# Presence Analytics: Making Sense of Humans Social Presence Within a Learning Environment

Muawya Habib Sarnoub Eldaw
*Dept. of Computer Science*
*Birkbeck, University of London*
London, UK
eldaw@dcs.bbk.ac.uk

Mark Levene
*Dept. of Computer Science*
*Birkbeck, University of London*
London, UK
mark@dcs.bbk.ac.uk

George Roussos
*Dept. of Computer Science*
*Birkbeck, University of London*
London, UK
g.roussos@dcs.bbk.ac.uk

***Abstract*—The various activities that take place within an observed environment such as a university campus, determine to a large extent, the kind of social interactions exhibited by the users in such environments. Using a big data set of wifi-traces, we attempt to understand the rules that governs these social interactions. We discovered that there are at least two types of social interactions within a university campus: *formal* such as attending a class and *informal* such as meeting friends at the cafeteria for coffee. Each of these two types of social interactions is tightly associated with a specific set of locations within the university campus. We also discovered that users tend to restrict their social interactions to a small set of geographical locations, where users revisited the same location to socialise with the same social group. Also, irrespective of the type of the social interactions, users tend to restrict their revisits to geographically nearby locations and only revisit locations that are further afield when they are in the company of their social group. These findings are based on the social groups detected by a new scalable density-based clustering method applied to a large data set of mobile users wifi traces. The results of the large experiments carried out in this research demonstrate how the proposed algorithm can noninvasively detect social groups on the basis of the activity performed at the selected location.**

***Keywords*-Big data, Human Presence Analytics, Social Interaction, Mobile Data, Wifi, Density-based Clustering, Social Groups**

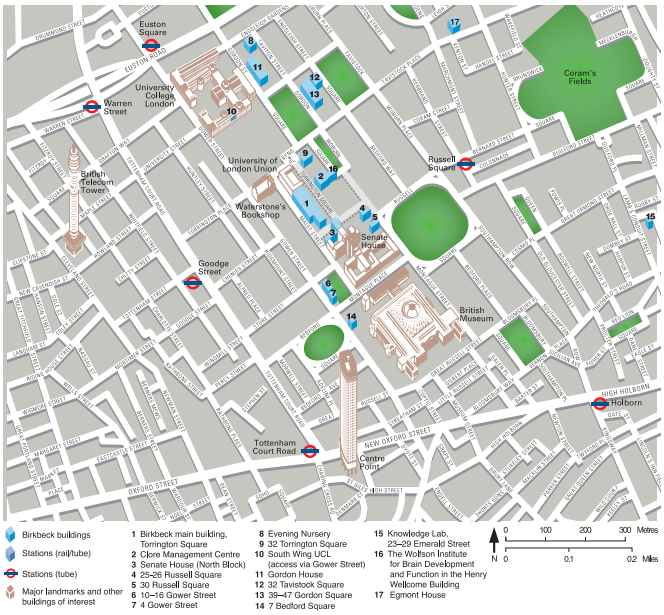



Figure 1: Birkbeck's Bloomsbury Campus in Central London.

## I. Introduction

The precipitously increasing amounts of detailed information generated by wifi and other mobile communication technologies, provide an invaluable opportunity to study different aspects of presence and movement behaviours of people within a given environment such as an organisation office complex or a university campus. Moreover, the pervasiveness of these technologies increases people's ability to access information, which undoubtedly influences the way the observed environment operates, and it is therefore essential that we develop the theoretical frameworks and the real-time monitoring systems in order to correctly understand how the presence of people and its dynamics reshape the structures of such environments.

With the aid of such tools, we can potentially discover hidden patterns of behaviour at both the collective and the individual user levels, thus increase our understanding about people's presence, and in turn, improve our ability to make informed decisions when we plan for our environments.

We summarise our contributions as follows:

1) Propose a density-based clustering method that discovers social groups by utilising activity traces of mobile users. We detect the social groups on the basis of the activities taking place at observed locations within a university campus. We provide a detailed description of this clustering method in Section III.
2) Develop a framework for inferring the type of an observed location, using the patterns of visit extracted from wifi activity traces recorded at that location. Here we have two main types of social activities: *formal* and *informal*, which we define in the next section II-E.
3) Investigate the similarities and differences between the *formal* and the *informal* social locations.

In the remainder of this article, the term *social activity* is used interchangeably with the term *event* to mean the same thing. Similarly, the terms *user* and *data point* are exchanged

with each other to mean the same thing.

## II. Characteristics of social activities within our case study environment

Understanding the social dynamics within an observed environment such as a university campus can be useful for a range of applications. In this paper, we study the social aspect of human presence with the aim of gaining a better understanding of the presence and movements of people within the case-study environment - the Bloomsbury campus of Birkbeck, University of London. By knowing *who* and *where* and *why* people spend their time, the university can plan for the most effective usage of space and allocation of services in the manner that creates a more positive attitude toward learning, and provides a richer and more rewarding experience.

### A. The case study environment

Birkbeck, University of London, the case-study university in this research, is a major provider of evening higher education in London. There are approximately 16,500 students attending the university, with approximately 88% of them enrolled on part-time programs [3]. Bloomsbury campus, the university's main campus location in central London, is situated adjacent to campuses of other universities, such as School of African and Oriental Studies (SOAS) and University College London (UCL). The close proximity to these other universities allowed a large amount of collaboration between them and as a result, Birkbeck's main campus is visited by thousands of students, researchers and academics on a daily basis.

*1) Eduroam data:* As a participant in Eduroam, a WLAN service developed for the international education and research community, Birkbeck offers secure roaming access to the Internet to all registered visitors [4]. The WLAN activity traces accumulated in the period from the 19th of September 2016 to 17th of December 2017, comprises 204,566 users and 223 locations that are divided between 11 sites within the main campus. Fig. 1 illustrates the location of Bloomsbury campus in central London.

