.

We propose an efficient and useful graph embedding for Wi-Fi conflict graphs, based on classical graph features, which proved to have a high discrimination power among the different schools Wi-Fi networks. Furthermore, the proposed graph embedding enabled us to study the graphs variations by means of clustering algorithms. First, we focused on the temporal dynamics of each school Wi-Fi network, analyzing the different conflict graph time series. The results allow us to identify which schools are more stable and which ones are more variable, and thus deserve more attention, so more technical resources should be assigned to them. That is to say, the thoughest scenarios should be prioritized to do the most time consuming tasks such as field surveys and RF analysis. On the other hand, we have studied how to group together different schools with similar conflict graphs patterns, which makes it possible to avoid the efforts involved to do field surveys and RF analysis at each particular location and use the optimal configuration parameters found for one site in all other similar schools.

A relevant goal for future work would be to integrate the analysis developed through the conflict graphs for the automatic generation of optimal configurations. The development of self-configuration capabilities for Wi-Fi networks would further reduce the technical resources required for maintenance and operation. In addition, other ways to extract useful information from conflict graphs could be explored, such as novel graph embeddings techniques and graph neural networks, which are showing promising results in other areas. This also opens the doors to the application of modern machine learning techniques, such as autoencoders and generative adversarial networks (GANs), not only for extracting graph embeddings, but also for simulation purposes, based on its ability to generate synthetic data from large real datasets.

**Acknowledgements** This work was partially supported by ANII (Grant FMV_3_2018_1_148149) and was approved by Plan Ceibal's ethical and data privacy committee.

**Funding** This work was partially supported by ANII (Grant FMV_3_2018_1_148149).

**Availability of data and material** This work was approved by Plan Ceibal's ethical and data privacy committee. We are currently in the authorization process to make the data public and open.

**Declarations**

**Conflicts of interest** The authors have no conflicts of interest to declare that are relevant to the content of this article.

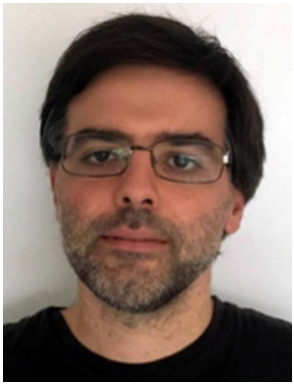
**Code availability** All the work done in this article is based on open source libraries as detailed in the manuscript. The code developed for the data analysis will be made public as soon as the data can be published.

## References

1. Sherlock, I., et al. (2020). Wi-Fi Alliance 2019 Annual Report (Published)
2. JP Vasseur: AI for Networking: Separating the Hype from Reality. Cisco Blogs, Networking (2020). https://blogs.cisco.com/networking/ai-for-networking-separating-the-hype-from-reality. Accessed from 8 Mar 2021.
3. Challita, U., Ryden, H., & Tullberg, H. (2020). When Machine Learning Meets Wireless Cellular Networks: Deployment, Challenges, and Applications. *IEEE Communications Magazine, 58*(6), 12–18.
4. Zappone, A., Di Renzo, M., & Debbah, M. (2019). Wireless networks design in the Era of deep learning: Model-based, AI-based, or both? *IEEE Transactions on Communications, 67*(10), 7331–7376.

