




# Leveraging WiFi network logs to infer student collocation and its relationship with academic performance

Vedant Das Swain<sup>1\*</sup> , Hyeokhyen Kwon<sup>1</sup>, Sonia Sargolzaei<sup>1</sup>, Bahador Saket<sup>1</sup>, Mehrab Bin Morshed<sup>1</sup>, Kathy Tran<sup>1</sup>, Devashru Patel<sup>1</sup>, Yexin Tian<sup>1</sup>, Joshua Philipose<sup>1</sup>, Yulai Cui<sup>1</sup>, Thomas Plötz<sup>1</sup>, Munmun De Choudhury<sup>1</sup> and Gregory D. Abowd<sup>1,2</sup>

\*Correspondence:

[vedantswain@gatech.edu](mailto:vedantswain@gatech.edu)

<sup>1</sup>Georgia Institute of Technology,  
North Ave NW, Atlanta, USA

Full list of author information is  
available at the end of the article

## Abstract

A comprehensive understanding of collocated social interactions can help campuses and organizations better support their community. Universities could determine new ways to conduct classes and design programs by studying how students have collocated in the past. However, this needs data that describe large groups over a long period. Harnessing user devices to infer collocation, while tempting, is challenged by privacy concerns, power consumption, and maintenance issues. Alternatively, embedding new sensors across the entire campus is expensive. Instead, we investigate an easily accessible data source that can retroactively depict multiple users on campus over a semester, a managed WiFi network. Despite the coarse approximations of collocation provided by WiFi network logs, we demonstrate that leveraging such data can express meaningful outcomes of collocated social interaction. Since a known outcome of collocating with peers is improved performance, we inspected if automatically-inferred collocation behaviors can indicate the individual performance of project group members on a campus. We studied 163 students (in 54 project groups) over 14 weeks. After describing how we determine collocation with the WiFi logs, we present a study to analyze how collocation within groups relates to a student's final score. We found that modeling collocation behaviors showed a significant correlation (*Pearson's*  $r = 0.24$ ) with performance (better than models of peer feedback or individual behaviors). These findings emphasize that it is feasible and valuable to characterize collocated social interactions with archived WiFi network logs. We conclude the paper with a discussion of applications for repurposing WiFi logs to describe collocation, along with privacy considerations, and directions for future work.

**Keywords:** Wireless sensor networks; Infrastructure sensing; Collocation; Social interactions; Student behavior; Academic performance

## 1 Introduction

Humans are social by nature; their functioning is related to behaviors that are interlinked with those of others [1]. One form of social interaction is the collocation of individuals

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

in the same space. Collocation provides the opportunity for synchronous interactions through multiple channels — voice, expressions, gestures and body posture — and for impromptu interactions that strengthen social ties. On campuses, understanding how students collocate can provide valuable insights to support academic success. Using traditional methods like surveys can be limited when trying to continuously evaluate how several individuals on a campus experience collocation. For these purposes, survey based approaches can often be obtrusive and do not scale to represent dynamic human functioning. Therefore, we need to identify a flexible approach to expressing collocation from an egocentric perspective.

The passive sensing community has introduced many automated and unobtrusive sensing methods to capture collocated social interactions [2–6]. However, approaches that require specialized client devices have limits [3, 4, 7]. Evolving mobile manufacturer specifications can critically disrupt long term data collection during mid-study with an update. In fact, the privacy concerns of installing such sensing firmware has also limited the meaningful data-streams available to researchers [8]. Insights on social behaviors require collective adoption from multiple socially related participants who must also consistently maintain the devices (e.g., keeping devices charged), thereby posing challenges to large-scale sensing and practical deployment of sensors. Together these factors challenge the scalability of such methods because they provide a sparse representation of the community.

An approach that mitigates some of the client-side challenges, is to use infrastructure-based techniques, such as installing Bluetooth beacons into the built environment [4, 9]. Nevertheless, these techniques can also rely on data being collected and processed through a client [6, 10]. Moreover, augmenting the entire infrastructure with new sensors can be expensive and, yet, cannot be used to inspect a history of campus-scale social behaviors (e.g., exam week, violent incidents, shutdowns, and global infectious disease-related pandemics). In contrast, many campuses maintain a managed WiFi access-point (AP) network that provides device association logs which can be repurposed to infer locations of users [11] and subsequently model individual behaviors [12, 13]. We posit that WiFi logs can be mined to express social interactions by characterizing how individuals collocate. We know that collocation presents avenues for social interactions that explain performance of individuals working in groups [14–16]. This aspect of human interactivity, is known as *spatiality* [17]. When individuals with a common intent gather in a space they interact via both verbal and non-verbal cues, that in turn influence individual functioning. We assess if WiFi logs meaningfully approximate a student's collocated social interaction patterns by testing if they represent a known outcome of such interactions — the student's performance [18–21]. Specifically, in this study, we investigate if collocation can explain the performance of students in project groups. We note that archival logs from managed WiFi network are a coarse descriptor of location (compared to other sensors they have lower spatio-temporal resolution). Yet, by examining if these logs can explain student performance, we seek to demonstrate the potential of a coarse sensor accessible to most modern campuses. Specifically, we pursue the following research goal: *To evaluate the ability of archival WiFi logs to retroactively express outcomes of collocated social interactions.*

The paper first provides a system description that elaborates how we determine collocation from association logs. Next, we explore if unobtrusively inferred collocation of

students in project groups can explain their individual performance. These students were distributed across 54 course project groups in a single course, and interacted over a 14-week period. Using statistical modeling approaches we examine if a student's collocation patterns are associated with an established outcome of social interactions—performance [14, 19, 22, 23]. Together, we present a case study that illustrates how WiFi association logs can model collocation patterns that capture signals of social interaction. Our findings encourage researchers to use their managed WiFi network as an instrument to explore collocated social interactions on campus. Our work can be extended to model many new aspects of collocation, such as where they occur (in classrooms, dorms, or cafes), who they occur with (familiar or unfamiliar peers), and for what purpose (other classes, extra-curricular activities, or serendipitous occurrences). Leveraging the archival logs enables us to test these hypotheses related to these behaviors at scale and over a long period of time. Lastly, the paper highlights applications of this data along with privacy and data-ethics concerns related to practical deployments.

## 2 Background and related work

When individuals with a common intent are in the same space at the same time, and are aware of it, they engage in some form of *collocated synchronous interactions*. Note, this paper adopts the definition for social interactions as, "...acts, actions, or practices of two or more people mutually oriented towards each other's selves..." [24]. Although these interactions can take place digitally, this paper focuses on automatically identifying synchronous social interactions in the physical world, i.e., the people interacting are collocated.

### 2.1 Collocation and performance

Literature on collocation describes the importance of intense interlinked activities in a dedicated physical space [17, 25] (e.g., "warrooms") as well as fluid activities in the presence of coworkers in a general physical space [14, 19] (e.g., open offices or adjacent cubicles). Both forms foster social interactions that are associated with individual and team performance [15, 16, 18, 25, 26].

Olson and Olson characterize multiple aspects of collocation at work and its implications [17]. Foremost, it is a synchronous social interaction that is not limited to verbal discussions and active sharing of resources. Even the presence of others working towards a common goal allows for subtle exchange of information through gestures and expressions [17] (e.g., is a teammate struggling, are they too absorbed or are they available for feedback). Additionally, collocation provides shared context that comprises common points of reference (e.g., whiteboards, post-it notes, or verbal concepts) [17]. Moreover, it supports informal interactions that can help "opportunistic information exchange" and improve social ties with teammates [17]. Prior work also posits several links between collocation and performance. Being physically situated in the same space keeps team members up-to-date, and therefore agile and innovative [14]. Staying collocated helps maintain common mental models of tasks, resources, skills, and problems [22]. In contrast, distance is known to elicit more conflict [26]. This is likely due to the non-uniform distribution of information that can lead to excluded members partaking in incomplete, inaccurate, or redundant tasks [27]. Distributed work is also related to heightened tensions between teammates, which affect wellbeing and impede individual performance [16]. On the other hand, collocation allows *team learning*, where members feel "safe" to seek feedback, experiment, and resolve errors [15]. Feedback from teammates is known to augment individual performance

[18]. Moreover, collocation can improve social ties between members [28] and therefore improve performance [23]. Related to performance, subtle cues of collocated social interactions are related to individuals focusing on single tasks for longer, continuous periods [19].

Traditional methods of evaluating collocated social interactions rely on survey instruments, but these are limited by recall and desirability biases [29, 30]. Moreover, self-reports are static assessments, while social interactions are fluid and vary over time [31]. One approach to studying human phenomena by avoiding such biases is with unobtrusive sensing. These automatic methods have the promise of dynamically sensing human behavior without interfering with an individual's natural functioning and are, therefore, more practical for gathering reliable insights.