### B. Types of Social Behaviour

The numerous daily activities that take place at the case-study environment, which include "learning classes", "meetings", "seminars" and "having lunch at the cafeteria", can be broadly divided into two main categories: *formal* and *informal* activities. Generally, in a *formal activity*, such as a learning class or a seminar, the social interaction is between a large group of individuals taking part in the activity, whereas in an *informal activity* we tend to find a close social interaction between a relatively smaller group of individuals. Moreover, individuals usually spend roughly the same duration of time when they attend a formal activity session whereas they tend to spend variable length of time when they are involved in an informal activity. Also, formal activities are usually linked to specific locations and appear to follow a regular pattern of occurrence whereas informal activities tend to not adhere to a fixed pattern of occurrence.

Generally, these two categories of activities underpin the different types of social behaviour that can be found at our chosen environment. In this research, we distinguish between two kinds of social presence: *formal* and *informal*, which we interpret as follows:

*a) Formal Social Presence:* is defined as the set of meetings that are attended by the same group of individuals, take place at the same location and occur regularly in sessions of fixed duration. For example: a three hour weekly lecture that take place at a specific lecture-room. We refer to the type of social relationship exhibited in such set of meetings as *formal social relationship* and the social group of users, who participate in such a relationship, as *formal social group*.

*b) Informal Social Presence:* is defined as the set of meetings that are attended by the same group of individuals and may take place at different locations. In contrast to the meetings of the *Formal Social Presence*, these meetings do not necessarily follow a regular patterns of occurrence or have a fixed duration. For example: having coffee at the cafeteria. We refer to the type of social relationship shown in such meetings as *informal social relationship* and the social group of mobile users, who take part in them, as *informal social group*.

*c) Visit:* We use the term *"visit"* to refer to an event when the time and the location of a particular user is recorded. This means that a user was at a specific location (i.e. a room) when they either initiated or received data using their mobile device over wifi.

### C. Types of Visited Locations

Unlike localisation techniques, which focus on discovering the exact location of the mobile device, in this research we are only interested in determining whether two, or more, devices are within the same room. We selected two types of locations for the evaluation of our proposed method: meeting rooms, where regular learning and administrative activities take place and leisure locations with food and drinks facilities. The details of these locations are given in Table I.

Table I: Selected Birkbeck Locations

| **Location** | **Site** | **Category** | **Number** of **Visitors** |
|---|---|---|---|
| Bar | Malet St. Ext. | Leisure (informal) | 4677 |
| Cinema | 43 Gordon Sq. | Leisure (informal) | 3035 |
| CoffeeShop | 43 Gordon Sq. | Leisure (informal) | 2967 |
| CoffeeShop | Malet St. | Leisure (informal) | 38520 |
| Room 102 | 10 Gower St. | Learning (formal) | 9963 |
| Room 301 | Malet St. | Learning (formal) | 4249 |
| Room 314 | Malet St. | Learning (formal) | 7076 |
| Room 413 | Malet St. | Learning (formal) | 665 |
| Room 417 | Malet St. | Learning (formal) | 189 |
| Room B29 | Malet St. | Learning (formal) | 16081 |
| Room 254 | Malet St. Ext. | Learning (formal) | 19051 |
| Room 456 | Malet St. Ext. | Learning (formal) | 12031 |

*1) Patterns of Visits:* We studied the number of revisits made to locations across campus and we observed that the distributions follow a power law for most locations. Fig. 2

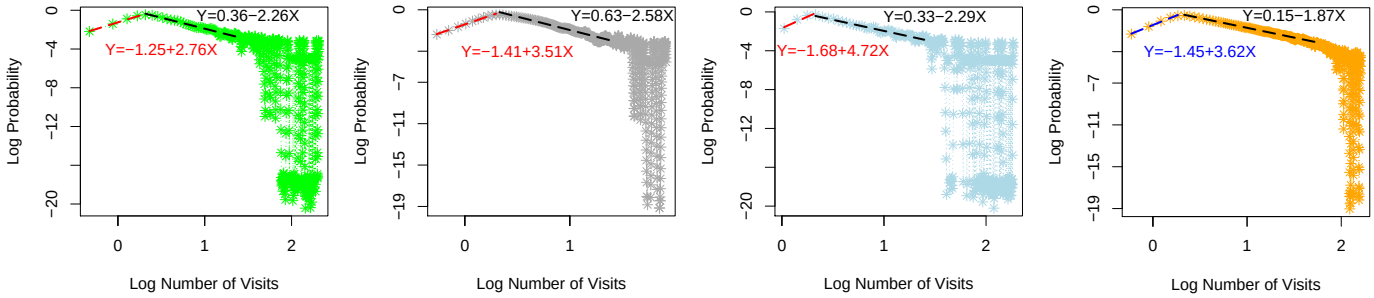



Figure 2: Distributions of number of revisits to the locations where *informal activities* occur. A revisiting user is one who made two or more visits to an observed location. Shown from left to right are the distributions for: the Coffee Shop, the Cinema at 43 Gordon Square, the Bar and the Coffee Shop at Malet Street. The two fitted straight lines indicate the broken power law relationship in each plot.

plots the distributions for the number of revisits made to locations where *informal activities* occur: the Coffee Shop and the Cinema at 43 Gordon Square, the Bar and the Coffee Shop at Malet Street. The log-log plots in this figure unanimously show that for the range up to about 25 revisits, the distributions follow a broken power law consisting of two power law regimes. Initially, for the first two visits, the distributions climb to their peak points at slopes 3.51, 2.76, 3.62 and 4.72, respectively. Then, for up to 25, 31, 25 and 63 visits, they descend gently at slopes -2.58, -2.26, -1.87 and -2.29, respectively. The four distributions jitter sharply for values of revisits beyond these ranges, which is a sign of an exponential cutoff. Interestingly, those individuals who made their first visit are more likely to revisit the observed location. This pattern suddenly reverses across the four locations, where for those individuals who made between 3 to 25, 31, 25 and 63 visits, respectively, the higher the number of their previous visits the less likely that they will revisit the observed location.

