TeamSense: Assessing Personal Affect and Group Cohesion in Small Teams Through Dyadic Interaction and Behavior Analysis with Wearable Sensors

YANXIA ZHANG*, FX Palo Alto Laboratory, USA
JEFFREY OLENICK, Michigan State University, USA
CHU-HSIANG CHANG, Michigan State University, USA
STEVE W. J. KOZLOWSKI, Michigan State University, USA
HAYLEY HUNG, Delft University of Technology, Netherlands

Continuous monitoring with unobtrusive wearable social sensors is becoming a popular method to assess individual affect states and team effectiveness in human research. A large number of applications have demonstrated the effectiveness of applying wearable sensing in corporate settings; for example, in short periodic social events or in a university campus. However, little is known of how we can automatically detect individual affect and group cohesion for long duration missions. Predicting negative affect states and low cohesiveness is vital for team missions. Knowing team members' negative states allows timely interventions to enhance their effectiveness. This work investigates whether sensing social interactions and individual behaviors with wearable sensors can provide insights into assessing individual affect states and group cohesion. We analyzed wearable sensor data from a team of six crew members who were deployed on a four-month simulation of a space exploration mission at a remote location. Our work proposes to recognize team members' affect states and group cohesion as a binary classification problem using novel behavior features that represent dyadic interaction and individual activities. Our method aggregates features from individual members into group levels to predict team cohesion. Our results show that the behavior features extracted from the wearable social sensors provide useful information in assessing personal affect and team cohesion. Group task cohesion can be predicted with a high performance of over 0.8 AUC. Our work demonstrates that we can extract social interactions from sensor data to predict group cohesion in longitudinal missions. We found that quantifying behavior patterns including dyadic interactions and face-to-face communications are important in assessing team process.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

Additional Key Words and Phrases: Behavior analysis; wearable social sensor; interaction; team cohesion; group dynamics

ACM Reference Format:
Yanxia Zhang, Jeffrey Olenick, Chu-Hsiang Chang, Steve W. J. Kozlowski, and Hayley Hung. 2018. TeamSense: Assessing Personal Affect and Group Cohesion in Small Teams Through Dyadic Interaction and Behavior Analysis with Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 39 (September 2018), 22 pages. https://doi.org/0000001.0000001

*Corresponding author

Authors' addresses: Yanxia Zhang, FX Palo Alto Laboratory, 3174 Porter Drive, Palo Alto, CA, USA, yzhang@fxpal.com; Jeffrey Olenick, Michigan State University, East Lansing, MI, USA, olenickj@msu.edu; Chu-Hsiang Chang, Michigan State University, East Lansing, MI, USA, cchang@msu.edu; Steve W. J. Kozlowski, Michigan State University, East Lansing, MI, USA, stevekoz@msu.edu; Hayley Hung, Delft University of Technology, Delft, Netherlands, h.hung@tudelft.nl.

2474-9567/2018/9-ART39 $15.00
https://doi.org/0000001.0000001

1 INTRODUCTION

Assessing team cohesion and effectiveness with pervasive sensing technologies in real-life settings is a growing field in human research [25, 37, 40]. Wearable social sensing platforms, such as Sociometric Badges [36, 48], have been developed to augment ID tags with multiple sensors, such as accelerometers, microphones and optical sensors. These platforms are unobtrusive tools to examine group behaviors and interactions in naturalistic environments without disrupting ongoing tasks.

Prior works demonstrated the effectiveness of using wearable badges to mine social networks in large multi-disciplinary teams, e.g., in corporate settings [15, 36, 48]. Social network structures have been shown to link to team productivity [48], creativity and team performance [15, 28, 36]. The majority of these studies focused on large scale and complex organization settings.

Besides studying groups in large scale corporate settings, long-term wearable recordings also provide an attractive opportunity to study the dynamics of team process, especially for small teams working in confined and isolated long duration missions [25]. Understanding long-term team cohesion and affect states is important as the team must remain together during the entire mission. It is vital to examine team processes and detect their critical states so that potential team dysfunctions can be anticipated. Prior works demonstrated that small groups have inherently different dynamics compared to large groups [13]. This paper focused on small team collaborations for long duration missions in confined spaces.

This paper studies whether individual behavior and social interactions provide useful information in assessing affective states and group cohesion. The topic is well known in social psychology that individual affect is influenced by peers' emotion [20] and the strength of social links in groups [27, 31]. Unlike prior studies focused on tracking individuals with mobile sensing [1, 3, 29, 47], here we focus on the problem of analyzing the affective state of team members in isolated, confined and extreme conditions [25]. Specifically, we analyze a team of six crew members who were in a four-month space exploration simulation mission. Group cohesion is critical in these settings, as a breakdown in the team can cause catastrophic consequences for the success of a mission, and waste years of planning and resources. Unlike typical scenarios in the wild, confined teams are unable to obtain help from outside to moderate their conflicts or to assess demotivation or frustrations of team members. Therefore, being able to infer these states from sensor data is highly desirable.

In such scenarios, the team effect on individual affective state is a bigger factor than can be observed from individual behavior alone. Our work proposes a method that treats team member's affect states and group cohesiveness as a two-class detection (negative vs. non-negative) problem. It takes into account of both the social interactions between team members and the individual behavior features extracted from wearable Sociometric Badges. We analyzed the team behaviors at both individual and group levels. At the individual level, we examined whether there exists any links between wearable social sensor features and personal affect and perceived cohesiveness. We first correlated different types of features to individuals' affect and cohesion ratings collected with experience sampling methods (ESM). We further performed a classification experiment to validate the effectiveness of our approach for detecting personal affect. At the group level, we generated group representations by aggregating features extracted from each team member. Finally, we conducted another classification experiment to validate the performance of detecting group cohesion.