## 2.2 WiFi-based sensing of collocation

Researchers have tried to determine collocation through sensors in the environment. For instance, WiFi-based fingerprinting can help identify ties between groups [10]. In such approaches, a user device can determine its own location by measuring the difference between signals from different Access Points (APs) [32]. Researchers would then need to install loggers on each individual's device to determine collocation. Deployment costs aside, to get comprehensive insights the client application will need to be trained over an entire network of APs and constantly update for persistence over long periods of time. Alternatively, enterprises have used WiFi router networks to develop Real-Time Location Systems (RTLS) [33, 34]. To infer location, these technologies store the Received Signal Strength Indicator (RSSI) values for any client-device within a neighborhood of APs. This could be extended to infer collocation but these solutions have a substantial cost for installation (requiring a full fingerprinting survey of the network). This, coupled with the privacy concerns of excessive precision, often outweighs the benefits of any realistic campus use-case. Yet, a common form of WiFi infrastructure deployment in university campuses [12, 13] only stores association logs describing which AP a client-device is connected to. Although it is relatively coarse [35], this parsimonious representation of location has been used to understand individual behavior — assist depression screening [12] and assign semantic tags to spaces [13]. These works motivate us to an approach that adheres to data minimization. While prior examples trace individual dwelling patterns across campus, few studies take an unsupervised approach to retroactively assess explicit social behaviors. We expand on such WiFi-based efforts to identify collocation between multiple students with a shared intent, such as a group project. Even though collocation does not necessitate verbal communication in the strict sense, it does serve a social function [17]. As discussed in Sect. 2.1, these social factors can affect the performance of collocated individuals. Therefore, to determine if our WiFi-based characterization of collocation meaningfully approximates social interactions, we inspect how it indicates performance. Prior work has explored various passively sensed phenomena as a proxy for social interactions between individuals. Mining WiFi network data can cluster people into social and behavioral groups [36, 37]. Even other infrastructure-based coarse location technologies, such as Bluetooth, have been used to capture subtle social interactions like synchrony within-group routines [9]. While these studies implicitly associate individuals together (e.g., distinguish students by dining hall), they do not explore collocation in physical spaces sufficiently. A recent work demonstrated how collocation can be used to predict stress by

harnessing the campus network infrastructure [38]. However, these systems either rely on additional augmentation of the infrastructure or knowledge of the network signal strength received by clients. In contrast, we examine human behaviors as evident in rudimentary raw network association logs that could be applied to almost any managed wireless network today.

Notably, De Montjoye et al., used WiFi-based collocation to approximate *instrumental* social ties [39]. However, their study does not describe how the logs were modelled. We also do not know if the WiFi-device infrastructure was engineered before the project semester, especially given project teams interactions were supervised by the researchers through the semester. By contrast, in our research both the participants and the infrastructure were completely unsupervised during the project semester. Therefore, through our system description we highlight many nuances of modelling such coarse data retroactively by triangulating multiple sources of archival data (e.g., raw WiFi logs, attendance records, and course grades). Additionally De Montjoye et al., assume collocation reflect instrumental ties but they do not validate these against any meaningful ground truth. Their approach leaves room for improvement. For instance, building-level collocation by De Montjoye et al., provides greater reliability but is likely lower in resolution. Consider students within the same majors, who visit the same building at the same time but for different classes. These students might appear to have strong ties because they appear collocated. However, they might also be students who are regular in their respective coursework and will do well in a team together regardless of how much they interact. Our research specifically tries to disentangle these individual behaviors from collocation by representing the room-level collocations. We also factor in when and where these collocations occur. Together, these aspects give stronger evidence that any existing WiFi infrastructure can be tapped to model its logs for collocated social interactions.

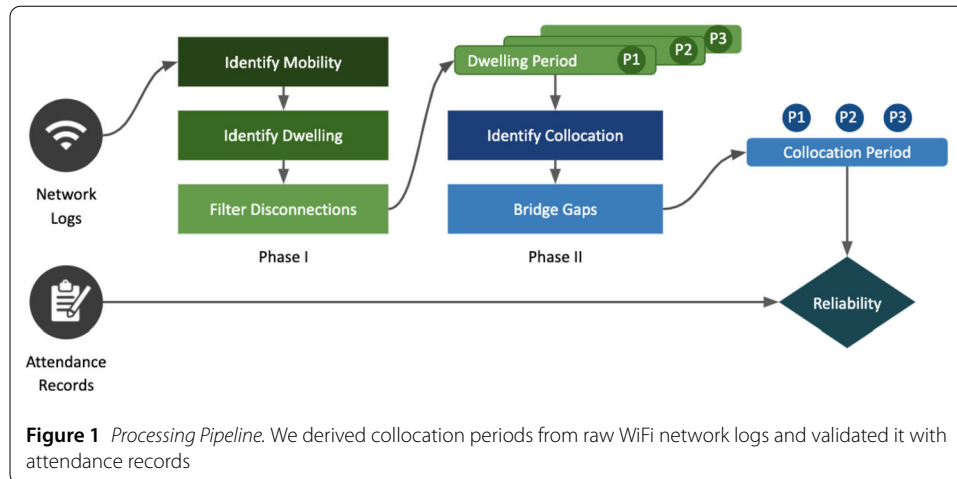
### 3 System description: identifying collocation with network logs

This study was done after the semester was over. Students' behaviors were not influenced by this study and their privacy was not compromised while they were enrolled in the course. Therefore, we use class attendance records to validate the system, as these can be obtained retroactively. This section describes a pipeline to determine collocation by leveraging WiFi network association logs and an evaluation of its reliability by comparing it to class attendance records (Fig. 1).

#### 3.1 Network data

##### 3.1.1 Sample association logs

We obtained consent from 46 students at a large public university in the United States, and then analyzed their anonymized WiFi association logs. These students belonged to two sections of a project-intensive course. Both sections were taught by the same instructor and had attendance data for each lecture. We refer to these sections as "1A" (22 students) and "1B" (24 students) throughout the paper. The instructor for the course provided each consenting student's attendance and group label, along with the course lecture schedule. We partner with the institute's IT management facility to obtain network log data for any device owned by a consenting student without requiring direct access to device MAC



**Table 1** Sample raw log

Field	Sample
Timestamp	Apr 1 00:10:51
Update Type	snmpupdate
Anon. User	2099
User Device	c4:7d:eb:0f:df:d5
AP ID	40:cd:14:b2:02:c0
AP Label	122S-209

addresses. This data was accessed at the end of the semester<sup>1</sup> and contains approximately 14 weeks of data, which spans 34 lectures for each section.<sup>2</sup>

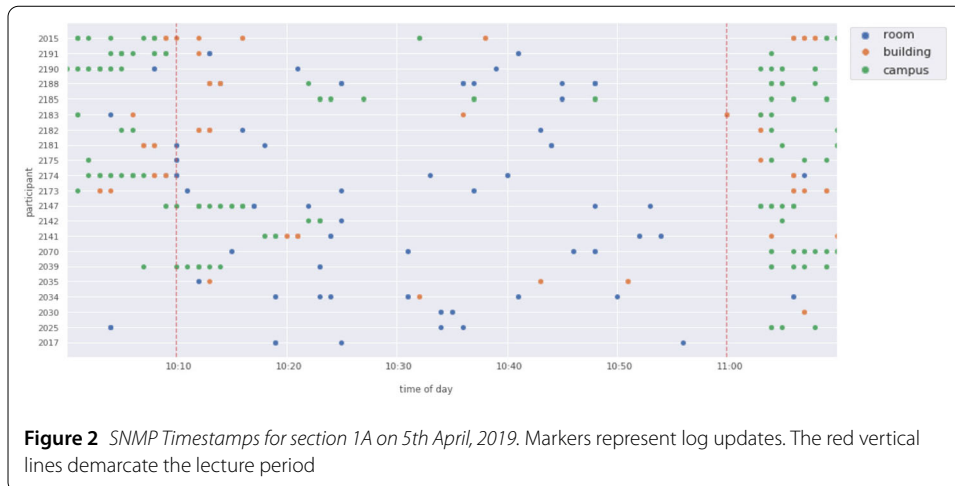
### 3.1.2 Managed WiFi network

Every AP installed on campus is mapped to a building ID and a room ID. The room ID indicates the room closest to the AP or the room that contains the AP (Table 1). Every entry in the log documents an SNMP (Simple Network Management Protocol) update in the network. This update is triggered when APs see a change, such as when a device connects. An update can also be triggered by an SNMP poll request to the AP. In response, the AP responds with devices connected to it, thus creating new logs. Therefore, the log itself indicates that a device is in the vicinity of an AP, but without information of the client RSSI, this inference has a low spatial resolution. Moreover, the logs for a connected device are erratic because of variable connectivity settings in the device agent (e.g., the WiFi turns off when inactive). The irregularity in log updates leads to a low temporal resolution. The low resolution is what introduces “coarseness” to this data. Outside of the specific association timestamps—when an AP responds to an SNMP poll or a client switches APs—the connected device is invisible in the logs.

<sup>1</sup>This analysis was approved by the Institutional Review Board (IRB) of the relevant institution, and the data was de-identified and secured in approved servers.

<sup>2</sup>No lectures took place 21st January 2019 (MLK Day), 1st week of January (winter break), and 3rd week of March (spring break).





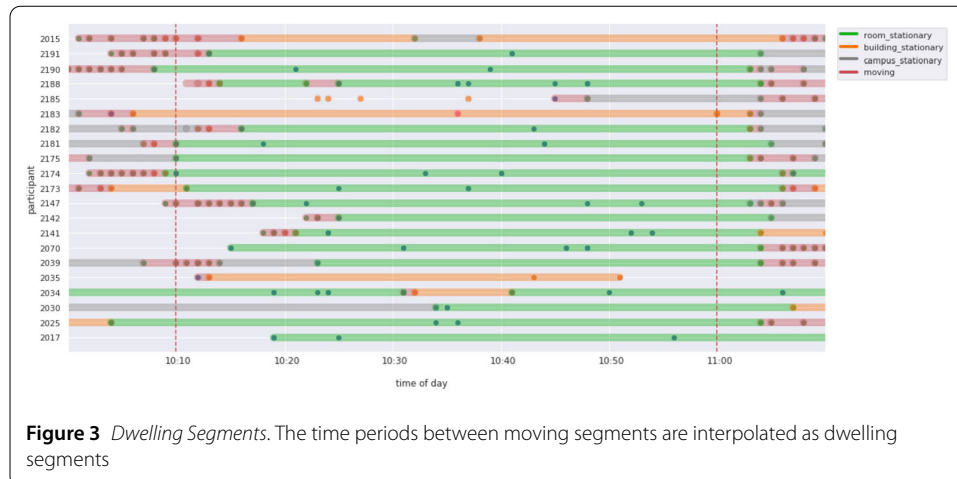
**Figure 2** *SNMP Timestamps for section 1A on 5th April, 2019. Markers represent log updates. The red vertical lines demarcate the lecture period*

### 3.2 Phase I: identifying dwelling segments from raw logs

(i) *Determine if an Individual is Mobile* — To assess how students move, we examine the logs accumulated in the 30 minutes before and after the lectures of sections 1A and 1B (Fig. 2). One of the classrooms had only 1 AP while the other one had 3 APs for coverage. Only less than 1% of the log entries showed concurrent updates at different APs from two or more devices owned by the same student. Thus we treated all log entries from a student's device as a proxy for the student. Since SNMP updates occur when a device roams, we measured the interval between two successive log entries from a user's device that associate with different APs. For example, from entering the building to entering class, devices will snap to different APs. This leads to 2 successive log entries at different locations. However, 2 such entries do not necessitate the time between them was spent moving. Consider Participant 2173 in 2, who associates with an AP outside the building, then logs an entry at an AP in the same building before logging an entry in the classroom, almost 8 minutes later. This raises the possibility that the student was dwelling in an adjacent area and then moved to class when it started. Figure 2 also illustrates that for most students the log updates before and after class times also exhibit higher update frequency in shorter intervals. We considered the 90th quantile of the intervals between 2 different logs as a reasonable threshold to reflect most of the instances where student is moving between APs. This was found to be 233 seconds. The high quantile heuristic lets us ignore anomalous or exceptional cases where a device might have lost connectivity — and failed to register any logs — while actually moving between APs. Specifically, we considered devices moving when different APs successively log the same student's device below this threshold.