### D. Detecting Different Types of Social Presence

Our intuition is that the activity taking place at an observed location determines, to a large extent, the kind of social interaction that occurs during the activity. Methods that only capitalise on temporal and spatial information to detect social groups of people visiting an observed location, may not always produce the desired accurate results. For example, during a formal meeting or a seminar, people may be seated far from one another despite being closely related to each other. Equally, they may be seated adjacent to one another despite the lack of a close relationship between them. A method that solely depends on proximity information to detect the social group attending a meeting or a seminar in which individual people are placed at distances greater than what is required to link them to one another, will most probably fail to detect the correct social grouping. Similarly, a clustering method that relies on a small distance between arrival times, will fail to correctly cluster two individuals that attended a meeting but arrived at times far apart from one another. Equally, a method that expects individuals' arrival times to be long apart from one another, will fail to detected social event that occur within shorter time intervals. For example a method designed to detect groups that attend social events in which individuals arrive an hour apart from one another, will most likely fail to discover short events such as a 15 minutes coffee-break meeting at the cafeteria. We argue here that in order to detect the correct social behaviour at a given location, it is imperative that, in addition to the temporal and spacial information, we take into consideration the semantic underpinning of the social interaction at that location. For example, a clustering method that adapts to different social activities will be able to adjust its temporal and spacial criteria in order to correctly detect the social group attending such meetings. Our proposed clustering method, which we discuss in the Section II-E, is parameterised with information about the kind of activity that take place at an observed location.

### E. Social Density-based Clustering

Building on the previously mentioned intuition, we propose a new scalable method that detects the social clustering of mobile users on the basis of the type of activity performed at an observed location. Given a database of users and a set of locations, we would like to discover the groups of users that visit these locations to participate in a social activity. For example, we would like to discover groups of students who attend lectures together as classes at different lecture-rooms, groups of researchers who hold regular seminars at particular meeting rooms or groups of friends who socialise at the Coffee Shop during break time.

In order to formulate how we would discover such social groups we would like to introduce the following notation:

The core concept of the proposed algorithm (SocialDBC) for social clustering is that a data point is assigned to a cluster/group if it is *socially-connected* to all the other member points of the cluster or the group. To explain this key idea, we give the following definitions of concepts that are common to many density-based clustering algorithms such as DBSCAN [14]:

Given a data set of points $D$, the proposed SocialDBC algorithm estimates the density around $p$ using the concept of $\epsilon$-restricted-neighbourhood, which is defined as follows:

Table II: Notation

| | |
|---|---|
| $U$ | Database of users. |
| $L$ | The set of locations. |
| $p, q, r$ | An $m$-dimensional point representing a user's set of visits to the locations given in $L$. |
| $v$ | A user's visit, to a given location, within a time interval $t$. |
| $D$ | The set of $m$-dimensional points representing the users in $U$. |
| $\theta_{q,r}$ | The Jaccard distance between $q$ and $r$. |
| $RN_\epsilon(p)$ | The neighbourhood of $p$ in which the maximum distance between any pair of points is $\epsilon$. |
| $\delta$ | The minimum number of joint visits |
| $A_G$ | The set of joint visits of all users in the set $G$. $A_G = \bigcap_{p \in G} p$. |
| $minPts$ | A density threshold. |
| $G$ | A social group of users. |

*Definition 2.1:* $\epsilon$-restricted-neighbourhood
An $\epsilon$-restricted-neighbourhood, $RN_\epsilon(p)$, is the neighbourhood of $p$ in which the maximum distance between any pair of points is $\epsilon$. This is formally defined as follows:

$$RN_\epsilon(p) = \{q, r \mid \theta_{q,r} \leq \epsilon\}. \quad (1)$$

Note that the point $p$ is always a member of its own $\epsilon$-restricted-neighbourhood, i.e., $p \in RN_\epsilon(p)$ always holds.

Given the above definition, one can see that the neighbourhood $RN_\epsilon(p)$ is a subset of the $\epsilon$-neighbourhood $N_\epsilon(p)$, in which any pair of points are within a maximum distance $\epsilon$, i.e.

$$|RN_\epsilon(p)| < |N_\epsilon(p)| \quad (2)$$

*Definition 2.2:* Core Points
A point $p \in D$ is classified as:

1) a *core point* if its neighbourhood $RN_\epsilon(p)$ has high density, i.e., $|RN_\epsilon(p)| \geq minPts$, where $minPts \in Z^+$ is a user-specified minimum density threshold,
2) a *noise point*, otherwise.

*Definition 2.3:* Social Connectivity
A point $p$ is *socially connected* to every point $q \in G$ if:

1) $p$ is a core point, and $q \in RN_\epsilon(p)$,
2) $|A_G| \geq \delta$, see Table II.

*Definition 2.4:* Social Groups
A social group $G$, is a *socially connected* set of points. An obvious example of a social group is the class of students that attended the same learning sessions, which are at least equal to $\delta$ sessions in total. Such a group is socially connected because every member of the group attended at least $\delta$ sessions that the other members attended irrespective of whether the sessions took place at one or several locations.

### F. Detection of social groups

SocialDBC uses the concept of $\epsilon$-restricted-neighbourhood and the thresholds: $\delta$, $\epsilon$ and $minPts$ to classify the points given in $D$ into *core* and *noise* points. Consequently, it links those core points that are *socially connected* into social groups. Fig. 3 illustrates the concepts of $\epsilon$-restricted-neighbourhood, the two classes of points: core and noise, as well as points multi-cluster membership.

Algorithm 1 gives the pseudo code for SocialDBC, which starts by declaring an empty set of core points (line 1). It then performs three tasks for each point given in $D$: it computes the neighbourhood $RN_\epsilon(p)$, if $p$ satisfies the requirement for core points, it adds $p$ to the set of core points, and then it declares that $p$ is assigned to none of the social groups by setting the set of ids, belonging to $p$, as being empty (lines 3-9).

In the next step, for each core point with no cluster assignment, SocialDBC finds the set of socially connected points for the given core point (line 12). If the detected set size is greater than or equal to the threshold $minPts$, the set is identified as a social group and as a result the set of ids associated with each point in the social group is amended to indicate that the point is a member of the newly detected social group.