The contributions of this paper are:

- A novel approach for extracting behavior features from wearable social sensors to assess team states. We extracted novel features at the individual level to capture communication and individual activities. The features include 1) $f2f$: amount of face-to-face interactions, which represents the strength of one's social link with other team members; 2) *Individual*: intrinsic features about individual's physical and vocal activities; and 3) *Dyadic*: interaction features which indicate how one person is influenced by and responds to their interaction partners. The features are for quantifying mirroring and influence interaction patterns.

We further constructed group-level features to capture team behaviors. Our approach is validated on a four-month recording of a realistic confined mission conducted in the wild.

- We found that equal number of features representing face-to-face communications and dyadic interactions are predictive for all three dimensions of affect, while less features representing individual activities are found to be predictive for valence compared to arousal. Furthermore, group-level f2f communications demonstrated to be the most significant factor in predicting both task and social cohesion. In sum, we advanced the understanding of group behaviors in small teams with wearable social sensing. We recommend to capture group interaction patterns in sensor design and quantify these behaviors in developing computational models for ubiquitous computing applications.
- To the best of our knowledge, this is the first work that quantifying behavior and interaction patterns from wearable social sensors are applied in assessing longitudinal team processes. Our results provided the initial evidence that predicting personal affect and group cohesion by quantifying behavior features outperformed the baseline methods.

2 RELATED WORK

2.1 Longitudinal Group Interaction Analysis with Social Sensing

Mobile phones have been adopted to study human social interactions in workplace and social events [8, 10, 11]. Eagle and Pentland [10] proposed the Serendipity system which used the Bluetooth in mobile phones to detect and identify people who are close by. Their system exploits the proximity information between two mobile phones to infer relationships among users. A Gaussian mixture model was trained to detect proximity patterns between users and then correlate these patterns with relationship types of users. Do and Gatica-Perez [8] proposed the GroupUs system which uses a probabilistic model to infer types of interaction such as whether the user is interacting with families or colleagues from longitudinal bluetooth data. However, bluetooth has a large scan radius of within 10 meters which does not provide reliable information for studying face-to-face social interactions at small spatial and temporal resolutions [9].

Wearable sensors such as the Sociometric badge are another type of tools for analyzing team interactions and group performance [6, 36, 46, 49]. Nguyen *et al.* applied unsupervised method to discover participants' affiliations and friendship information using the reality mining dataset that was collected by the MIT media Lab from over 100 people wearing Sociometric badges [35]. In these studies, social network analysis based on networks and graph theory is widely used. Waber *et al.* conducted an experiment that collects data from 22 employees wearing the Sociometric badges from a bank in Germany for a month [46]. They correlated temporal changes in social network patterns with individual and groups' performance obtained from e-mail logs and individual self-reports. Their results show that summary statistics computed from their data can only capture ten percent of the variance of responses. They further suggest that a fine level of analysis is needed to discover the important factors from the data. Similar studies were conducted with a group of 67 nurses in a hospital for 27 days and 52 employees from three branches of a bank in the Czech Republic for 20 working days [36]. Through analyzing social network structures and individual behaviors, they identified a negative correlation between amount of face-to-face communication and job satisfaction, and a negative correlation between centrality and group interaction satisfaction.

Tripathi and Burleson studied the relationship of face-to-face interaction strength and individual movement in relation to creativity in teams [44]. In their experiments, 3 teams of 5 to 7 member's movement and interaction data were recorded with sociometric badges and a daily survey for two to four work weeks. They found that average movement is significantly higher for creative days than non-creative days. Using average movement and frequency of face-to-face interactions can predict creativity with 87.5% and 91% accuracy respectively. Another related study from Gloor *et al.* [15] collected sociometric data from 14 graduate students and their instructor

during a one-week seminar. They extract each participant's social network position represented with three metrics: degree centrality, betweenness centrality and contribution index. These metrics were aggregated over the entire five-day period and correlated with individual personality, trust between members and creativity of the team. Their results indicate that more interactions with central people and more members facing each other might predict openness and higher creativity. Recent works applied machine learning to mine longitudinal team interaction data. Zhang *et al.* used topic models to extract social interaction routines from Sociometric badges and identified correlations between team interaction events and their perceived affect and cohesion [49].

2.2 Detecting Affective States in Pervasive Computing

Different methods have been proposed to assess affective states (i.e., stress, emotion and mood), including using biosignals, acoustic, visual and human activities analysis [5]. The majority of these systems rely on sensing biosignals such as Electrodermal Activity (i.e., skin conductance), heart rate and blood flow [22, 41]. The biosignals are usually measured with wearable sensors that are close to human body. Recent works have proposed multi-modal approaches to detect affect states such as depression by combing both visual appearances and dynamics from faces captured at a distance with video cameras [7] or the combination with microphones [33].

In uncontrolled out of lab settings, recent works investigated how daily activities are linked to people's emotional states through continuous unobtrusive sensing via digital traces (e.g., device usage, email, social media, locations) [2, 3, 16, 29, 47]. The majority of these studies focus on assessing mental health such as depression. Hernandez *et al.* studied computer usage in offices to determine stress levels [16]. A number of studies found that mobile phone usage patterns such as calls, GPS location and application usage are correlated to people's depression states [1, 30, 45, 47]. The StudentLife study collected smartphones data from 48 Dartmouth students over 10 weeks to assess their mental health [47]. They analyzed how academic activities have an effect on students' affective mood. Breda *et al.* [45] compared different classification methods for predicting short-term mood from mobile phone data, e.g., number of calls. They found out that contextual data about individuals are useful and can increase predictive performance. Another closely related work is that from [31]. Their study linked the overall sociability (co-locations detected via mobile phone's bluetooth proximity) of 54 participants in the same residential complex to their daily mood. They found that people who exhibited poor mood are the group who have lower overall sociability in that community. However, they did not find correlations between daily variation in sociability and poor mood. They concluded that individual's overall sociability is more important than daily variations in its effect on mood.