(ii) *Determine if an Individual is Dwelling in Place* — The user was considered to be dwelling for any time segment when they are not mobile. Based on the criteria for moving, a user was considered stationary in 2 cases, (i) when successive log entries were at the same location, (ii) the time before the next entry exceeds the threshold. Contiguous dwelling segments where the AP does not change were combined to represent longer dwelling segments (Fig. 3).

(iii) *Filtering Out Disconnection Periods* — When students exit campus they disconnect from the network. Individuals can be lost to the network and then be “visible” when they enter a building after a period of time. Due to our threshold, the time period between these two mobility phases could be erroneously labeled as dwelling, whereas the user was

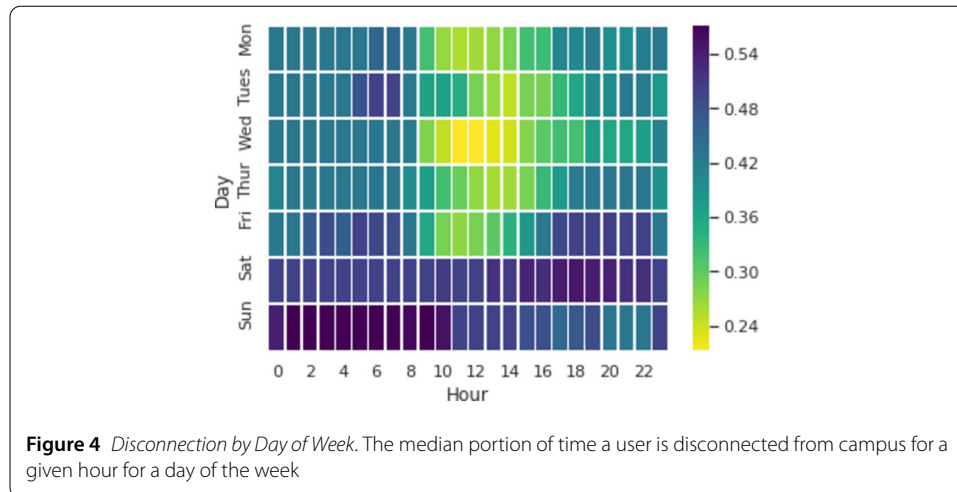


actually disconnected from the network. This large interval needed to be distinguished from actual dwelling periods. For this, we inspect the device dwelling times of students who were actually present in class (based on their attendance). Modern devices conserve their processes in various ways, such as turning off certain functions during inactivity. If a student does not use their connected device, it can appear disconnected simply because it has no active processes. Having said that, devices will “wake” from time to time to check connectivity, or when it switches to a different AP, and therefore register a log. These behaviors are unique to different connected devices and their use. Accordingly, we take a datadriven approach to identify the longest period a device was connected to an AP while inactive. We found the longest interval between two successive log entries was 76 minutes. We consider this as a heuristic threshold to filter disconnections. With this, we marked any periods of dwelling as disconnected (or inactive) where the log entries were timestamped at intervals exceeding the threshold. It is possible for a user be away from campus and back within 76 minutes. These will be the false positives of our filter. We anticipate this noise to be minimal. These occurrences will be constrained to the APs at the edge of campus, such as outdoor APs. For face validity we examined how connection and disconnection varies by the hour of the day and days of week. The disconnection periods we identified were predominantly on weekends and before or after class times (Fig. 4). This observation gives us confidence that our filter provides a reasonable estimate of location.

### 3.3 Phase II: identifying collocation

After Phase I identified individual dwelling periods, in Phase II we identified coinciding dwelling periods to describe collocation. Simply considering the overlapping dwelling segments could have breaks when even one of the collocated members inadvertently switches between AP and then returns (e.g., participant 2034 in Fig. 3). This could occur either when they took a break or if they are in place but their device intermittently found a better connection to a different AP. Since the aim of obtaining collocation segments, is to use it as a proxy for collocated social interactions (Sect. 2.1), we consider a liberal approach to characterize collocation. Moreover, just because an individual is not in sight, it does not signify the conclusion of social interactions [24]. For example, when an individual takes a brief break from a meeting, say, to grab coffee or use the restroom. Therefore, instead of dissecting the collocation around such short-lived absences, these gaps in the segments





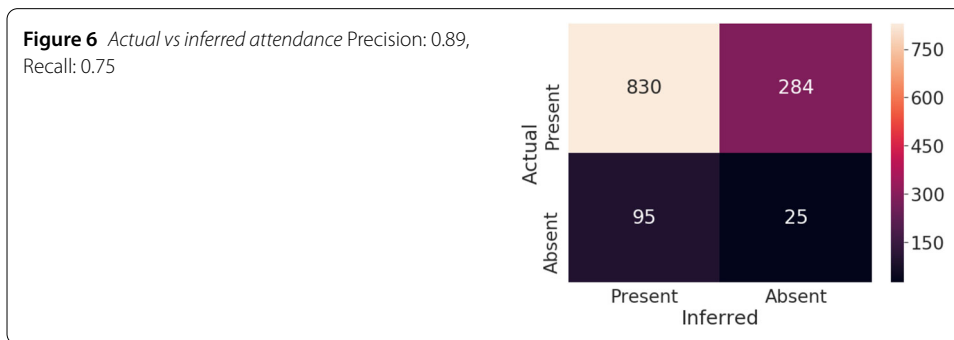
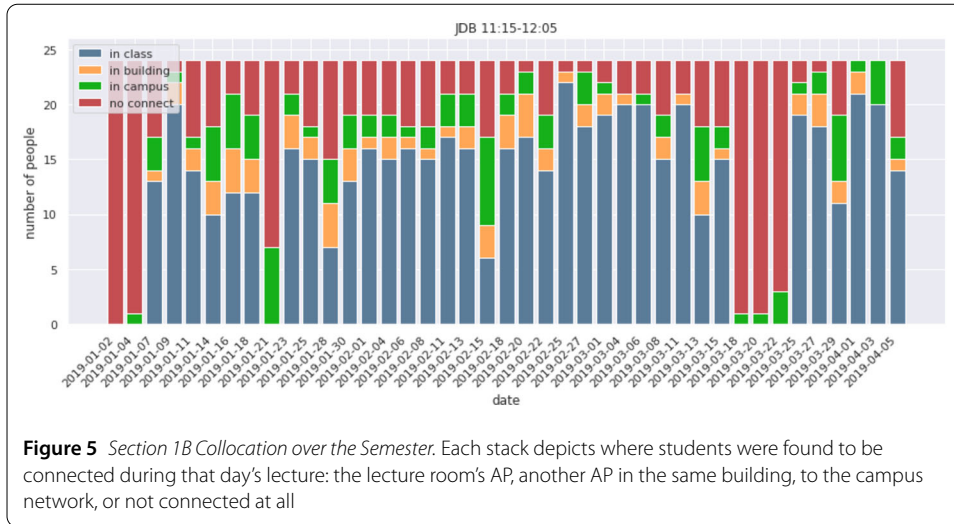
were bridged. In particular, these gaps were characterized by (i) common members of a group are collocated before and after a gap; and (ii) during the gap some subset of members are still dwelling or collocated. After identifying such overlapping segments, we first found the median duration of these gaps. In this case we consider the median as the absence of a member may or may not indicate the termination of social interaction. The median in our data for such occurrences was 11 m 7 s. Any gaps less than this threshold were resolved by considering all members to be collocated throughout, including the break period.

### 3.4 System reliability

To quantify the reliability of this coarse collocation inference, we evaluated the attendance of 46 students in 2 sections for the 34 lectures that occurred in the sample data period. Each section had 3 classes a week and but met in different buildings. For both sections, the instructor provided us with lecture-by-lecture records of each consenting student's attendance. Attending class is one form of collocation on campus that involves students gathered around a WiFi AP. Even though every AP's coverage on campus might vary, when students collocate to work outside lecture times they typically gather in breakout rooms, empty classrooms, library spaces, or other similar indoor spaces. Thus, we assume client device behaviors to be consistent between classrooms and other spaces where students would likely collocate. Hence, we consider presence in class a reasonable ground truth to evaluate the reliability of our proposed automated method for the purposes of our study.<sup>3</sup>

*Missing data* On certain lecture days, we did not find any entry for some students. The red stacks in Fig. 5 show the number of students per lecture with no log entries for section 1B. On comparing this to the attendance records, we learned that 93% of the times a student does not appear in the logs, they were actually recorded as present by the instructor. One possibility is that the student either had all their devices turned off or connected to a different network (e.g., cellular data, or the campus guest/visitor network). Every student in our sample had no WiFi log entries on at least one lecture they attended (the median was five lectures). Therefore, despite its pervasiveness, leveraging the managed network

<sup>3</sup>Note that our work centers on collocation indoors and therefore, the heuristics used in the pipeline are applicable only to similar scenarios.



can still miss out on students who were actually present. For such occurrences, the automated method cannot ascertain presence or absence and therefore, we exclude these student records (for that lecture) from further analysis.