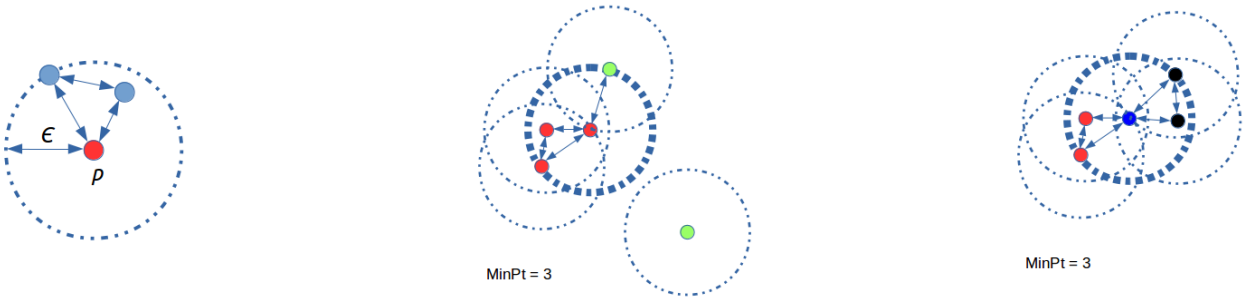
A core point may be socially connected to other core points in other social groups. Such a point will be added to all of those social groups. Any point that has not been assigned to a social group is considered to be noise.

### G. SocialDBC vs DBSCAN

A major distinction between the proposed SocialDBC algorithm and the many DBSCAN versions that exist in the literature is that the former discovers only convex clusters of points. A fundamental concept of the social grouping discussed in this research is that detected social groups must not include *a-friend-of-a-friend* relationships, which DBSCAN inherently allows through the creation of elongated non-convex clusters.

Another subtle difference between the two methods manifests in how multi-cluster participation is perceived. Overlapping of clusters conforms with how social grouping is defined in this research, where an individual can be a member of multiple social groups irrespective of the type of social interaction. While the social groups discovered by SocialDBC are not exclusive, i.e. SocialDBC permits the participation of points in multiple clusters, DBSCAN produces exclusive clusters where overlapping is not permitted.

One important feature of the SocialDBC method is the two level computation of the distance between two points. In addition to the usage of Jaccard distance to find the neighbourhood of a given point, we apply a minimum number of visits threshold to filter out those neighbouring points that do not belong to the social group. For example, to detect the group of students that attend the same class, we first find all the students that are part of the neighbourhood of an observed student. To do this we compute the Jaccard distance between the set of locations that the observed student visited and the set of visited locations of each of the students recorded in the database [1]. From the obtained neighbourhood we further filter the group of students that made joint visits greater than or equal to a minimum threshold of joint visits. This group of




(a) (b) (c)

Figure 3: (a) The $\epsilon$-restricted-neighbourhood of $p$. (b) Core and noise points. (c) Multi-cluster membership. The unlabelled double-sided arrow represents a distance that is less than or equal to $\epsilon$. The three red points in the sub-figure (b) are *core* points whereas the ones coloured in green are classified as *noise*. In the sub-figure (c), the point coloured in blue is a member of two clusters: the red and the black clusters of points.

**Algorithm 1** Social Density-based Clustering

**function** $SocialDBC(D, \epsilon, \delta, minPts = 2)$
  Core $\leftarrow \phi$
  **for each** $p \in D$ **do**
    Compute $RN_\epsilon(p)$
    **if** $RN_\epsilon(p) \geq minPts$ **then**
      Core $\leftarrow Core \cup \{p\}$
    **end if**
    $id_p \leftarrow \phi$
  **end for**
  $k \leftarrow 0$
  **for each** $p \in Core$ **do**
    $G \leftarrow$ FindSocialGroup($p$, $\epsilon$, $\delta$, $minPts$)
    **if** $|G| \geq minPts$ **then**
      $k = k + 1$
      **for each** $p \in G$ **do**
        $id_p \leftarrow id_p \cup \{k\}$
      **end for**
    **end if**
  **end for**
  $Groups \leftarrow \{G_i \mid G_i = \{p \mid p \in D,\ i \in id_p\}\}$
  $Noise \leftarrow \{p \in D \mid id_p = \phi\}$
**return** $Groups$, $Noise$

**function** $FindSocialGroup(p, \epsilon, \delta, minPts)$
  $G \leftarrow \phi$
  $A_G \leftarrow p$
  **for each** $q \in RN_\epsilon(p)$ **do**
    $\psi_i \leftarrow A_G$
    $\psi_{i+1} \leftarrow \psi_i \cap q$
    **if** $|\psi_{i+1}| \geq \delta$ **then**
      G $\leftarrow G \cup \{q\}$
      $A_G \leftarrow \psi_{i+1}$
    **end if**
  **end for**
**return** $G$

students that meets the joint visit criterion is considered to be a social group.

The key limitation, which both methods share, is the sensitivity of the result of clustering to the value of $\epsilon$, specially when the underlying clustering that we seek to discover has a wide range of density values.

The two methods have similar complexity due to the computation of the neighbourhood for each point in the data set. Thus, the worst-case complexity for SocialDBC is $O(n^2)$.

## III. Modelling Social Presence

### A. The Social Presence Model (SPM)

We propose the SPM model, which classifies locations into *formal* and *informal* locations on the basis of the visiting patterns detected at those locations. We have learnt so far how social groups can be detected using spacial and temporal information extracted from wifi activity traces and we would like to formulate a model that exploits these visiting patterns to predict the type of location where people socialise.

Based on our definition of *formal social presence* II-B0a, the visits made to an observed location by the same social group represent a set of uniformly distributed points in the visit space. Consequently, for each social group we can test for a discrete uniform distribution applied to the group's set of visits, recorded at the observed location. To illustrate the idea, we proceed as follows.

Given a location $l$, for each detected social group, we compute the length of the time period between each visit and the next. The data set made of these period lengths can be regarded as a sample $s$, which we hypothesise to be uniformly distributed. Formally, for each social group that visited the location $l$ we find the set of visits $v_1, \ldots, v_n$, arranged in chronological order. We compute the number of days between each two consecutive visits to create the set $s$. We denote the set comprising all the sets of in-between visits gaps for the current location as $S$, thus $|S|$ denotes the number of social groups that visited the observed location.