Besides using digital traces from mobile phones, Bogomolov *et al.* [3] used behavioral metrics obtained from seven month of 111 subject's mobile phone activities, personality traits and additional context indicators, such as weather information, to recognize daily stress levels. Their model of using the combination of all features achieved a person-independent accuracy of 72.28% for a binary classification of daily stress. Suhara *et al.* [43] investigated predicting depression with self report histories using Recurrent Neural Networks. They used 345,158 total days of logs collected from more than 2,382 users over 22 months. The results show that long-term historical information of a user can improve the prediction accuracy for detecting depression.

2.3 Cohesion Estimation in Teams

From the organizational psychology domain, many studies have been carried out to analyze cohesion. See the review by Salas et al. for a recent review [39]. Cohesion in teams has been studied extensively because of its link with team effectiveness and performance, particularly when conceptualized in terms of the task and social components [39]. In terms of computational approaches, we are aware of three. The first by Olguin-Olguin and Pentland [36], proposed a framework to perform organizational design engineering using the sociometric badge to measure frequency of face to face interaction in combination with other sources such as emails, surveys and

performance data. They define cohesion to be how well acquaintances are connected together based on the frequency of infrared sensing. Therefore the frequency of observation of face to face interactions acted as a proxy of cohesion. We argue in this paper that the concept of cohesion may be more complex. That is aside from frequency of interaction, a number of features related to the coordination of conversational behavior such as turn-taking (as shown by Hung and Gatica-Perez [17]) or paralinguistic mimicry (as shown by Nanninga *et al.* [34]) during meetings are indicative of social or task cohesive behavior. Both works used third party perceptions of cohesion on time slices of 2-5 minutes during a meeting.

In sum, our work differs from existing research in that: firstly, little work investigated the link between affect and face-to-face social interactions in longitudinal team missions. There are no prior studies that apply machine learning for detecting affect states from combining both social interactions and individual activities. Secondly, we are focusing on a small team longitudinal context which has a much smaller scale group structure compared to existing applications in corporate[36], hospital and campus settings [47]. Lastly, compared to analyzing social network structure in complex organizations, our study focuses on the quality of dyadic human-human interactions as an indicator of individual affect and team perceptions of cohesion. Unlike prior work typically using only frequency of face-to-face interactions, we proposed a novel approach of using mimicry features to quantify interaction quality extracted from dyadic interactions.

3 RESEARCH QUESTIONS

The main objective of this work is to explore the value of monitoring long-term team behaviors with wearable social sensors. Conventionally, teams use self-report measures and video recordings. The logged data were usually analyzed off-line. Extensive evidence from behavior research in group studies showed that team states are associated with unique patterns of individual behaviors and social interactions, despite of the tasks that team members are involved in [26]. Maintaining healthy affective states and cohesion within the team are vital for a high performance team. Our aim is to quantify behaviors using proxy extracted from wearable social sensors and conduct empirical evaluations on to what extent we can make inferences about team processes (i.e., individual affective states and group cohesion).

Our work focuses on three types of behavior: *strength of face-to-face communications*, *individual activities* and *dyadic interaction patterns*. The frequency counts of face-to-face interactions – a proxy for the strength of face-to-face communications – were demonstrated to relate to several group concepts. For example, the amount of face-to-face interactions were shown to infer personal relationships [11], roles within groups and group performance in prior works [15, 36]. The second type of behaviors is related to individual activities, i.e., physical and vocal activities. A high base level of physical and vocal activities are typically associated with being excited and interested. Variations in physical and vocal activities usually correspond to fluctuations in people's mood [36]. Recent works have showed initial evidence that individual's daily physical and vocal activities are linked to depression for mental health monitoring [29, 47] and creativity in teams [3]. The last behavior patterns reflect the dyadic social interactions (i.e., mimicry and influence). For example, it is well studied that more mimicry behavior would improve engagement and lead to positive effect in human-human interactions [18, 38, 42]. The mimicry behavior is related to the emotion contagion phenomenon that emotion can be shared among groups [20]. Although it is a crucial criterion in examining group activities rely on unconscious interaction patterns between team members, researchers seldom explored such dyadic interaction patterns as computational models.

Our research questions are: 1) whether the behavior features extracted from the wearable social sensors are linked to personal affect and perceived cohesiveness; 2) at individual level, how accurate and which behavior features are effective for assessing different dimensions of personal affective states; 3) at group level, how we can predict group cohesion by aggregating individual behavior features based on within-team variations?

4 THE SPACE EXPLORATION SIMULATION DATASET

Team cohesion is considered to be a dynamic phenomenon rather than a stable construct [23]. However, currently, research have focused on short period of team collaborations, e.g., meetings [34]. Team cohesion has rarely been investigated for longitudinal settings (such as months or years) [23, 24, 40]. To investigate team cohesion in such settings, the space exploration simulation dataset was collected from a team of six members over four months.

4.1 Subjects

The team is a set of volunteers which were selected from a pool of candidates as we would select for a functional team, that is they were screened on personality, skills and abilities, and an interview process. While in the habitat, they perform a variety of team-oriented research tasks. For example, they complete team missions out of the habitat where they must work together to complete assigned objectives, while working under an approximation of the constraints which would be faced by a Mars flight crew. For example, any outside communication is set to a 20 minute time delay to simulate the communication delays between Earth and Mars. Crews also have some more unstructured personal time where they can work on their own research, or engage in social activities as a group. This more unstructured time occurs more in the evenings.

Crews have a formal structure assigned prior to the mission by the lead scientists on site. Roles are designed to approximate what would be expected on a six person flight crew. They include a formal commander, medical officer, engineer, science officer, architect and biologist. The commander is the one generally to keep the crew on schedule and task, as well as monitor the general state of the crew and liaison with mission control. The medical officer takes the lead on monitoring and maintaining the physical health of the crew. The engineer is tasked with addressing any physical problems with the habitat, such as plumbing leads, electrical outages, etc. The science officer liaisons with the various outside researchers collecting data in the habitat to ensure equipment and protocols are working correctly.