*Performance measures* We considered a student to be in class if any time during class they were “seen” as connected to the AP associated with the room of the lecture. For every time our system identified a student to be present, the likelihood they were actually in class was 0.89. This describes the *precision* of our system. On the other hand, our system rarely indicated a student is at a location when they were not physically present. The *false discovery rate* was only 0.11 We speculate, the false positives could be due to students failing to record their name on the attendance sign-up sheet (e.g., if they showed up late to class). Alternatively, for every instance when the student was present in class, the likelihood our system inferred their presence was 0.75. This is the *recall* of our system. For reference the *false negative rate* was 0.25. A false negative could occur when a student’s device connects to a different AP on the network. Figure 3 denotes these as the orange segments. A device could also connect to an AP that is physically further away because the signal from their closest WiFi was attenuated [40].

To summarize, the *F1-score* of such a system can be interpreted as 0.81 (Fig. 6). It has high precision, but with a *sensitivity* of (0.75), it does run the risk of erroneously marking students as absent. In the future, this can be addressed by deploying a broader set of APs for a given location.

## 4 Case study: collocation and performance in groups

So far, we have shown how a managed WiFi network can be repurposed to describe collocation among individuals connected to the network. Although this system illustrates a “coarse” sensor, it also presents an easily accessible data source that can describe many social interactions over a long period of time. Therefore, it is important to examine if these retroactively inferred collocation patterns are indicative of meaningful social interactions among students on campus. We know that collocated social interactions are associated with performance [17–19, 25, 26]. If the collocation described by our system is meaningful, it should be able to reflect these latent aspects of social interaction. We evaluated the viability of archival WiFi logs for to describe these latent social interactions through this case study.

### 4.1 Study

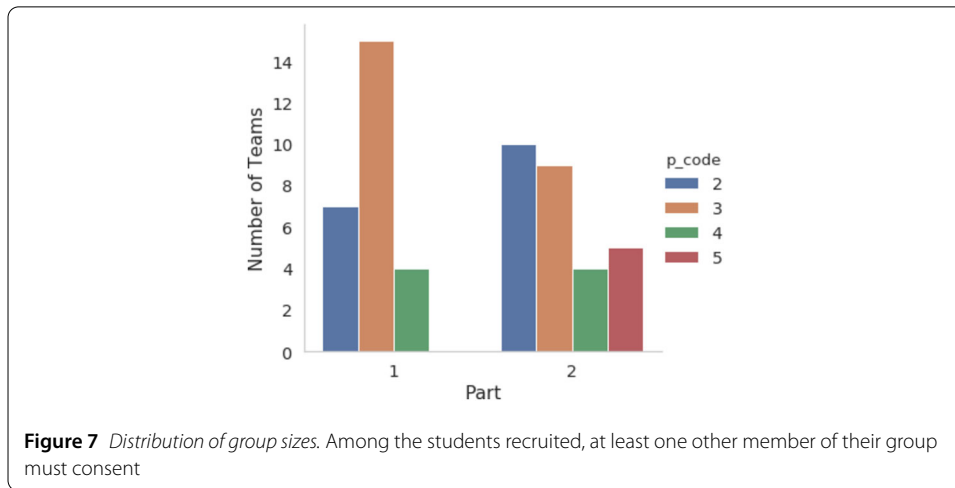
The participants were enrolled in an undergraduate design course for computer science (CS) majors. The course is offered every semester and is a two-semester sequence. Students in this course were expected to work with a team of four to six students over two semesters (Part 1 and Part 2) on a single design project. In Spring 2019, this course had four sections for Part 1 and five sections for Part 2. The system reliability described in Sect. 3.4 was based on 2 sections from Part 1 where the instructors recorded attendance for each lecture. This case study includes those students along with students from other sections and Part 2 for the analysis. Each section had an enrollment of about 40 students. In terms of course structure, Part 1 involved both lectures as well as project milestones. In contrast, Part 2 had fewer lectures and expected students to allocate scheduled class-times for project-related efforts. Students in both parts were expected to collaborate on project work outside scheduled lectures. It is not generally known how often student teams met outside of class, nor is it known how much those collocations impacted performance. Instructors of the course used various peer-evaluation surveys to characterize the psychosocial experiences of students. Although it is unclear how explanatory these surveys are, we know the their responses were subjectively factored into the final score assigned to each student. Given the biases that surveys have [29, 30], we expect objective measures of social interaction to be more indicative of performance. Therefore, in this analysis we considered a model of these peer-evaluations as a baseline in comparison to a model of automatically inferred collocation patterns.

#### 4.1.1 Participants

*Recruitment* The recruitment took place in Spring 2019 in collaboration with the course instructors. The research team advertised the study during the lectures and online outreach through the instructors. Upon enrollment, participants provided consent for the researchers to access their anonymized WiFi AP log data as well as their course data after completion of the semester. During enrollment, participants also completed an entry survey where they reported their group ID along with describing when, where, and how often they interacted with their group members face-to-face for class purposes. Participants were remunerated with a \$5 gift-card for enrolling. In total, we received consent from 186 students, which was about 51% of all students enrolled (Table 2). Of these, 170

**Table 2** Participants in the study with complete data

Section	Part 1	Part 2
A	22	21
B	24	27
C	18	31
D	20	12
E	–	11
Total	84	102



students were in the age range of 18-24 years, and 16 were of age 25 and above. Among these students, 59 reported female (32%).<sup>4</sup>

*Privacy* Participant privacy was a key concern for us. The two core streams of data, course outcomes and WiFi AP logs, are both de-identified and stored in secured databases and servers which were physically located in the researchers’ institute and had limited access privileges. The study and safeguards were approved by the Institutional Review Board of the authors’ institution.

**4.1.2 Course data**

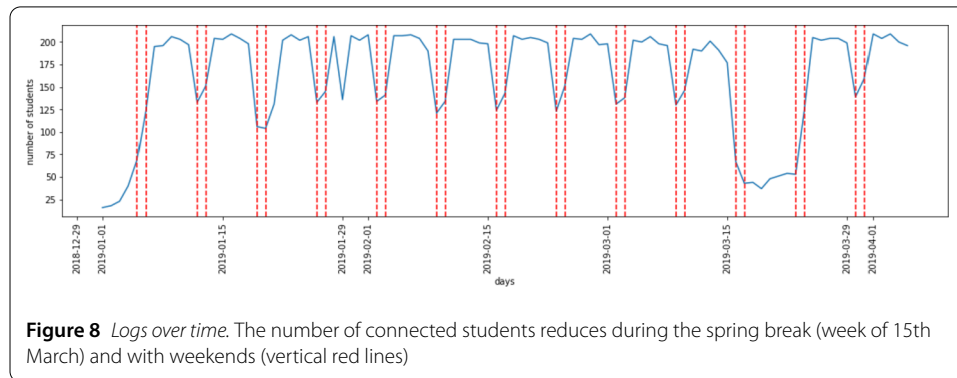
The course instructors provided course-related data for 186 consenting students along with course lecture times (Table 2). Among these students, 23 students did not have any other member from their group in our study and thus were dropped from this analysis. These remaining 163 students were in 54 separate groups (Fig. 7).

*Final score* This is a numerical score between 0 and 100 that informs the eventual letter grade based on the instructor’s grading scheme. This final score is dominated by the project outcomes but students are assessed individually. These variations are introduced by participation as well as the instructor’s subjective assessment of peer evaluation. Among the recruited group members, the range of scores between members could be as large as 6.5 points. Differences of this magnitude can be the difference between the grades of students. This final score represents the ground truth for academic performance.

<sup>4</sup>As per the official headcount 25% of the students within the CS major have been recorded as female

**Table 3** Peer-Evaluation Scales (1-5); Psychological Safety (1-7)

Construct	Mean	Med	Std
Member Effectiveness	4.36	4.45	0.51
Team Satisfaction	4.44	5.00	0.76
Conflict (Task)	1.64	1.67	0.62
Conflict (Relation)	1.26	1.00	0.51
Conflict (Process)	1.41	1.00	0.59
Psychological Safety	6.12	6.29	0.80



**Peer evaluation** Students completed an extensive peer-evaluation battery at the end of the semester (Table 3). This battery comprises of validated survey instruments to quantify aspects of social interactions that are expected to be related to individual performance:

- *Team Conflict* [41] — Conflict represents the perception of incompatible goals or beliefs between individuals that cannot be trivially reconciled. Less conflict leads to more motivation and satisfaction and therefore associated with performance enhancement [42, 43].
- *Team Satisfaction* [44] — Satisfaction reflects the contentment in terms of expectations. Dissatisfaction with one’s team can lead to lower levels of task performance [43, 45].
- *Psychological Safety* [46] — This captures a “shared belief held by members of a team that the team is safe for interpersonal risk taking” [46]. It is associated with individual learning progress as they are more amicable to experiments and feedback [46].
- *Team Member Effectiveness* [47] — This measure encompasses five dimensions<sup>5</sup> related to “team member effectiveness”

We used a participant’s responses to these surveys to build a gold-standard baseline model to infer their final score.

#### 4.1.3 Network data

The WiFi access point log data for consenting students was obtained from the institute’s IT management facility. This data is richer compared to the sample data for the processing pipeline (Sect. 3.1.2) — it includes more individuals and a larger set of APs. The data spans a time frame of 95 days between January 1 2019 and April 5 2019 (Fig. 8). On average, the time between the first log entry for any one of a participant’s devices and the last is approximately 90 days. The logs in this study included 204 unique buildings with 4865

<sup>5</sup>While the other scales were self-evaluations, this score is the average of how their peers evaluated a team member

unique APs. We only found multiple APs to be in the same room for 803 rooms. Additionally, the 204 buildings were manually categorized to best express the purpose of that space [5, 13] — for example, “academic”, “dining”, “green spaces”, “recreation”, and “residential”. Two researchers referred to campus resources to independently assigned categories to these buildings. Only two of the building labels disagreed, which was resolved by a third researcher. The raw logs of the consenting students was processed as described in Sect. 3.3 to obtain periods when students were dwelling and collocated. Over the semester, the median collocation duration was about 70 hrs.