Assigning $l$ to the class of *formal locations* can be estimated by counting how many sets $s \in S$ are approximately uniformly distributed. Therefore, the probability of the observed location $l$ being classified as a *formal location* can be computed as the proportion of the number of uniformly distributed sets $s \in S$, compared to the number of social groups that visited the observed location.

$$\Pr\big(Y = formal\big) = \frac{\sum_{s \in S} I(s \sim U(a,b))}{|S|} \tag{3}$$

where $I$ is an indicator function that has value 1 only when its argument is true, and 0 otherwise. $a$ and $b$ are the minimum and maximum number of days between two consecutive visits.

Since we only have two types of locations: *formal* and *informal*, classifying a location as *formal* corresponds to predicting that its type is *formal* if $\Pr\big(Y = formal\big) > 0.5$, and *informal* otherwise.

To verify the uniformity of $s \in S$, we use the following hypotheses:

$H_0$: The periods lengths in $s$ are uniformly distributed.
$H_1$: The periods lengths in $s$ are not uniformly distributed.

In order to test these hypotheses, we compute the chi-square goodness of fit statistic as shown below [24].

$$T = \frac{\sum_{i=1}^{d}(O_i - E_i)^2}{E_i} \approx \chi^2_{d-1} \tag{4}$$

where $O_i$ is the observed count of the period length $i$, $E_i$ denotes the expected count, $E_i = \frac{1}{|s|}\sum_{i=1}^{d} O_i$, and $d$ is the number of count values $O_i$ based on the observed $s$.

### B. Baseline Model

We use a multiple logistic regression model as a baseline model for comparison. The model directly infers the type of an observed location based on a set of features, which describe each social group that attended the location: the size of the group, number of visits made by the group, minimum and maximum number of days between two consecutive visits. It is a global model in the sense that the model is fitted using information from all formal and informal locations in our data. We estimate the probability of whether an observed location can be classified as *formal* or *informal* using the following equation.

$$\Pr\big(Y = formal\big) = \frac{e^{\beta_0+\beta_1 x_1+\beta_1 x_2+\beta_p x_3+\beta_p x_4}}{1 + e^{\beta_0+\beta_1 x_1+\beta_1 x_2+\beta_p x_3+\beta_p x_4}} \tag{5}$$

The maximum likelihood method is used to estimate the parameters $\beta_0,\ \beta_1, \ldots,\ \beta_4$. $x_1,\ x_2,\ x_3\, x_4$ denote the size of the group, number of visits made by the group, the minimum and the maximum number of days between two consecutive visits, respectively.

An observed location is classified as *formal* if $\Pr\big(Y = formal\big) > 0.5$, and *informal* otherwise.

## IV. EXPERIMENTAL EVALUATION

### A. Data Processing

To detect students social groups, which we subsequently used in the evaluation of our proposed models, the raw eduroam data was processed to create *m*-dimensional points. Each point denotes the *visits* made, by one of the users, to different locations across the university campus.

*1) Data Privacy:* All sensitive personally identifiable information, such as the user's email address, have been removed from the data set. All other items of personally identifiable information of the participants in this study, notably the device MAC address, have been replaced with pseudonymised identifiers. Information relating to physical locations, such as the locations associated with specific eduroam access point BSSIDs, has not been anonymised. However, we do not reveal these locations, for example by placing them on a map, and we ensure that access information is aggregated by BSSID but not per user - specifically no attempt has been made to create individual user fingerprints. The specific provisions followed for data collection, management and processing, and associated security provisions have been approved by the university's research ethics committee as compliant with our research integrity code of practice (for more details see [12]).

### B. Models Evaluation

In this section, we evaluate the proposed SPM and baseline models on the eduroam data set, which we describe in section II-A1. We are particularly interested in the predictive performance of the models, i.e., given the information about the visits made by different social groups, our goal is to accurately predict the type of each location visited.

*1) Evaluation metrics:* To measure the performance, we consider the mean prediction accuracy as an evaluation metric, i.e. an accuracy of 0.1 means that only 10% of the time the proposed model successfully predicts the correct type of the observed location. Also for each model, we provide a table of confusion (a confusion matrix) to report the number of false positives, false negatives, true positives, and true negatives. The significance level of 0.01 is used for performing the statistical hypothesis testing.

*2) Experimental setup:* Our experiments were based only on detected social groups which visited the set of locations given in Table I. These social groups were detected using our proposed clustering method SocialDBC. Each social group had a least two visits to same observed location. To evaluate the accuracy of the SPM model, we used all the data without division into training and testing data sets. However, for the evaluation of the baseline model we divided the data into training and testing sets, where from each location we used 80% of data for training and the remaining 20% for testing.

### C. Uniformity of Social Presence

Our initial intuition is that the locations associated with formal activities are visited in a regular manner by social groups with uniform periods between visits. In contrast, those locations that are linked to informal activities have irregular

patterns of visits. To verify this intuition, we evaluated the proposed models, described in III, on the visit data of each of the locations given in Table I

Table III: Table of Confusion

| | Predicted Location Type | Actual Location Type: Formal | Actual Location Type: Informal |
|---|---|---|---|
| SPM | Formal | 8 | 3 |
| | Informal | 0 | 1 |
| Baseline | Formal | 0 | 1 |
| | Informal | 8 | 3 |

Table. III reports the number of false positives, false negatives, true positives, and true negatives from the evaluation of these models. The reported results show the superiority of the SPM over the baseline model, which it outperforms by a factor of three in terms of accuracy: 0.75 and 0.25 for SPM and baseline models respectively. The SPM model correctly classified all the formal locations whereas the baseline model failed to correctly classify any of them. Interestingly, the baseline model correctly classified three out of the four informal locations where as the SPM model only classified one. Fig. 4, plots, per location, the distribution of social groups into uniform and non-uniform on the basis of the results obtained from the SPM and the baseline models, as shown in the sub-figures (a) and (b) respectively. On the one hand we see in the sub-figures (a) that the SPM model detected the uniformity in the visiting behaviour of the social groups that attended the formal locations, but on the other hand we see that the model also detected a similar visiting behaviour from the social groups that attended the informal locations. For example, the Coffee Shop at Malet Street, which the model incorrectly classified as a formal location, we have a high number of social groups that made uniform visits to the location. We also observe a similar result for the Coffee Shop at 43 Gordon Square, which is not a location where *formal* activities occur, but nonetheless we see that the location was visited in a regular manner by a significant number of social groups. One interpretation of such results is that many social groups visit these two locations at lunchtime and in coffee breaks during lectures and other learning sessions. Since break times are usually dictated by the teaching timetable, it is not strange that we observe a uniform visiting behaviour for many of the social groups that attend these locations. The visits made to the Bar at Malet Street, which was classified correctly by the model as an informal meeting location for social groups, do not seem to be dictated by the teaching timetable. This is most probably because people rarely socialise there during the teaching hours.