The original six individuals in the crew range in age from 26-60, with a mean of 34. They were 2 males, 4 females, all identified as white and have at least a bachelor's degree. Two individuals indicated some minor prior experience in ICE (Isolated, Confined, and Extreme) conditions (8 months and a half month respectively). Participants completed informed consents for all relevant parties prior to beginning the study. One participant was forced to retire from the experiment early for personal reasons.

Table 1. The space exploration simulation dataset collected over a 4-month mission

User ID	Start - End date	Days	Entries	ESM
1	03-31 to 07-24	107	32308	193 (165)
2	03-31 to 07-24	101	29856	197 (182)
3	03-31 to 07-24	94	43980	151 (150)
4	03-31 to 07-24	54	10868	140 (113)
6	03-31 to 07-24	95	32270	190 (177)

4.2 Data Collection

The data for this study were collected from a team of six individuals who were placed in isolation for a period of four months (see Table 1), with ID *1*, *2*, *3*, *4* and *6*. Subject *5* is identified from the face-to-face interactions, while the associated data is not available due to early withdrawal from the study. The team members live within a confined, shuttle-like structure. The structure is simulating the environment that a crew would inhabit during short-duration space flight. The structure is similar to a container three stories tall and around 6 meters in

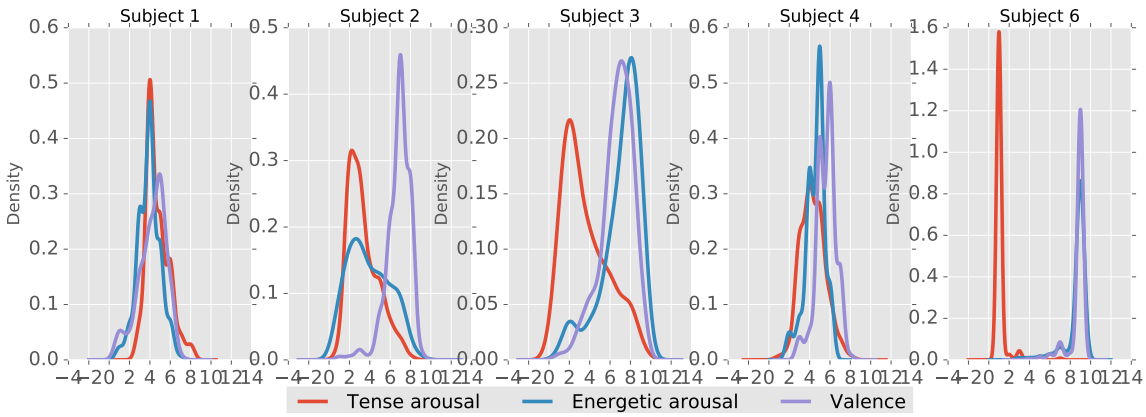

Fig. 1. ESM affective states responses on daily average feelings (1 - strongly disagree; 9 - strongly agree).

diameter. Participants wore badges from Sociometric Solutions during waking hours whenever engaged in potentially social activities (therefore not while exercising, or showering) for multiple hours every day.

The experience sampling method (ESM) was used to collect daily reports. The subjects were asked to fill in the survey twice a day, once in the morning and once in the evening. There are a few times that users provide more than two responses within a day. Due to the schedule of team tasks, the majority of the responses were submitted in the evening. The mornings and afternoons for the team were largely taken by mandatory tasks on more rigid timelines which may not always have left time for filling out the survey at the ideal time. While in the evenings, the team was more free to structure their own time and complete general tasks. In addition, the team is largely autonomous once in the habitat and is not under direct supervision by outside researchers. This is a form of non-compliance from our standpoint, and is a decided limitation with this data set. However, we feel this would represent a clearer problem if we were attempting to reach a level of granularity where we were predicting morning and evening responses separately rather than an aggregate day measurement due to potential issues with recall accuracy when trying to rate their feelings from that morning. In aggregating to the day level, we receive a more global picture of their assessment for the day regardless of when they ratings occurred.

4.3 ESM Data

During the 4 month mission, all members were asked to fill in surveys about their average feelings and experiences with their team mates.

4.3.1 Affective states. The ESM questionnaire asked the participants three questions (9-item scales) that are related to average daily feelings, which measure affect using the Positive and Negative Affect Schedule PANAS survey [14]. The first question of *tense arousal* is whether the subject feel "very relaxed, calm or very nervous and stressed". The second question of *energetic arousal* is asking whether they feel "very sluggish, dull or very active and bright". The last one of *valence* is about feeling "very sad, low or happy and pleased". Figure 1 illustrates the distribution of each subject's responses.

4.3.2 Team cohesiveness. The ESM questionnaire asked about members' experiences with the team over the last hour whether they felt cohesive in terms of "being mutually committed to our work" and "maintaining our interpersonal relationships" (5-item scales). Figure 2 illustrates the distribution of each subject's responses on perceived cohesion. The correlations between these two variables is 0.4 in our data set. This indicates a moderate

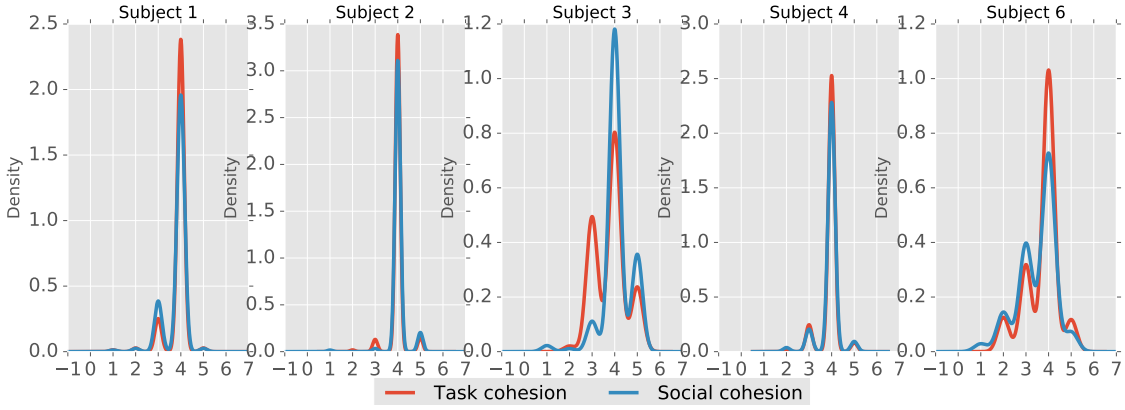

Fig. 2. ESM cohesive responses about perceived cohesion with teammates (1 - strongly disagree; 5 - strongly agree).

correlation between the two types of cohesion, but is not high enough to suggest that they are the same variable. When assessing distributions this distinction is masked as individuals tend to have the same response patterns for these variables across their time in the habitat, but from day to day their ratings on the two fluctuate somewhat, but not completely, independently.