## 4.2 Feature engineering

The low spatial resolution of the collocation makes it insufficient to assert from isolated instances if collocation of group members were connected to their performance. However, processing multiple collocation periods over the semester can approximate collocated interactions. For instance, members of the same group might collocate regularly at a specific type of building. Therefore, we engineered features that captured such patterns.

### 4.2.1 Feature descriptions

We extracted relevant information at a week-level based on various semantically labelled behaviors (Table 4). For feature crafting we used the schedule of lectures provided by the instructors to distinguish behaviors during class from those outside class. Moreover, every AP in our dataset was manually annotated to describe the intended purpose of the building. Therefore, we could outline if a student was dwelling or collocated in either an academic space, a residential one, or a recreational space. We first extracted a set of “Individual features” that characterized behaviors which are not explicitly social, but could impact performance (e.g, attendance.). Then we separately characterized “Group features” or “Collocation features” captured the behaviors of individuals related to their group, such as time spent collocated with other group members. This dissociation of features helps provide discriminant validity. Essentially, it helps us assert that coarse collocation-based features are not confounded by an individual’s general behavior, such as the time spent in academic spaces. Fundamentally, all features measure the duration of various dwelling and collocation activities. Since this study was primarily intended to showcase feasibility we only focus on these basic measures. However, the dataset can be used to craft more nuanced features (such as punctuality or timeseries characteristics). To craft the collocation features, we used computed both absolute duration and a relative percentage (of collocation time spent by all members of the group).

**Table 4** Raw features derived from the collocation data at a weekly level

Type	Description	Spatial Variants		
		Any Academic	Residential	Recreational
<i>Individual Features</i>				
Attendance	Present at lecture room during scheduled time	○	○	○
Dwell	Time spent at a place while stationary			
<i>Collocation Features—Measured as absolute duration and relative to the group</i>				
Scheduled	Time spent with group members during reported weekly meeting times			
Class	Time spent with group members during class hours	○	○	○
Other	Time spent with group members at other times			



For collocation, we delineated three types of behaviors based on when the student was collocated with others from their group:

1. *Scheduled*: Groups reported their regular meetings in a free-form response field during enrollment (Sect. 4.1.1). Responses typically indicated a primary building (e.g., learning commons) along with a potential backup (e.g., library). However, teams also expressed meetings could take place at undetermined locations on campus. Moreover, groups often provided multiple tentative meeting times and places for a week. To accommodate all possibilities, this feature captured the collocations between group members that occurred during any of the reported periods.
2. *Class*: This captured collocations with group members during class times. This is different from the attendance feature because it considered collocation outside the assigned lecture room. For instance, students in Part 2 were expected to meet during class time, and not necessarily in the scheduled room for the class. Based on student reports, Part 2 teams did not necessarily use all class times in a week for meetings.
1. [3] *Other*: This is a catch-all bucket to capture all other ad-hoc collocations. Only 4 groups in our study reported interacting with group members for non-academic reasons (e.g., “lived together”). Students could be found to be collocated outside of class and scheduled meetings due to various factors. On one hand, this could indicate extra effort as meetings could occur outside schedule. On the other hand, students could collocate to complete course work together, or be serendipitously in the same space. In that case collocation can serve to improve social bonds and in turn improve performance [17, 28]. Thus we include this bucket of features to represent social interactions beyond structured and anticipated meetings.

#### 4.2.2 Feature processing

The features described above were aggregated by week over the 14 weeks in the semester. Table 4 shows green ticks to indicate the features we compute. Considering a unique feature for each week, we computed  $5 \times 14$  for *individual features* and  $(9 \times 2) \times 14$  for *group features*. However, this creates a large feature space (196 dimensions) given our sample of 163 students. To reduce the feature space we calculate summary features to describe the entire semester of the individual. Specifically for each feature extracted at a week level, we computed the *median*, the *mean* and the *standard deviation* for the study period. In addition to these, we also computed the *approximate entropy* of the feature per individual [48]. This aggregation reduced the overall feature count to 20 and 36 for individual and group features, respectively.

### 4.3 Training and estimation

We built multiple regression models to investigate how the collocation-based features estimate final scores in comparison to survey-based peer evaluation scores.

#### 4.3.1 Model descriptions

$M_{\text{Peer}}$  denotes the model trained on peer-evaluation scores (Sect. 4.1.2) based on the self-reported survey responses provided by the instructors.  $M_{\text{Indi}}$  refers to the model trained on individual features and  $M_{\text{Colloc}}$  describes the model trained only features that represent

collocation among group members — potentially describing collocated social interactions. We assessed the discriminant validity in predicting final course scores with each subset of features without confounding effects from other features. Furthermore, we develop combination models to comprehensively understand how a combination of automatically generated features estimate academic performance ( $M_{\text{Indi+Colloc}}$ ).

#### 4.3.2 Estimators and validation

We evaluated all models through a 5-fold cross-validation process. Since the scores of members in the same groups are likely to be closer to each other, we ensured that members of the same project group remain in the same fold. This ascertained that no data leaked inadvertently between training and testing. Specifically, we did not train on one student and test on their group member who was likely to have a similar score. In this way, our models were independent of which groups we trained on. To estimate the target variable (the final score), for each model described, we trained a Linear Regressor [49] to represent linear relationships between features and a Decision Tree Regressor [50] for non-linear relationships. Additionally, we also train a Gradient Boost Regressor [51], i.e., an ensemble learner. To determine the relationship between model features and final scores, we measured the correlation between the predicted value and the actual values. The correlation results we report are pooled over all the folds. Pooling ensures that our results are robust to the heterogeneity between splits in the cross-validation process. Therefore, the pooled correlation between values can provide a more generic relationship over the entire sample of observations [52]. For internal validation, we compared these models to a rudimentary baseline  $M_0$ , which always estimated the median of the target variable from the training set.

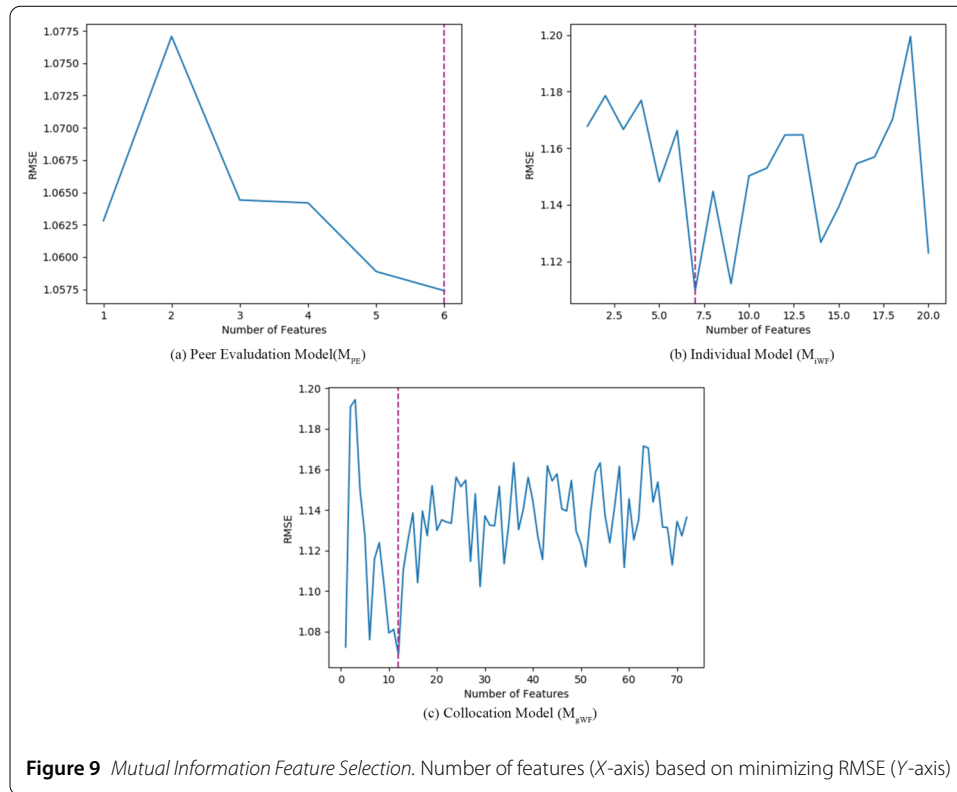
#### 4.3.3 Feature transformations and selection

We performed the following transformations (fitted only on the training folds):

1. *Scaling Final Scores by Instructor* — Since the final score varies based on the instructor, we standardized the final scores based on the distribution of scores for each instructor in the training data.
2. *Impute Missing Data* — Some students had not have completed all survey instruments, or a few project teams did not report their scheduled meeting times (7 students). We imputed these missing values with the mean of the feature.
3. *Standardize the Features* — Converted to zero mean and unit variance [53].
4. *Mutual Information Regression* — We used the mutual information between the training features and the target variable for univariate feature selection [54]. The number of features selected were varied from 1 to  $k$ , where  $k$  was the total number of features in the model (Fig. 9). We selected the  $k$  that minimized the RMSE (Root Mean Square Error) [55].