### *D. The Geographical Spread of Visits Across Campus*

We studied the number of locations visited by social groups across campus and we discovered that around 83% of the detected groups visited only one location to socialise. Fig. 5, shows the distribution of the number of locations visited by

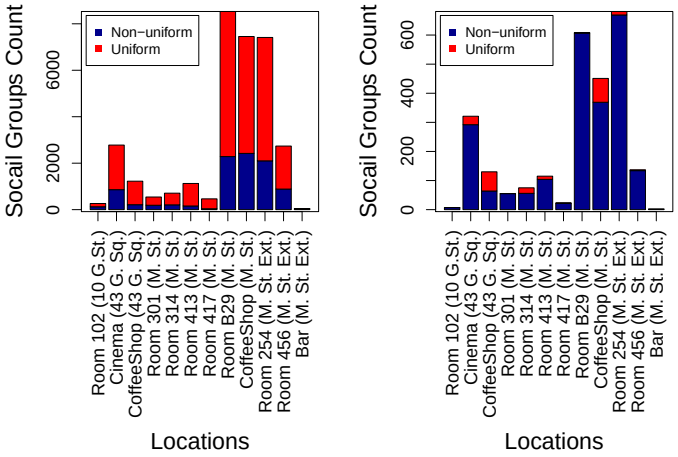



Figure 4: Distribution of count of social groups and how likely that the revisits made to the observed locations follow a uniform or a non-uniform distribution. The distributions are based on the predictions made by (a) the SPM model and (b) the baseline model. The plotted location names correspond with the names given in Table I.

social groups across locations where informal activities occur. We also examined the number of locations visited by social groups that attended the Coffee-shop and the Bar at Malet Street and for each group we counted the number of visited locations from other sites of the campus, i.e. places located offsite Malet Street. As shown in Fig. 6, 91% and 99% of social groups that visited the informal locations at Malet Street and Gordon Square, restricted their visits to nearby locations, i.e. locations within the same site, as opposed to locations that are further afield. One interpretation of such result is that many social groups visit *informal locations* at lunchtime and in coffee breaks during lectures and other learning sessions. These breaks usually last for short periods, and consequently do not provide enough time for groups to socialise offsite far from their prime location of work or study.

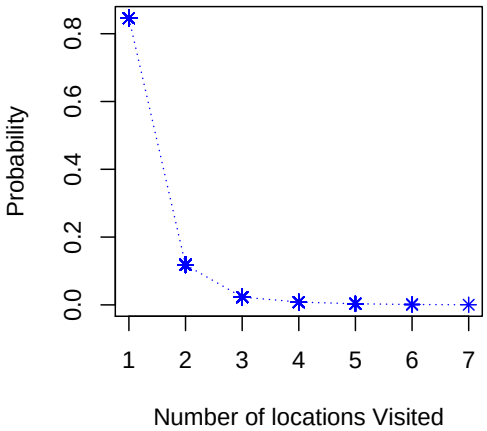
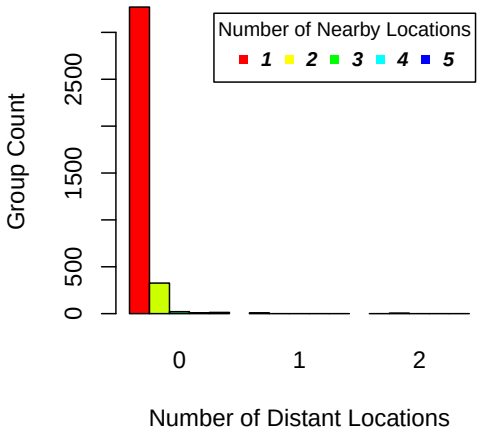
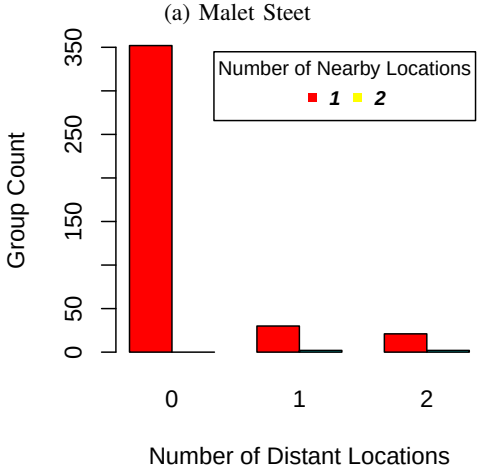



Figure 5: Distribution of number of locations visited by social groups detected across all locations.




(a) Malet Steet




(b) Gordon Square

Figure 6: Number of distant locations visited by social groups that visited Malet St and Gordon Sq informal locations. In this experiment, a distant location is any Birkbeck location excluding the ones situated at Malet St and Gordon Sq.

### *E. Visiting Behaviour Across Locations*

Intuitively, formal locations, where activities such as learning classes and lab sessions take place, are usually attended by groups as opposed to individual users. To find out whether users visit a given location as a group or individually, we calculate the *social weight*, which compares the number of shared visits made by the social group to the total number of visits made by the individual user, including the visits they made with their social group:

$$SocialWeight = \frac{Number\ of\ group\ visits}{Number\ of\ individual\ user\ visits} \quad (6)$$

In ideal settings, a *social weight* value that is equal/close to 1 demonstrates the superiority of group visits over the individual user visits. In contrast, a significantly smaller value is a clear indication that the user prefers to visit the observed location as an individual as opposed to visiting it with a group. Fig. 7 illustrates such scenarios where the skewness of the distribution indicates the superiority of one type of visiting behaviour over the other.