5. Zhang, C., Patras, P., & Haddadi, H. (2019). Deep learning in mobile and wireless networking: A survey. *IEEE Communications Surveys Tutorials, 21*(3), 2224–2287.
6. Capdehourat, G., Larroca, F., & Morales, G. (2020). A nation-wide Wi-Fi RSSI dataset: Statistical analysis and resulting insights. In: 2020 IFIP Networking Conference, Networking 2020, Paris, France, June 22–26, 2020, IEEE 370–378.
7. Kolaczyk, E. D. (2009). *Statistical analysis of network data: Methods and models* (1st ed.). Berlin: Incorporated: Springer Publishing Company.
8. Haenggi, M., Andrews, J. G., Baccelli, F., Dousse, O., & Franceschetti, M. (2009). Stochastic geometry and random graphs for the analysis and design of wireless networks. *IEEE Journal on Selected Areas in Communications, 27*(7), 1029–1046.
9. Ramachandran, K., Sheriff, I., Belding, E. M., & Almeroth, K. C. (2008). A multi-radio 802.11 mesh network architecture. *Mobile Networks and Applications, 13*(1–2), 132–146.
10. Jarupan, B., & Ekici, E. (2011). A survey of cross-layer design for VANETs. *Ad Hoc Networks, 9*(5), 966–983.
11. Katzela, I., & Naghshineh, M. (1996). Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey. *IEEE Personal Communications, 3*(3), 10–31.
12. de Oliveira, C.T., Theoleyre, F., & Duda, A. (2012). Channel assignment strategies for optimal network capacity of IEEE 802.11s. In: Proceedings of the 9th ACM Symposium on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks, Paphos, Cyprus p. 53-60.
13. Marina, M. K., Das, S. R., & Subramanian, A. P. (2010). A topology control approach for utilizing multiple channels in multi-radio wireless mesh networks. *Computer Networks, 54*(2), 241–256.
14. Ramachandran, K.N., Belding, E.M., Almeroth, K.C., & Buddhikot, M.M. (2006). Interference-aware channel assignment in multi-radio wireless mesh networks. In: Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications, pp. 1–12
15. Plummer, A., Wu, T., & Biswas, S. (2007). A cognitive spectrum assignment protocol using distributed conflict graph construction. In: MILCOM 2007 - IEEE Military Communications Conference, pp. 1–7.
16. Kala, S., Reddy, M., Musham, R., & Tamma, B. (2016). Interference mitigation in wireless mesh networks through radio co-location aware conflict graphs. *Wireless Networks, 22,* 679–702.
17. Cui, Y., Li, W., & Cheng, X. (2011). Partially overlapping channel assignment based on "node orthogonality" for 802.11 wireless networks. In: 2011 Proceedings IEEE INFOCOM, pp. 361–365.
18. Kim, S. H., Kim, D. W., & Suh, Y. J. (2012). A group-based channel assignment protocol for rate separation in IEEE 802.11-based multi-radio multi-rate ad hoc networks. *Ad Hoc Networks, 10*(1), 95–110.
19. Jeunen, O., Bosch, P., Herwegen, M.V., Doorselaer, K.V., Godman, N., & Latré, S. (2018). A machine learning approach for IEEE 802.11 channel allocation. In: 14th International Conference on Network and Service Management (CNSM), pp. 28–36.
20. Nakashima, K., Kamiya, S., Ohtsu, K., Yamamoto, K., Nishio, T., & Morikura, M. (2020). Deep reinforcement learning-based channel allocation for wireless lans with graph convolutional networks. *IEEE Access, 8,* 31823–31834.
21. Zeng, K., Lou, W., & Zhai, H. (2008). On end-to-end throughput of opportunistic routing in multirate and multihop wireless networks. In: IEEE INFOCOM 2008 - The 27th Conference on Computer Communications, pp. 816–824.
22. Cerdà-Alabern, L., Neumann, A., & Maccari, L. (2015). Experimental evaluation of bmx6 routing metrics in a 802.11an wireless-community mesh network. In: 2015 3rd International Conference on Future Internet of Things and Cloud, pp. 770–775.
23. Wang, K., Yang, F., Zhang, Q., Wu, D. O., & Xu, Y. (2007). Distributed cooperative rate adaptation for energy efficiency in IEEE 802.11-based multihop networks. *IEEE Transactions on Vehicular Technology, 56*(2), 888–898.
24. Chan, A., & Liew, S. C. (2009). Performance of VoIP over Multiple Co-Located IEEE 802.11 Wireless LANs. *IEEE Transactions on Mobile Computing, 8*(8), 1063–1076.
25. Amer, M., Busson, A., & Lassous, I.G. (2018). Association optimization in wi-fi networks based on the channel busy time estimation. In: 2018 IFIP Networking Conference (IFIP Networking) and Workshops, pp. 298–306.
26. Cheng, X., Mohapatra, P., Lee, S., & Banerjee, S. (2008). MARIA: Interference-aware admission control and QoS routing in wireless mesh networks. In: 2008 IEEE International Conference on Communications, pp. 2865–2870.
27. Gupta, R., Musacchio, J., & Walrand, J. (2007). Sufficient rate constraints for qos flows in ad-hoc networks. *Ad Hoc Networks, 5*(4), 429–443.
28. Lin, Y., & Wong, V. W. (2008). An admission control algorithm for multi-hop 802.11e-based wlans. *Computer Communications, 31*(14), 3510–3520.