## 4.4 Results: model comparison

Table 5 summarizes the results with the best estimator for each model. For any set of features, only the estimator that minimized the RMSE was considered for comparison between models. To compare models we used *Pearson's r* to describe the correlation of each model's estimate with the final scores of the students. This coefficient characterizes the complete association by considering all observations and does not assume normality



**Figure 9** Mutual Information Feature Selection. Number of features (X-axis) based on minimizing RMSE (Y-axis)

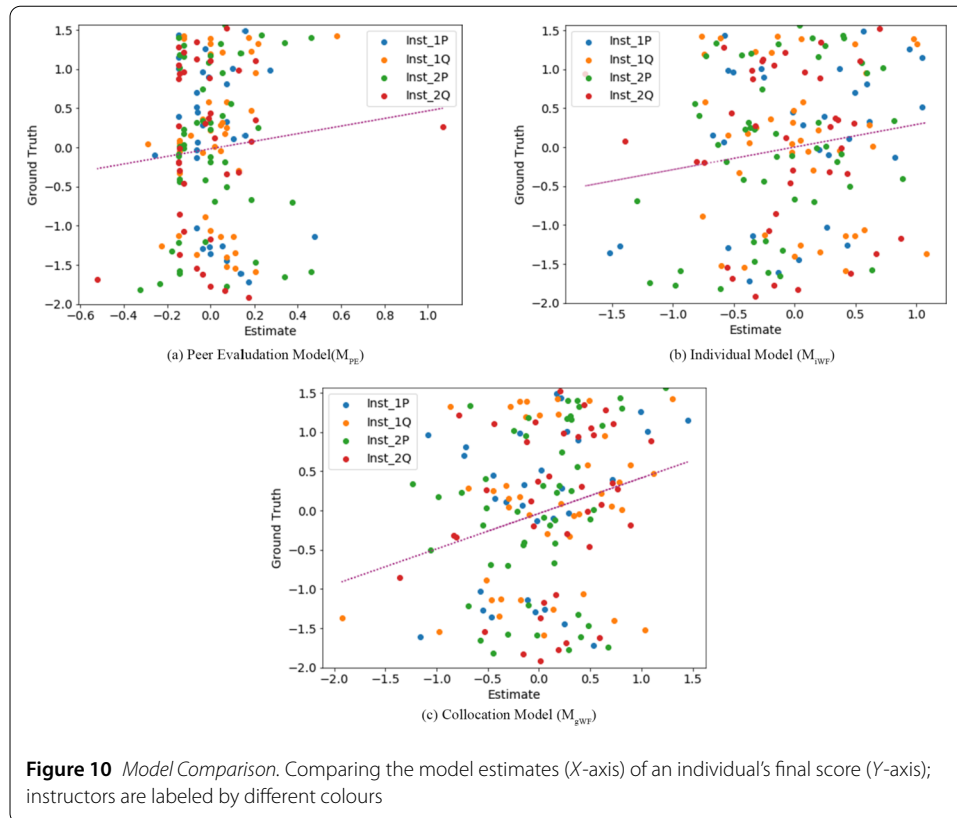
**Table 5** Model Performance. (‘.’:  $p < 1$ , ‘.’:  $p < 0.1$ , ‘\*’:  $p < 0.05$ , ‘\*\*\*’:  $p < 0.01$ )

Model	Training Data	Estimator	Pearson’s R	
$M_{Peer}$	Peer Evaluation	LR	0.08	–
$M_{Indi}$	Individual Behavior	GB	0.14	.
$M_{Colloc}$	Collocation Behavior	GB	0.24	**
$M_{Indi+Colloc}$	Individual + Collocation	GB	0.25	**

[56]. All models exhibited an improvement over  $M_0$  — the rudimentary median estimator. None of the models based on peer evaluation features ( $M_{Peer}$ ) were found to be significant, but among them Linear Regression showed the most error reduction. For  $M_{Indi}$  the best estimator used Gradient Boost. Its estimates were more significant but with a weak correlation of 0.14. In comparison, for  $M_{Colloc}$  the best estimator, which used Gradient Boost, exhibited a very significant correlation of 0.24. We also compared the dependent overlapping correlations [57] of  $M_{Colloc}$  against  $M_{Peer}$  and  $M_{Indi}$  (with a confidence-interval of 90%). In both cases, the correlation of  $M_{Colloc}$  with the final score was significantly different than that of  $M_{Peer}$  ( $p = 0.02$ ) and  $M_{Indi}$  ( $p = 0.08$ ) (Fig. 10). Additionally, incorporating both individual and within–group behaviors showed minor improvement. This improvement was not significant in comparison to  $M_{Colloc}$  [57].

### 4.5 Interpretation of results

We know from literature that collocation can provides opportunity for a variety of social interaction that are linked to an individual’s performance [14, 16, 18, 19, 46]. Our model trained on students’ collocation behaviors ( $M_{Colloc}$ ) could estimate academic performance with a significant correlation of 0.24. In the context of social sciences this would be consid-



ered a moderate effect [58, 59]. Moreover, other work specific to the domain of academic performance report similar magnitude of correlation [60] and consequently validate the importance of our results. This implies that a managed WiFi network can be retroactively leveraged to describe meaningful collocated social interactions.

We compared our model of collocation behaviors with other models for discriminant validity. The results show that the model trained on students' collocation behaviors ( $M_{Colloc}$ ) outperformed the correlation of estimates obtained by modeling peer-evaluation ( $M_{Peer}$ ) and individual behaviors ( $M_{Indi}$ ). The features in  $M_{Colloc}$  aggregate collocation behaviors of students known to be socially connected over multiple weeks. This emphasizes that collocation captures aspects of performance that cannot be captured by self-reported surveys or individual variances. We believe this difference in models is because  $M_{Colloc}$  reflects social interactions. While peer evaluation scores are expected to yield better correlations [15, 43, 45, 61], the social desirability bias in manually reporting team experiences can wash out the intricacies of actual team behavior [29, 30].  $M_{Indi}$  was also found to be somewhat better than the peer-evaluation model. This itself implies that dynamic objective measures of individual behaviors can explain performance in groups better than surveys. In the larger context of academic progress, individual behaviors could have a much larger role in explaining performance. A study of a variety of students across different majors over two years shows that attendance measures can explain final grades with a correlation of 0.24 [60]. By contrast, our study is focused on a specific group-based, project-intensive course for CS majors. Attendance to lectures was only required for students in Part 1 of the course (about 45% of our participants). Our participants were expected to meet in person to work on their project towards their final score. For the most part, students were

expected to self-determine their meetings and work towards the course outside traditional meeting times. Given the collaborative nature of the course, it is not surprising that  $M_{\text{Indi}}$  falls short of  $M_{\text{Colloc}}$ . Note, our passive inference of collocation does not explicitly discern what transpired when a student and their group members were in the same space. Although it is not necessary that the collocated students verbally interacted, in line with the concept of *spatiality*, even the presence of peers in the vicinity can affect individual performance through non-verbal cues [17].

The coarse nature of this sensor makes it challenging to unpack the exact nature of collocated social interactions. However, given the ubiquity of such managed WiFi networks, we can now conceive this sensor as a complement to other methods of understanding social interactions. By retroactively describing collocation patterns, researchers have the opportunity to devise new hypotheses given their specific situated communities.

## 5 Discussion

We presented empirical results that even coarsely inferred collocation of related individuals is linked to their academic outcomes. This validates that we can characterize aspects of collocated social interactions by retroactively studying group behaviors in student cohorts on campus. Accessing network logs is not uncommon at universities and has no additional overhead. In fact, this alternative provides an additional benefit to universities without excessive spending.

### 5.1 Applications of inferring collocations for academic experiences

Our system demonstrates the potential of a new analytical lens to understand social behaviors on campus. This enables instructors to provide data-driven insights to a new cohort based on actual behaviors of successful teams. However, collocation is only beneficial for certain kinds of projects [14, 17, 19], such as software development, or, as in our case, design. To understand the transferability of our results to other forms of academic work, researchers need to further inspect what occurs between the group members during collocation. Identifying these activities can help define which characteristics of collocated synchronous interactions [17] are actually associated with higher performance. For example, project members might just be more dedicated to their tasks in the presence of others [19], or collocation might improve their social bond and make them more comfortable about feedback [16]. Interviews along with momentary assessments can guide researchers to automatically infer the social importance of different collocations based on the location, time, and history of collocated individuals. This knowledge could be used to augment the static semantic labels of places and instead illustrate a more dynamic social blueprint of campus. Moreover, since these logs can be retroactively obtained, it can provide data to explore new questions that help determine student outcomes. For instance, how do members of teams with prior collocations work in comparison to teams of strangers [62], or how different are collocation patterns in a new cohort for a student from a marginalized community [63]. Practically, these results also have implications for remote learning as more universities have embraced distributed classrooms. This helps universities consider the trade-offs for using spaces for collocated group activities while also promoting the need for remote collaboration technologies that can approximate collocation behavior — similar to what has been advocated by research on dispersed information work [15, 17, 18, 26, 27]. Theoretically, our work begs to question the relationship between collocation of students

and social relationships outside curricular activities. While collocation of team members can build stronger social ties [28], it is yet to be determined if the same can be said for students not associated through projects or academic outcomes. Subsequent research can consider combining other kinds of archival data to gain a richer understanding of on-campus social behaviors. Data logged in learning management systems can be incorporated to examine more short term effects of collocation. Similarly, processing data from access cards can help distinguish the purpose of certain campus collocation events (e.g., purchasing lunch together). With appropriate procedures for consent, many such archival sources of data on campus can be combined with collocation data to provide novel insights on social behaviors.

## 5.2 Privacy, policy and ethics

Any pervasive technology with the potential of large-scale passive sensing faces privacy concerns [64]. The use of the WiFi association logs is confined to campus and does not elicit anxieties related to a client-side applications leaking data from other sensors. However, automatic computation of where individuals are and whom they interact with can be considered sensitive by students [65]. Therefore, when adopting such approaches to infer interactions, stakeholders need to consider approaches like *differential privacy* to obfuscate sensitive data [66].

Even when anonymous, predators can mine collocation patterns to identify individuals [67]. To protect against this, more data can be abstracted, i.e., the AP locations can be anonymized as well (while retaining category, floor, and relative information). Yet, it still needs to be established who has the privileges to query for information and what the queries can be [68]. Moreover, campuses can adapt existing access policies for student records to protect student collocation patterns.