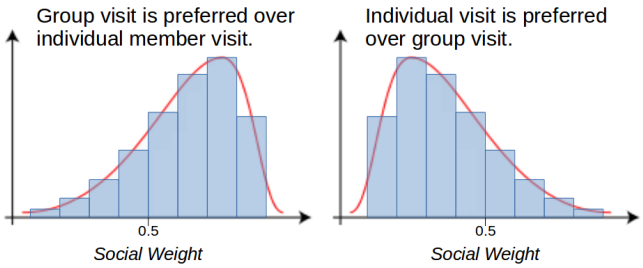



Figure 7: Types of visiting behaviour as seen through the distribution of ratio of number of group visits compared to the number of individual member visits.

As shown in the Fig. 8 and Fig. 9, the *social weight* value varies from one observed location to another but generally those locations that are linked with formal activities seem to be favoured by social groups as opposed to individual users. With exception of the distribution for Room B11 at 43 Gordon Square site, it is clearly evident from the negative skewness of the peaked distributions shown in Fig. 8 that more visits were made, to these locations, by social groups as opposed to individual users. Although the distribution for Room B11 has a positive skewness but the social weight values shown range between 0.6 and 1.0, which clearly indicates that the location was visited by groups of users more than it was visited by individual users.

Similar to formal activity locations, most of the observed locations associated with informal activities seem to have the group behaviour of visit as the favoured mode of visit. As shown in Fig. 9, the negatively skewed and highly peaked distributions for locations such as the Coffee Shops suggest that they are preferred locations for social groups. Despite the positive skewness of distribution for the Bar at Malet Street Extension, the social weight values shown are greater than 0.5, which strongly indicates that the location was visited by groups of users more than it was visited by individual users. The Cinema at 43 Gordon Square seems to have a large proportion of its visits made by individual users but it nonetheless remains a favoured destination for social groups.

## V. Related Work

Numerous research investigated the possibility of exploiting wifi traces in order to obtain an up-to-date view of the human presence within an academic environment. We review some of these works in relation to the four data aspects: the social, the spatial, the temporal and the semantic aspects.

The works presented in [2] and [22] investigated how density-based clustering of WLAN traces can be utilised to discover social groups of students within a university campus. In these research works, information extracted from the timetable as well as the teaching practices at the case-study university, was leveraged to inform the proposed models about the patterns of visit made to targeted locations. Although such

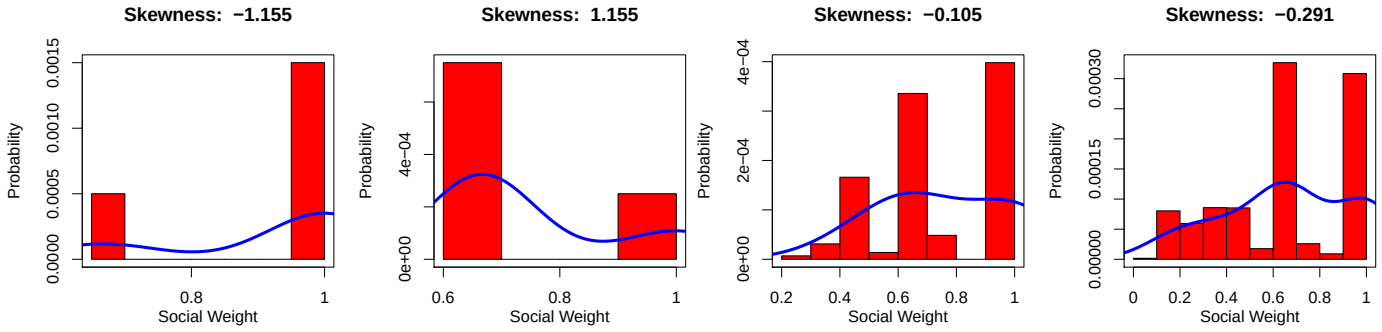



Figure 8: Distributions of *social weight* for formal activity locations. Shown from left to right are the distributions for: (a) Room 102 at 10 Gower Street, (b) Room B11 at 43 Gordon Square, (c) Room 314 at Malet Street and (d) Room 254 at Malet Street Extension.

extra information increases the accuracy of the obtained results, it limits the application domain of the proposed methods to environments, where such kind of information is available. This limitation is addressed in this research work, where the newly proposed clustering method does not depend on such information to detect social groups.

In a study discussed in [16] it was shown that it is possible to identify social groups amongst the observed set of mobile users. The study was based on WLAN mobility traces that were collected over a period of one month. A key finding in this study was the significant difference found between male and female session duration. In [9], which discusses another study involving university students, Eagle and Pentland identified activity patterns related to the users daily behaviour. They further discovered that the daily patterns can be associated with the user's major of study and, in turn, linked to the level of employment.

The article given in [21] describes a study in which the data set was collected in early 2003 over a period of only one week. The carefully designed study was used to investigate the usage of the network before an expansion plan was drawn. The key objective of the research was to find out information about usage the network. In [23], which is a relatively larger study as opposed to those aforementioned studies, the authors described the growth of the network from spatial and temporal perspectives. The research described in [17] estimated the long-term network usage among different *access points*, and predicted the distribution of future user locations in order to estimate the redistribution of loads among neighbouring *access points* at those locations.

All the studies discussed herein do not provide a comprehensive analysis of the four dimensions of the human presence within an academic environment. With exception to [22], the analysis provided in these studies overlooks one or more aspects of the human presence, in particular the semantic aspect, which has not been discussed in any of them. The analysis and the discussion presented in this paper is an attempt to bridge this gap. It is based on a large amount of wifi traces, recently collected at Birkbeck, University of London, which is one of the participant universities in Eduroam. Furthermore, this analysis provides a current perspective about the recent trend in Eduroam usage.

## VI. Conclusion

We investigated patterns of human social presence using a large data set of Eduroam activity traces. These traces were collected at a busy university campus in central London. We developed a clustering method that leverages on the type of activity performed at observed location in order to detect visiting social groups. We discovered that people generally socialise at a very small set of nearby locations within campus - within the same building or site. Generally, people visited a distant location, i.e. another Birkbeck site, when they were in the company of their social group. Our analysis also revealed that locations can be categorised into two main types: formal and informal locations. Based on the visiting behaviours exhibited at an observed location, our proposed model of human social presence (SPM) can infer the type of any observed location across the university campus. This seemingly simple model reliably predicts the type of visited location and outperforms the nontrivial baseline model by a factor of three.