4.4 Wearable Sociometric Badge Data

During the data collection, each participant was wearing a Sociometric Badge (SS Badge) around their neck that records data about movements, audio and social interactions [36].

- *Movement:* The SS badge consists of an accelerometer that computes information on body movement. The accelerometer data is sampled at 20 Hz. The raw data is processed online and only the average values of features computed over a configured time resolution are logged.
- *Audio:* Similar to the accelerometer, the badge has a microphone which samples the audio data at a frequency of 8kHz. The microphone is used to record vocal features, but not the conversation content. The average values of features within the time resolution are saved.
- *Face-to-face interaction:* The badge has an infrared sensor facing outwards with a unique identifier. The number of infrared detections of an id corresponds to the number of seconds (at a configured time resolution) that one subject is within the face-to-face interaction range.
- *Location:* The bluetooth in the badge measures received signal strength to estimate proximity and location information of people nearby wearing a SS badge.

5 METHOD

5.1 Preprocessing the ESM data

Figure 3 illustrates the steps of preprocessing the ESM data on a daily level. We merge and take the average of ratings submitted within a day to represent daily responses. We consider responses submitted after midnight referred to their experiences in the previous day as some subjects were often active in the late evening. In total, we have 451 days of sensor recordings and the corresponding ESM data from 5 subjects.

Affect and cohesiveness at individual levels: for affect state responses, we denote the three survey answers as $\{A_1, A_2, A_3\}$: A_1 (tense arousal) refers to average feelings from very stressed to relaxed, A_2 (energetic arousal)

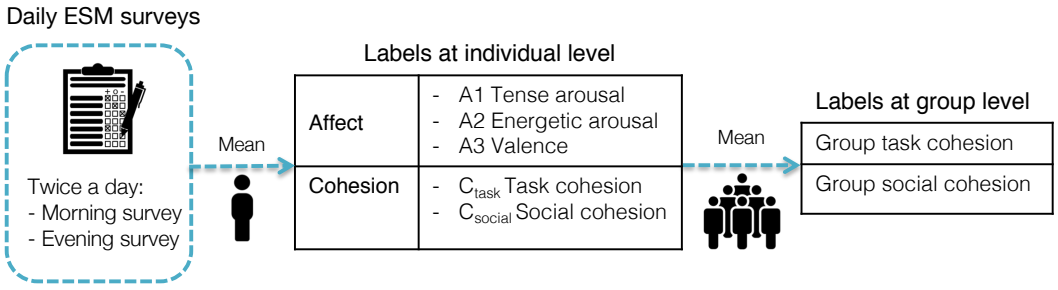

Fig. 3. Preprocessing the ESM data collected daily from each team member.

refers to feelings from very sluggish to bright, and A_3 (valence) refers to feelings from very sad to pleased. We inverted the scores in the "relaxed to stressed" question so that all responses have consistent negative to positive (scores from low to high) directions. The cohesive responses are denoted as $\{C_{task}, C_{social}\}$ where C_{task} represents task cohesion in "mutually committed to our work" and C_{social} represents social cohesion in "interpersonal relationships". The final measurements are the average of all responses submitted in a day.

Cohesiveness at group levels: Figure 4 illustrates the daily fluctuations of group task and group social cohesion of the team over the entire mission. The group task and social cohesion are aggregated from the team members' daily responses of C_{task} and C_{social} respectively. The group cohesion is determined by taking the average of data from days that at least 4 or 5 members' data are available. From the histogram plot of group cohesion distributions (see Figure 4), we can observe that variations exist around their mean values and the average cohesion is high. In total, we obtained 95 days of sensor recordings and the corresponding group cohesion ratings.

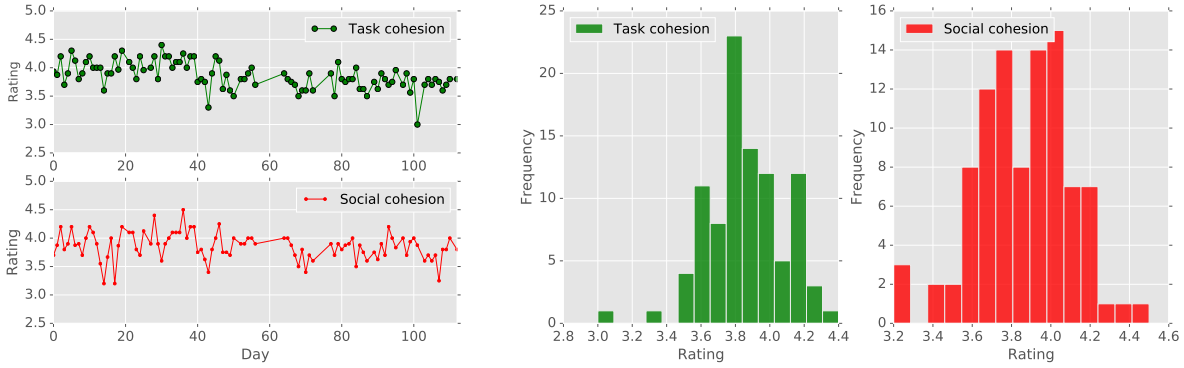

Fig. 4. Group task and social cohesion over the entire mission (left) and their distributions (right).