Since accumulation of network association logs is not uncommon at universities, it does not present any new surveillance infrastructure and instead posits reusing existing methods. As per the *principle of proportionality*, collocating on campus can be considered public information [69]. However, it is not the localization that is sensitive, but the accumulation and aggregation of such data that makes privacy negotiations challenging [70]. These differing expectations of how this data is used can be considered concerning by the campus community. Therefore, for practical deployment, it is imperative for any community that seeks to use this data to secure some form of consent. Opt-out procedures need to provide explicit notice e.g., notify a change in terms of service.

However, choosing to opt-out can be considered an unfair choice that limits students right to self-determine [71] — campus' managed network provides access to key resources. If instructors use this data for intervening with certain groups during midterms, those not on the network would not have the same opportunity of improvement. Campuses must establish safeguard policies that ensure no individual is penalized for their choices [72] Further, while repurposing these logs is a form of data minimization, we propose that stakeholders define paradigms for “use minimization” — when, what, and how much of such data can be processed for applications. In many cases, this data should only be accessed retroactively, only for public areas (which excludes housing), and span limited periods.



### 5.3 Limitations and future work

The most apparent limitation of using these association logs to determine collocation is its low spatio-temporal resolution. This introduces reasonable uncertainty in determining the exact location of individuals [12, 47]. Moreover, since every connecting client device is unique in its connectivity management, it is non-trivial to establish exactly how every device would be located. For our system we introduce various data-driven heuristics to tackle this coarseness (Sect. 3). These heuristics were informed by a small set of students (and their devices) as well as limited access points. Apart from idiosyncratic attendance patterns, even the specific device make, and the physical obstacles around the AP can impact how logs are updated. Since our heuristics are trying to model specific sets of students at specific locations, their unique patterns could have downstream effects on reliability and pattern modeling. Further research should consider reevaluating the system reliability over a more diverse set of observations in terms of users, devices, and locations.

We know from our system reliability tests (Sect. 3.4) that our system is more likely to label false-negatives. Therefore, our system might be underestimating the actual collocation incidents. According to the literature, collocation supports social interaction and in turn is linked to performance. We studied this phenomenon on students in a group project to test the feasibility of leveraging WiFi network logs to indicate collocated social interactions. Taking into account missed instances of collocation, a more true measure might indicate a stronger relationship to performance than what we found. On the contrary, our system also can generate false positives. One of the reasons for this could be the challenge in distinguishing if a user is dwelling or disconnected. Within certain thresholds, a user could go outside the network and return in between two phases of movement. For instance, they went to get coffee outside the campus and returned. These episodes can appear like collocation when non occurred. Overall, these incidents are unlikely and only expected to happen in APs at the edge of campus. Yet, this could contribute to the noise in our data which could get modelled erroneously. One way to mitigate this would be to maintain a variable set of heuristics given different APs and their locations on campus. Extensively studying crowd behavior with some self-reported ground-truth can help reduce this noise.

Some of the major challenges to leveraging WiFi is related to the infrastructure it is a part of. Indoor setups present several challenges that can lead to unexpected device associations [40]. As a result, an individual could be in a room and not be associated with the physically closest AP, but rather another AP node that found a stronger signal to the client. This creates an opportunity to deal with this noise by modeling the probability of displaced connections. and incorporating the size configuration of rooms and neighborhood maps of the APs. Furthermore, advanced off-the-shelf methods to study archival data can be developed to make AP nodes aware of other APs visible to a client — similar to RTLS approaches [33, 34].

For identifying collocation of group members, we assume that students devices are connected to the network during collocation. This is a reasonable assumption since our participants comprised Computer Science majors enrolled in a design course. In contrast, this might not be true for other forms of group work. For instance, projects at hardware workshops could have extended periods where digital devices are untouched and appear disconnected. Therefore, it is important to be aware of the expected device use during collaboration in physical spaces to estimate if WiFi logs can approximate user location.

Our study is aimed to demonstrate the feasibility of a managed WiFi network to retroactively describe collocation behaviors that can approximate the effects of social interactions. We showed this by focusing on a finite set of features to describe collocation behaviors over the semester. The scope of our study restricted us to a course that involved interactions once a week. This informed us to aggregate our features by week (Sect. 4.2). While this decision was apt for our study, it also reduced the number of longitudinal observations we had (14 data points for each week). Having said that, with the appropriate consent, the managed WiFi network can help explain collocation across more frequently (examine classes that occur every day) and over longer periods (full academic year). Depending on the specific hypotheses, future work can craft more nuanced collocation features (e.g., seasonal change in the number of unique individuals collocating with the student).

Theoretically, our findings coincide with ideas of *spatiality* [17]. According to this, when collaborators are present near each other, they are interacting through observations and an increased sense of accountability. A key assumption to our study was that we examined behaviors of participants who were expected to have collocated social interactions. However, it is possible for individuals to be collocated and yet unaware of each other's presence. These episodes could have other unseen relationships with their performance. In our modelling we distinguish collocation instances by location (academic, residential, or recreational) and schedule (class, meetings, or other). One might assume that collocation in "recreational" areas outside classes or meetings might be coincidental and unlikely to foster any real social interaction. In the same vein, subsequent work can introduce additional dimensions to semantically categorize different collocation instances (e.g., collocation at a specific time or place that only occurred once could be ignored). As researchers consider further studies in this space they can refine what can be learned from WiFi-inferred collocation. For future studies in this direction, we encourage researchers to include other sources of data that can mitigate the blind spots of WiFi logs and also qualify them better.

## 6 Conclusion

Collocated social interactions are valuable to the experience of students, workers, and other communities that share a common physical space. The more we can understand these social behaviors, the better we can support the community. This paper examined the feasibility of expressing collocation from WiFi network logs. We established the reliability of computing collocation of students on campus. Then we demonstrated that this characterization of collocation behaviors can reflect individual outcomes of collocated social interactions, particularly, success in a group project. Our work encourages future opportunities to apply such a data source to support the campus community.

### Acknowledgements

We thank Matt Sanders and the Office of Information Technology at Georgia Tech for collaborating with to access WiFi log data under secure protocols. Similarly, we thank the course instructors who helped us in recruitment and obtaining consented course data. Also, we thank Aaron Striegel and his lab at University of Notre Dame for reviewing our early draft. Lastly, we thank members of the Social Dynamics and Wellbeing Lab and Ubicomp group at Georgia Tech for their feedback throughout this research.

### Funding

Some research personnel were partially supported by funding for endowed chairs from the Georgia Institute of Technology.

### Abbreviations

AP, Access Point; RTLS, Real-Time Location Systems; RSSI, Received Signal Strength Indicator; SNMP, Simple Network Management Protocol;  $M_0$ , Rudimentary baseline;  $M_{\text{peer}}$ , Model trained on peer-evaluation scores;  $M_{\text{indi}}$ , Model trained on individual features;  $M_{\text{colloc}}$ , Model trained on collocation among group members.

### Availability of data and materials

The data used in this paper are not available for public use. These data were accessed under specific data–use agreements that restrict data sharing. For further information about data, access and collaboration, contact VDS.

### Declarations

#### Competing interests

The authors declare that they have no competing interests.

#### Author contributions

VDS, HK, BS, MBM, TP, MDC, and GDA designed the research; VDS, HK, BS, MBM, and GDA acquired the data; VDS, SS, KT, DP, YT, JP, and YC analyzed, and interpreted data for the work; VDS, TP, MDC, and GDA wrote the paper. All authors read and approved the final manuscript.

#### Author details

<sup>1</sup>Georgia Institute of Technology, North Ave NW, Atlanta, USA. <sup>2</sup>Northeastern University, 360 Huntington Ave, Boston, USA.

Received: 14 May 2021 Accepted: 9 June 2023 Published online: 07 July 2023

### References

1. Homans GC (1974) Social behavior: its elementary forms
2. Lukowicz P, Pentland S, Ferscha A (2011) From context awareness to socially aware computing. *IEEE Pervasive Comput* 11(1):32–41
3. Olgú in DO, Waber BN, Kim T, Mohan A, Ara K, Pentland A (2008) Sensible organizations: technology and methodology for automatically measuring organizational behavior. *IEEE Trans Syst Man Cybern, Part B, Cybern* 39(1):43–55
4. Eagle N, Pentland AS (2006) Reality mining: sensing complex social systems. *Pers Ubiquitous Comput* 10(4):255–268
5. Wang R, Harari G, Hao P, Zhou X, Campbell AT (2015) Smartgpa: how smartphones can assess and predict academic performance of college students. In: *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pp 295–306
6. Das S, Chatterjee S, Chakraborty S, Mitra B (2018) Groupsense: a lightweight framework for group identification. *IEEE Trans Mob Comput* 18(12):2856–2870
7. Nguyen T, Phung D, Gupta S, Venkatesh S (2013) Extraction of latent patterns and contexts from social honest signals using hierarchical Dirichlet processes. In: *2013 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE, Los Alamitos, pp 47–55
8. Shilton K (2009) Four billion little brothers? Privacy, mobile phones, and ubiquitous data collection. *Commun ACM* 52(11):48–53
9. Das Swain V, Reddy MD, Nies KA, Tay L, De Choudhury M, Abowd GD (2019) Birds of a feather flock together: a study of person–organization fit through latent activity routines. In: *Proc. ACM hum.-comput. Interact (CSCW)*
10. Hong H, Luo C, Chan MC (2016) Socialprobe: understanding social interaction through passive wifi monitoring. In: *Proceedings of the 13th international conference on mobile and ubiquitous systems: computing, networking and services*, pp 94–103
11. Shi J, Meng L, Striegel A, Qiao C, Koutsoukolas D, Challen G (2016) A walk on the client side: monitoring enterprise wifi networks using smartphone channel scans. In: *IEEE INFOCOM 2016—the 35th annual IEEE international conference on computer communications*. IEEE, Los Alamitos, pp 1–9
12. Ware S, Yue C, Morillo R, Lu J, Shang C, Kamath J, Bamis A, Bi J, Russell A, Wang B (2018) Large-scale automatic depression screening using meta-data from wifi infrastructure. *Proc ACM Interact Mob Wearable Ubiquitous Technol* 2(4):1–27
13. Eldaw MHS, Levene M, Roussos G (2018) Presence analytics: making sense of human social presence within a learning environment. In: *2018 IEEE/ACM 5th international conference on big data computing applications and technologies (BDCAT)*. IEEE, Los Alamitos, pp 174–183
14. Kozlowski SW, Hulst BM (1987) An exploration of climates for technical updating and performance. *Pers Psychol* 40(3):539–563
15. Edmondson AC, Bohmer RM, Pisano GP (2001) Disrupted routines: team learning and new technology implementation in hospitals. *Adm Sci Q* 46(4):685–716
16. Fruchter R, Bosch-Sijtsema P, Ruohomäki V (2010) Tension between perceived collocation and actual geographic distribution in project teams. *AI Soc* 25(2):183–192
17. Olson GM, Olson JS (2000) Distance matters. *Hum-Comput Interact* 15(2–3):139–178
18. Geister S, Konradt U, Hertel G (2006) Effects of process feedback on motivation, satisfaction, and performance in virtual teams. *Small Group Res* 37(5):459–489
19. Mark G, Gonzalez VM, Harris J (2005) No task left behind? Examining the nature of fragmented work. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, pp 321–330
20. Finch JF, Barrera M Jr, Okun MA, Bryant WH, Pool GJ, Snow-Turek AL (1997) The factor structure of received social support: dimensionality and the prediction of depression and life satisfaction. *J Soc Clin Psychol* 16(3):323–342