29. Zuyuan, Fang, & Bensaou, B. (2004). Fair bandwidth sharing algorithms based on game theory frameworks for wireless ad-hoc networks. *IEEE INFOCOM, 2004*(2), 1284–1295.
30. Niculescu, D. (2007). Interference map for 802.11 networks. In: Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC'07, p. 339-350. Association for Computing Machinery, New York, NY, USA.
31. Cheng, Y., Li, H., Wan, P., & Wang, X. (2010). Capacity region of a wireless mesh backhaul network over the csma/ca mac. In: 2010 Proceedings IEEE INFOCOM, pp. 1–5.
32. Margolis, A., Vijayakumar, R., & Roy, S. (2007). Modelling throughput and starvation in 802.11 wireless networks with multiple flows. In: IEEE GLOBECOM 2007 - IEEE Global Telecommunications Conference, pp. 5123–5127
33. Stojanova, M., Begin, T., & Busson, A. (2019). Conflict graph-based model for IEEE 802.11 networks: A divide-and-conquer approach. *Performance Evaluation, 130,* 64–85.
34. Broustis, I., Papagiannaki, K., Krishnamurthy, S. V., Faloutsos, M., & Mhatre, V. P. (2010). Measurement-driven guidelines for 802.11 wlan design. *IEEE/ACM Transactions on Networking, 18*(3), 722–735.
35. Zhou, X., Zhang, Z., Wang, G., Yu, X., Zhao, B. Y., & Zheng, H. (2015). Practical conflict graphs in the wild. *IEEE/ACM Transactions on Networking, 23*(3), 824–835.
36. Cisco: Radio Resource Management White Paper. (2016). https://www.cisco.com/c/en/us/td/docs/wireless/controller/technotes/8-3/b_RRM_White_Paper.html Accessed from 8 Mar 2021
37. Li, W., Zhang, J., & Zhao, Y. (2017). Conflict graph embedding for wireless network optimization. In: IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, pp. 1–9.
38. Plan Ceibal: About Plan Ceibal. https://www.ceibal.edu.uy/en/institucional (2018). Accessed from 8 Mar 2021.
39. Conte, D., Foggia, P., Sansone, C., & Vento, M. (2004). Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence, 18*(03), 265–298.
40. Shervashidze, N., Schweitzer, P., Jan, E., Leeuwen, V., Mehlhorn, K., & Borgwardt, K. (2010). Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research, 1,* 1–48.
41. Newman, M. E. J. (2010). *Networks: An introduction*. Oxford: Oxford University Press.
42. Hagberg, A.A., Schult, D.A., & Swart, P.J. (2008). Exploring network structure, dynamics, and function using networkx. In: G. Varoquaux, T. Vaught, J. Millman (eds.) Proceedings of the 7th Python in Science Conference, Pasadena, CA USA, pp. 11 – 15
43. Afaqui, M.S., Garcia-Villegas, E., Lopez-Aguilera, E., Smith, G., & Camps, D. (2015). Evaluation of dynamic sensitivity control algorithm for IEEE 802.11ax. In: 2015 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1060–1065.
44. Boppana, R., & Halldórsson, M. (1992). Approximating maximum independent sets by excluding subgraphs. *BIT Numerical Mathematics, 32,* 180–196.
45. Kosowski, A., & Manuszewski, K. (2004). Classical coloring of graphs "Graph colorings." *Contemporary Mathematics, 352,* 1–19.
46. Jolliffe, I. (2011). *Principal component analysis* (pp. 1094–1096). Berlin: Springer.
47. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12,* 2825–2830.
48. Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, Ld. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLOS One, 14*(1), 1–34.
49. Wierzchoń, S. T., & Klopotek, M. A. (2018). *Modern algorithms of cluster analysis, Studies in Big Data*. Cham: Springer.
50. Cuesta-Albertos, J. A., Gordaliza, A., & Matrán, C. (1997). Trimmed *k*-means: An attempt to robustify quantizers. *Annals of Statistics, 25*(2), 553–576.
51. Gonzalez, J.D., Yohai, V.J., & Zamar, R.H. (2019). Robust clustering using tau-scales.
52. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2013). http://www.R-project.org/. Accessed from 8 Mar 2021.
53. Thorndike, R. L. (1953). Who belongs in the family. *Psychometrika, 18*(4), 267–276.
54. Satopaa, V., Albrecht, J., Irwin, D., Raghavan, B. (2011). Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In: 2011 31st International Conference on Distributed Computing Systems Workshops, pp. 166–171.
55. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20,* 53–65.
56. Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63*(2), 411–423.