5.2 Feature extraction from the sociometric badges

We extract three types of features including face-to-face interactions, individual activities and dyadic interactions. The features were first computed every minute from the raw sensor log data. The daily representation for each subject is a vector that concatenates the mean of each feature over the entire day. Figure 5 illustrates our processing pipeline for feature extractions from the wearable sensor data.

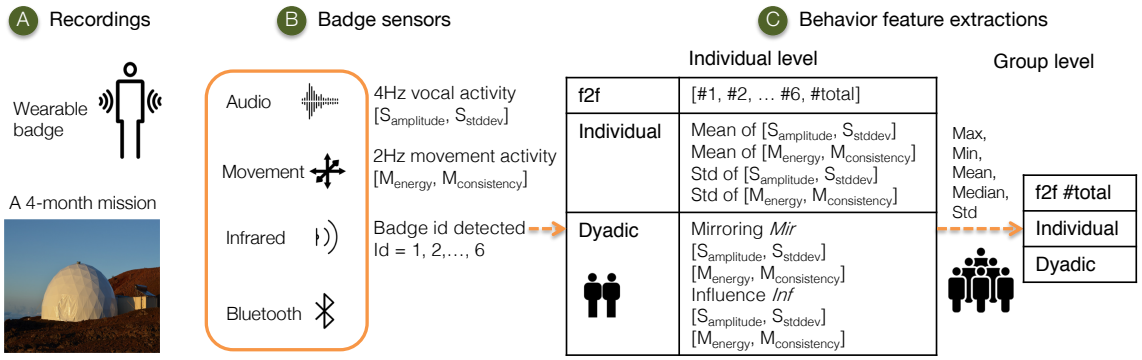

Fig. 5. Extraction of behavior features from the badges wore by each participant.

5.2.1 Strength of face-to-face interactions. Although the sociometric badges are bluetooth-enabled and can detect approximate location information such as a device's MAC address and RSSI, bluetooth signals have a large range of 10 meters which are not reliable for inferring face-to-face social interactions [4, 9]. Dyads might be detected in neighboring rooms as the bluetooth range might be greater than the room distance. The bluetooth information is not suitable in this study as the team lived within a confined, shuttle-like structure that is less than 6 meters in diameter. Instead, the sociometric badges use IR transmissions as a proxy for face-to-face interactions [6]. The infrared sensors on the badge face outwards and transmit IR signals with a cone of height 1 to 2 meters and a radius of ±15 degrees. For one badge to be detected, two badges must have a direct line of sight within the IR signal cone range [46].

For each subject, we extracted the infrared detections that consist of another's badge ID being detected $u \in \{1, 2, 3, 4, 5, 6\}$ and its associated timestamp t. We used the number of detected pings from another badge to represent the frequency of face-to-face interactions. We constructed face-to-face social strength features using an adjacency matrix similar to previous works [37, 44]. The matrix exploits all interaction links at certain time points. Interactions within a sliding window ($W = 1min$) are grouped together. The number of pings detected for every pair of subjects represents strength of communications similar to the network edge strength in the social network structure. For each subject, the face-to-face communication strength is represented using a 6x1 row vector of the adjacency matrix to indicate all possible pairwise interactions. The number of total pings with id u_i detected is represented as $\#u_i$. Besides examining the communication strength between each pair of team members, we also include total amounts of face-to-face interactions denoted as $\#total$.

5.2.2 Individual activities. We extracted individual vocal and movement activities and consistency level features that are shown to be good features in previous works using sociometric badges [35, 36, 44]. The sociometric badge logged the vocal activity features $[S_{amplitude}, S_{stddev}]$ for every 0.25 s from the audio data. We extracted the mean and standard deviation of these two vocal features to represent the distribution of the vocal activity levels, resulting in four features $[S_{amplitude\_avg}, S_{amplitude\_std}, S_{stddev\_avg}, S_{stddev\_std}]$. The movement features $[M_{energy}, M_{consistency}]$ representing physical activities were generated every 0.5 s. The movement energy is computed from the amplitude over three axes of the accelerometer which corresponds to motion intensities. The consistency values are the stability of the movement energy. It has a range between 0 and 1, where 1 indicates no change and 0 indicates maximum variance. Similar to vocal activity features, we computed the mean and

standard deviation of two movement features indicating the distribution of physical activity levels denoted as $[M_{energy\_avg}, M_{energy\_std}, M_{consistency\_avg}, M_{consistency\_std}]$.

5.2.3 Dyadic interactions. Mimicry is one of the most important cues in signaling emotions in dyadic interactions and plays an important role in enabling effective communication and collaboration. People naturally mimic via either verbal content or nonverbal cues, such as postures, facial expressions and aligning movement or speech patterns [18, 42]. We propose to use mimicry features as a proxy to capture the quality of dyadic interactions. Mimicry features are usually measured in two ways: event-based and correlation-based [32, 34]. The event-based measure is commonly adopted in applications that focus on recurring events of pre-defined behaviors, such as laughter, head nodding and hand gestures. The correlation-based measures examine recurrence and similarities between two time series in raw signal and feature space without explicit event detections.

We computed the correlation metrics for quantifying mimicry [12]. For every pair of time series data from all subjects, we segmented dyadic interactions based on the face-to-face events detected from the IR sensors. A sliding window of 10 minutes was used to classify the time series segment into dyadic interaction if face-to-face interactions were detected within the window. Then we computed the minute by minute Pearson correlations using the time series data from the two interaction partner, denoted as the *mirroring Mir* feature. The *Mir* feature is used to quantify the similarity of two individuals' behavior patterns over time. The second metric we proposed is the *influence* Inf feature where we compute the minute by minute cross correlations between two time series data using Fourier transform. The Inf feature represents whether one individual's behavior would lead another individual to follow and exhibit similar behaviors. The resulting dyadic interaction features were computed from all pairwise individual activities including vocal activities $[S_{amplitude}, S_{stddev}]$ and movement activities $[M_{energy}, M_{consistency}]$.