21. Ford GG, Procidano ME (1990) The relationship of self-actualization to social support, life stress, and adjustment. *Soc Behav Pers Int J* 18(1):41–51
22. Cannon-Bowers J, Tannenbaum S (1995) Defining team competencies and establishing team training requirements. In: Guzzo R, Salas E (eds) *Team effectiveness and decision making in organizations*. Jossey-Bass, San Francisco, pp 333–380
23. Sparrowe RT, Liden RC, Wayne SJ, Kraimer ML (2001) Social networks and the performance of individuals and groups. *Acad Manag J* 44(2):316–325
24. Rummel RJ (1976) *Understanding conflict and war: vol. 2: the conflict helix*. Sage, Beverly Hills
25. Kozlowski SW, Ilgen DR (2006) Enhancing the effectiveness of work groups and teams. *Psychol Sci Public Interest* 7(3):77–124
26. Hinds PJ, Bailey DE (2003) Out of sight, out of sync: understanding conflict in distributed teams. *Organ Sci* 14(6):615–632
27. Cramton CD (2001) The mutual knowledge problem and its consequences for dispersed collaboration. *Organ Sci* 12(3):346–371
28. Trainer EH, Kalyanasundaram A, Chaihirunkarn C, Herbsleb JD (2016) How to hackathon: socio-technical tradeoffs in brief, intensive collocation. In: *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pp 1118–1130
29. Aiken LS, West SG (1990) Invalidity of true experiments: self-report pretest biases. *Eval Rev* 14(4):374–390
30. Krumpal I (2013) Determinants of social desirability bias in sensitive surveys: a literature review. *Qual Quant* 47(4):2025–2047
31. Schröder T, Hoey J, Rogers KB (2016) Modeling dynamic identities and uncertainty in social interactions: Bayesian affect control theory. *Am Sociol Rev* 81(4):828–855
32. Yang C, Shao H-R (2015) Wifi-based indoor positioning. *IEEE Commun Mag* 53(3):150–157
33. Cisco C (2014) <https://www.cisco.com/en/US/docs/solutions/Enterprise/Mobility/emob30dg/Locatn.html>. Accessed: 2023-07-05
34. Accuware AW (2016) <https://www.accuware.com/support/wi-fi-location-monitor-accuracy/>. Accessed: 2020-05-10
35. Martani C, Lee D, Robinson P, Britter R, Ratti C (2012) Enernet: studying the dynamic relationship between building occupancy and energy consumption. *Energy Build* 47:584–591
36. Jiang S, Zhu X, Huang J, Shou G (2015) Mining social groups in campus based on wireless detection. In: *2015 IEEE international conference on smart city/SocialCom/SustainCom (SmartCity)*. IEEE, Los Alamitos, pp 285–288
37. Wang Y, Shao L (2018) Understanding occupancy and user behaviour through wi-fi-based indoor positioning. *Build Res Inf* 46(7):725–737
38. Zakaria C, Balan R, Lee Y (2019) Stressmon: scalable detection of perceived stress and depression using passive sensing of changes in work routines and group interactions. In: *Proceedings of the ACM on human-computer interaction* 3(CSCW), pp 1–29
39. De Montjoye Y-A, Stopczynski A, Shmueli E, Pentland A, Lehmann S (2014) The strength of the strongest ties in collaborative problem solving. *Sci Rep* 4(1):1–6
40. Kjaergaard MB, Nurmi P (2012) Challenges for social sensing using wifi signals. In: *Proceedings of the 1st ACM workshop on mobile systems for computational social science*, pp 17–21
41. Jehn KA, Mannix EA (2001) The dynamic nature of conflict: a longitudinal study of intragroup conflict and group performance. *Acad Manag J* 44(2):238–251
42. Carnevale PJ, Probst TM (1998) Social values and social conflict in creative problem solving and categorization. *J Pers Soc Psychol* 74(5):1300
43. Taylor SE, Brown JD (1988) Illusion and well-being: a social psychological perspective on mental health. *Psychol Bull* 103(2):193
44. Van der Vegt GS, Emans BJ, Van De Vliert E (2001) Patterns of interdependence in work teams: a two-level investigation of the relations with job and team satisfaction. *Pers Psychol* 54(1):51–69
45. Jiang JY, Zhang X, Tjosvold D (2013) Emotion regulation as a boundary condition of the relationship between team conflict and performance: a multi-level examination. *J Organ Behav* 34(5):714–734
46. Edmondson A (1999) Psychological safety and learning behavior in work teams. *Adm Sci Q* 44(2):350–383
47. Loughry ML, Ohland MW, DeWayne Moore D (2007) Development of a theory-based assessment of team member effectiveness. *Educ Psychol Meas* 67(3):505–524
48. Pincus SM, Gladstone IM, Ehrenkranz RA (1991) A regularity statistic for medical data analysis. *J Clin Monit Comput* 7(4):335–345
49. Seber GA, Lee AJ (2012) *Linear regression analysis*, vol 329. Wiley, New York
50. Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. *IEEE Trans Syst Man Cybern* 21(3):660–674
51. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 29(5):1189–1232
52. Airola A, Pahikkala T, Waegeman W, De Baets B, Salakoski T (2011) An experimental comparison of cross-validation techniques for estimating the area under the roc curve. *Comput Stat Data Anal* 55(4):1828–1844
53. Kreyszig E (2010) *Advanced engineering mathematics*. Wiley, New York
54. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. *Phys Rev E* 69(6):066138
55. Chai T, Draxler RR (2014) Root mean square error (rmse) or mean absolute error (mae)?—arguments against avoiding rmse in the literature. *Geosci Model Dev* 7(3):1247–1250
56. Nefzger M, Drasgow J (1957) The needless assumption of normality in Pearson's *r*. *Am Psychol* 12(10):623
57. Zou GY (2007) Toward using confidence intervals to compare correlations. *Psychol Methods* 12(4):399
58. Lovakov A, Agadullina ER (2021) Empirically derived guidelines for effect size interpretation in social psychology. *Eur J Soc Psychol* 51(3):485–504
59. Funder DC, Ozer DJ (2019) Evaluating effect size in psychological research: sense and nonsense. *Adv Methods Pract Psychol Sci* 2(2):156–168
60. Kassarnig V, Bjerre-Nielsen A, Mones E, Lehmann S, Lassen DD (2017) Class attendance, peer similarity, and academic performance in a large field study. *PLoS ONE* 12(11):0187078

61. Jehn KA (1997) A qualitative analysis of conflict types and dimensions in organizational groups. *Adm Sci Q* 42(3):530–557
62. Hasan S, Koning R (2019) Prior ties and the limits of peer effects on startup team performance. *Strateg Manag J* 40(9):1394–1416
63. Prakash R, Beattie T, Javalkar P, Bhattacharjee P, Ramanak S, Thalinja R, Murthy S, Davey C, Blanchard J, Watts C et al (2017) Correlates of school dropout and absenteeism among adolescent girls from marginalized community in North Karnataka, South India. *J Adolesc* 61:64–76
64. Onnela J-P, Rauch SL (2016) Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. *Neuropsychopharmacology* 41(7):1691–1696
65. Rooksby J, Morrison A, Murray-Rust D (2019) Student perspectives on digital phenotyping: the acceptability of using smartphone data to assess mental health. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp 1–14
66. Dwork C, Roth A et al (2014) The algorithmic foundations of differential privacy. *Found Trends Theor Comput Sci* 9(3–4):211–407
67. Hubaux P (2020) Decentralized privacy-preserving proximity tracing. PhD thesis, Fraunhofer HHI
68. Bagdasaryan E, Berlstein G, Waterman J, Birrell E, Foster N, Schneider FB, Estrin D (2019) Ancile: enhancing privacy for ubiquitous computing with use-based privacy. In: *Proceedings of the 18th ACM workshop on privacy in the electronic society*, pp 111–124
69. Langheinrich M (2009) Privacy in ubiquitous computing. In: *Ubiquitous computing*. CRC Press, Boca Raton, pp 95–160
70. Wang JL, Loui MC (2009) Privacy and ethical issues in location-based tracking systems. In: *2009 IEEE international symposium on technology and society*. IEEE, Los Alamitos, pp 1–4
71. Rössler B et al (2001) Der wert des privaten. Suhrkamp, Frankfurt am Main
72. Sweeney Y (2020) Tracking the debate on Covid-19 surveillance tools. *Nat Mach Intell* 2(6):301–304

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)

---