**Germán Capdehourat** holds a PhD in Electrical Engineering (PhD) from UDELAR (Uruguay). He isProfessor at UDELAR and also member of the Uruguay National Research System since 2012. Heparticipated in R&D projects in the areas of image processing and wireless networks, with severalpublications in international conferences and journals. In the professional field he has been workingsince 2004 in the ICT sector and since 2007 in Plan Ceibal, the one laptop per child programnationwide deployed in the country. In this context he has several years of experience working withWiFi deployments both indoor and outdoor, as well as last-mile technologies for rural areas internetaccess.

**Paola Bermolen** was born in Montevideo, Uruguay in 1976. In 2004 she obtained a degree in Mathematics from the University of the Republic. In 1998 she joined the Institute of Mathematics and Statistics Prof. Rafael Laguardia of the Faculty of Engineering as an assistant. She has been an associate professor there since 2018. She obtained her PhD in 2010, at Telecom ParisTech, France, under the tutorship of Prof. François Baccelli and Prof. Dario Rossi. She has been responsible for several national and international projects. His areas of interest are related to stochastic modeling of telecommunication networks. More recently, the focus of his research is on the performance evaluation of wireless networks, including random geometry models and random graphs.

**Marcelo Fiori** received the Electrical Engineering, MSc, and PhD degrees from the Universidad de la República, Uruguay (UdelaR), in 2008, 2011, and 2015 respectively. He holds an Assistant Professor position at the Institute of Mathematics, UdelaR. His main research interests include machine learning, graph matching problems, and sparse representations, with special focus in signal processing.

**Nicolás Frevenza** is an Assistant Professor at the School of Economics and Administration of the Universidad de la República (Uruguay). He was a postdoctoral fellow at the Department of Mathematics of the Universidad de Buenos Aires (Argentina) from 2017 to 2019, working on probability theory and discrete analysis. He obtained his Ph.D. degree in Mathematics at the University of Buenos Aires under the supervision of Inés Armendáriz and Pablo Ferrari in March 2017. He received the degree in Mathematics in 2011 from the Universidad de la República.

**Federico Larroca** Federico 'Larroca' La Rocca is an Assistant Professor at the Engineering School of the Universidad de la República (Uruguay). He was a research engineering (PostDoc) at Telecom ParisTech (ex ENST) during the first quarter of 2010, where he obtained his Ph.D. degree in Computer Science and Networking under the advisoring of Prof. Jean-Louis Rougier in December 2009. He received the degree in Telecommunication Engineering in 2006 from the Universidad de la República. From 2004 to 2011 he held a teaching assistant position at the Universidad de la República.

**Gastón Morales** is an electrical engineer graduated from UdelaR (Uruguay) in 2020. He has been an Assistant Professor at the Engineering School of the UdelaR since 2018 in the area of telecommunications. He holds a research position at the same university and participates in several projects nowadays. His role in those projects include testing and data analysis.

**Claudina Rattaro** is an Assistant Professor at the Engineering School of the Universidad de la República (UdelaR). She has a degree in Electrical Engineering (speciality Telecommunications) from UdelaR since July 2008. She received her M.Sc. in Electrical Engineering in 2012 with a thesis concerning statistical tools in wireless networks and her Ph. D degree in 2017 with a thesis titled "Stochastic models for Cognitive Radio Networks", both from UdelaR. Her current research interests are related to the analysis and modeling of communication systems and artificial intelligence for networking.

**Gianina Zunino** is graduated in Telecomunication Engineering from Universidad Católica (Uruguay). Since 2013 she has been working in Plan Ceibal, the one laptop per child program nationwide deployed in the country. His mains tasks include management of network equipment and infrastructure, mainly wireless LAN controllers, access points and routers.