5.3 Group feature extraction

We extracted group level features to capture distributions of individual behaviors within the team. From the badge data, we used the combined feature set (f2f, Individual, Dyadic). In the f2f category, we used only the $\#total$ feature that represents the total amount of face-to-face interactions. A vector of size 17x1 was generated per day for each team member. From all team members, the daily representation of group feature was extracted by computing the max, min, median, mean and standard deviation per feature category. This resulted in a group feature vector of size 85x1 from the badge data. For the baseline method that uses ESM affect ratings, the group level feature was generated by using the daily responses $\{A_1, A_2, A_3\}$ from all team members. Similar to using the badge data, we computed the max, min, median, mean and standard deviation in each affect rating category. The final group feature vector size is 15x1 for the ESM affect baseline.

6 EVALUATIONS

6.1 RQ1: which features are linked to personal affect and cohesion?

The first experiment examines whether the proposed behavior feature representations are linked to individual affect and perceived cohesion. This is achieved by correlating the average daily representation of each feature to self-reported ratings obtained from the ESM data. We obtained in total of 451 samples from all 5 subjects. This includes 451 daily ESM ratings on affect $\{A_1, A_2, A_3\}$ and cohesion $\{C_{task}, C_{social}\}$ and features of three categories as described in the previous section: 1) strength of face-to-face interactions (7 variables); 2) individual activities (8 variables) and 3) dyadic interactions (8 variables).

We used the Pearson correlation coefficients to identify important features that are correlated to each ESM rating. For face-to-face interaction features from subject u_i, the correlations were only computed using number of interactions with other team members, because $\#u_i$ is zero for that subject. We also removed samples for days that interactions do not exist in computing correlations for interaction related features. For each ESM rating, we

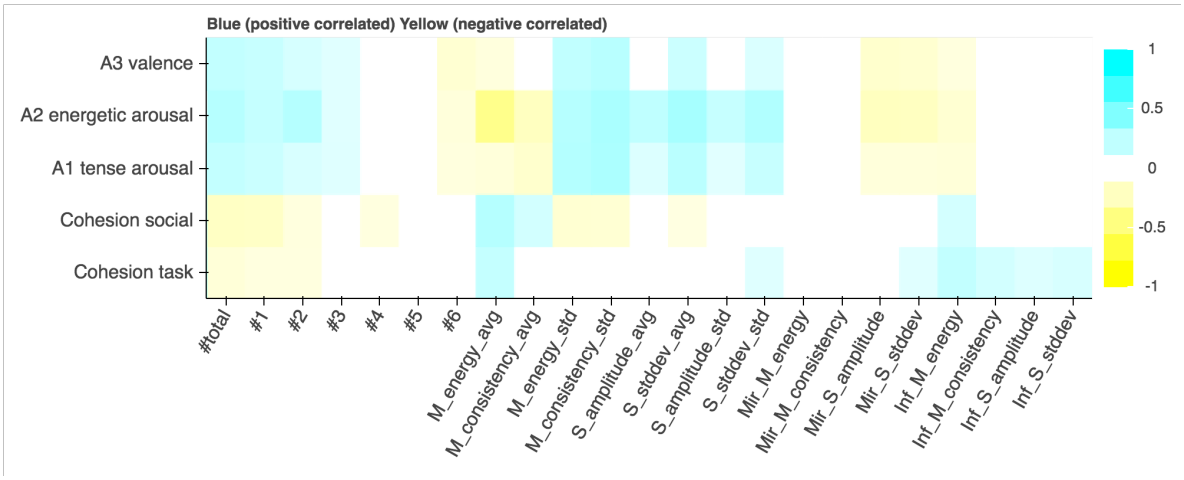

Fig. 6. Correlations between daily ESM responses (y axis) and proposed individual behavior and interaction features (x axis). Blue: positive correlation; Yellow: negative correlation; White: no significant correlations.

computed a list of correlation and p-values generated from independent tests of a single feature. We applied the Benjamini-Hochberg procedure to control the False Discovery Rate (FDR) for multiple hypothesis tests at level $\alpha = 0.05$. The correlation results are shown in Figure 6. We explain the details of our findings as below:

Affect: The correlations between each feature and the three affect categories show consistent relationships. There are weak positive correlations ($r \in (0.1, 0.3), p < 0.05$) between the total number of face-to-face interactions $\#total$, interactions with three members (#1, #2, #3) in the team and affective states. In contrast, amount of interactions with team member #6 has a significant weak negative correlation ($r \in (-0.2, -0.1), p < 0.05$) with individual's affect. There is no correlations found for the amount of interactions with subject 5, who withdrew early from the study. We note that the majority of individual activity variables are correlated to affect but with different effects. The average of movement energy and consistency are negatively correlated to affect, while the rest of the individual activity features are positively correlated to affect. In particular, average daily movement energy is negatively correlated ($r = -0.46, p < 0.001$) to A_2 of feelings from very sluggish to bright. The average fluctuations of vocal activities S_{stddev} is found to be positively correlated to A_1 of feelings from very stressed to relaxed ($r = 0.27, p < 0.001$) and A_2 of feelings from very sluggish to bright ($r = 0.35, p < 0.001$). In terms of the dyadic interaction features, mirroring in vocal activities Mir_S are consistently shown to be negatively correlated ($r \in (-0.3, -0.1), p < 0.001$) to affect. Another feature Inf_M representing influences in movement M_{energy} is shown to be weakly and negatively correlated ($r \in (-0.2, -0.1), p < 0.05$) to affect.

Individual perception of cohesion: Task cohesion C_{task} shows weak and negative correlation ($r \in (-0.2, -0.1), p < 0.05$) to the amount of face to face interactions with two team members and the total amounts (#1, #2, $\#total$). Two individual activity features are positively correlated ($r \in (0.1, 0.3), p < 0.05$) to task cohesion C_{task}, including mean movement energy and vocal variations. All four Inf features and one Mir feature extracted from dyadic interactions are positively correlated ($r \in (0.1, 0.3), p < 0.05$) to C_{task}.