## Acknowledgment

The authors would like to thank the members of Birkbeck IT Services who extracted the data set for us for the purpose of this research.

## References

[1] Charikar, Moses S.: Similarity estimation techniques from rounding algorithms. In Proceedings of the Thirty-fourth Annual ACM Symposium on Theory of Computing. ACM, pages 380-388, (2002)

[2] Eldaw, M. H. S., Levene, M., and Roussos, G.: Density-based Social Clustering for Mobile Users. In Proceedings of The 13th International Joint Conference on e-Business and Telecommunications, Volume 6: WINSYS, pages 52–62, (2016)

[3] Birkbeck in numbers, http://www.bbk.ac.uk/about-us/bbk/downloads/2014-articles/bbk33-53-numbers.pdf

[4] Eduroam at Birkbeck, http://www.bbk.ac.uk/its/services/wam/Eduroam

[5] Allahdadi, A., Morla, R., Aguiar, A., and Cardoso, J. S.: Predicting short 802.11 sessions from radius usage data. In: 9th IEEE International Workshop on Performance and Management of Wireless and Mobile Networks, pages 1–8. IEEE, (2013)

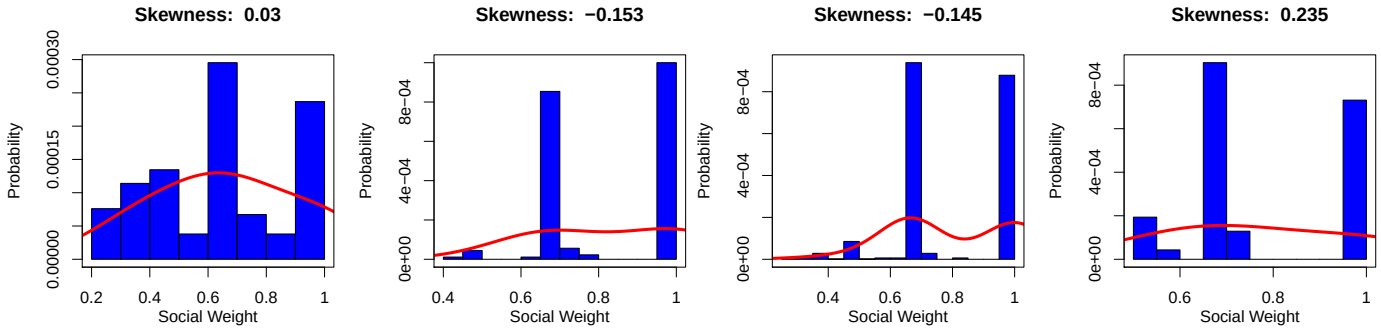



Figure 9: Distributions of *social weight* for informal activity locations. Shown from left to right are the distributions for: the Cinema, the CoffeeShop at 43 Gordon Square, the Coffee Shop and the Bar at Malet Street.


[6] Cheetham, A. H. and Hazel, J. E.: Binary (presence-absence) similarity coefficients. Journal of Paleontology, pages 1130–1136, (1969)

[7] Clauset, A., Shalizi, C. R., and Newman, M. E.: Power-law distributions in empirical data. SIAM review, 51(4): 661–703, (2009)

[8] Späth, H.: Cluster analysis algorithms for data reduction and classification of objects, (1980)

[9] Eagle, N. and Pentland, A.: Reality mining: sensing complex social systems. Personal and ubiquitous computing, 10(4):255–268, (2006)

[10] Eduroam, https://www.Eduroam.org/

[11] Birkbeck Rooms, http://www.bbk.ac.uk/roombookings/rooms

[12] Research Integrity at Birkbeck, http://www.bbk.ac.uk/committees/research-integrity

[13] Eldaw, M. H. S., Levene, M., and Roussos, G.: Presence analytics: Discovering meaningful patterns about human presence using wlan digital imprints. In Proceedings of The International Conference on Internet of Things and Cloud Computing, page 53, ACM, (2016)

[14] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd, volume 96, pages 226–231, (1996)

[15] Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A.: Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, volume 1(3), pages 231–240, (2011)

[16] Kumar, U., Yadav, N., and Helmy, A.: Gender-based feature analysis in campus-wide wlans. ACM SIGMOBILE Mobile Computing and Communications Review, volume 12(1), pages 40–42, (2008)

[17] Lee, J.-K. and Hou, J. C.: Modeling steady-state and transient behaviors of user mobility: formulation, analysis, and application, In Proceedings of the 7th ACM international symposium on Mobile ad hoc networking and computing, pages 85–96, ACM, (2006)

[18] Mulhanga, M. M., Lima, S. R., and Carvalho, P.: Characterising university wlans within Eduroam context. In Smart Spaces and Next Generation Wired/Wireless Networking, pages 382–394. Springer, (2011)

[19] Sneath, P. H.: The application of computers to taxonomy. Journal of general microbiology, volume 17(1), pages 201–226, (1957)

[20] Hamilton, James Douglas.: Time series analysis, Princeton: Princeton university press, volume (2), (1994).

[21] Schwab, David, and Rick Bunt. :Characterising the use of a campus wireless network. INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies. Vol. 2. IEEE, (2004).

[22] Eldaw, Muawya Habib Sarnoub, Mark Levene, and George Roussos. Social-DBSCAN: A Presence Analytics Approach for Mobile Users' Social Clustering. International Conference on E-Business and Telecommunications. Springer, Cham, (2016).

[23] Papadopouli, Maria, Michael Moudatsos, and Merkourios Karaliopoulos. Modeling roaming in large-scalewireless networks using real measurements. Proceedings of the 2006 International Symposium on on World of Wireless, Mobile and Multimedia Networks. IEEE Computer Society, (2006).

[24] Read, Timothy RC, and Noel AC Cressie. Goodness-of-fit statistics for discrete multivariate data. Springer Science  Business Media, (2012).