Similar to task cohesion, the strength of social interaction in total and with four subjects are found to be weakly negatively correlated ($r \in (-0.2, -0.1), p < 0.05$) to social cohesion. For individual activities, movement levels play an important role. All four features are found to have positive correlations especially the average movement energy ($r = 0.29, p < 0.001$). Only one out of eight dyadic interaction features, influence in movement energy Inf_M_energy, is weakly and positively correlated ($r = 0.17, p < 0.01$) to social cohesion C_{social}.

Table 2. Affect classification results of mean accuracies using our methods and the baseline method. The Majority Classifier (MC) assigns all predictions to the majority class.

	Our method				Baseline
Category	Combined	f2f	Individual	Dyadic	MC
A_1: *tense arousal*	**84.05%***	83.57%	83.09%	83.04%	82.78%
A_2: *energetic arousal*	73.20%	70.52%	**76.32%***	70.10%	70.59%
A_3: *valence*	**71.88%** *	68.54%	66.88%	69.13%	69.05%

Notes: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 3. Affect classification results of F1 values using different feature sets and the baseline Majority Classifier (MC) method. Besides results indicated as $^{**}p < .01$, all differences between the best and the MC classifier are significant with $p < .001$

	Our method				Baseline
Category	Combined	f2f	Individual	Dyadic	MC
A_1: *tense arousal*	0.771	**0.772****	0.764	0.762	0.763
A_2: *energetic arousal*	**0.696*****	0.642	0.690	0.645	0.599
A_3: *valence*	**0.628*****	0.602	0.597	0.619	0.574

Notes: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

In this analysis, we are interested in whether there are general effects between the wearable sensor features and affect and perceived cohesion. We want to screen the extracted sensor features using all the data and have an initial idea whether it is sensible to explore further to perform predictions. In our correlation analysis, we pooled all the feature data from different subjects and estimated a single ESM rating. Our assumptions are that all individuals from this team share similar characteristics and work towards a common team goal, as they have been trained to form a highly cohesive team. The limitation of this assumption is that we consider subjects to be sampled from the same distribution and ignore any variations among the subjects. This approach might be limited in groups where individuals share no similarity and therefore should be treated independently.

6.2 RQ2: can we detect individual affect states with these features?

In the second experiment, we investigate whether the proposed behavior and interaction features are effective in assessing daily affective states. We conducted a classification experiment to evaluate the detection performance of person-dependent affect states. The affect data was split into two classes of negative and non-negative groups. The affect state A of negative class include all cases with scores < 5.

We compared four feature sets based on information described in Section 5.2. The first feature set $f2f$ (7x1) includes information only from the infrared detections that represent the strength of face-to-face interactions. The second feature set *Individual* (8x1) includes information only from individual activities. The third feature set *Dyadic* (8x1) contains all Mir and Inf features detected from dyadic interactions. The fourth set *Combined* (23x1) uses all features from the previous three sets. A feature vector is extracted for each day per subject.

To evaluate the classification performance, we compared three supervised learning methods including logistic regression classifier using the 'sag' solver, support vector machine (SVM) with a linear kernel and random forest trees. The dataset is split into 70% for training and 30% for testing for each subject. During the training process, the input feature vectors were preprocessed and standardized by the mean and standard deviation on the training

Table 4. Affect classification results of AUC values using different feature sets and the baseline random classifier.

	Our method				Baseline
Category	Combined	f2f	Individual	Dyadic	
A_1: *tense arousal*	0.564	**0.576*****	0.542	0.543	0.5
A_2: *energetic arousal*	**0.556****	0.531	0.536	0.542	0.5
A_3: *valence*	0.527	**0.540*****	0.526	0.521	0.5

Notes: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

set. The same transformation is applied on feature vectors in the testing set using mean and standard deviation obtained from the training set. A feature selection procedure was performed to remove features with zero variance. For SVM, C parameter was selected from $\{10^{-2}, 10^{-1}, \ldots, 10^{3}\}$. For random forest classifier, the maximum depth of the tree was selected from {3, 10, None}. The model parameters were selected with fivefold cross validations on the training data by grid search. We conducted 50 iterations of randomly sampling the training and testing sets.

Table 2 presents the mean classification accuracies over all subjects using support vector machine which gives the best results. We did not evaluate the classification performance for subject 6 due to extremely small number of negative samples (less than 1) in his data. We used the majority classifier (MC) that assigns all predictions to the majority class as the baseline. The baseline accuracy is very high because the class distributions are highly imbalanced. From the results, we can see that using all the features combined achieved highest accuracies in classifying A_1 tense arousal and A_3 valence. Interestingly, using only individual activity feature set achieves the highest accuracy in classifying A_2 energetic arousal. We also computed the weighted F1 scores from two classes using different feature sets as shown in Table 3, where F1 score conveys the balance between the precision and the recall. The F1 scores confirmed that the Combined feature set achieved better or comparable performance compared to each individual feature set. Remarkably, detecting A_2 energetic arousal achieved the best performance with both higher accuracy (76.32% vs 70.59%) and F1 score (0.696 vs 0.599) compared to the baseline MC classifier. The AUC values of using different feature sets are shown in Table 4, compared to a baseline random guessing classifier (AUC = 0.5). To compare the performance of the best results of our approach with the baseline method, we used the paired t-test to compare the differences of classification results with exactly the same training data and test with the same data during the random sampling.

6.3 RQ3: can we classify group cohesion states with aggregated group features?

The previous two experiments examined how effective the proposed features perform at individual levels. Our last experiment evaluates whether aggregated group level features are effective in assessing team cohesion. Specifically, we are interested in group cohesion over the entire mission and how individual's behavior has an effect on the team process. In total, we have 95 days of samples after preprocessing.

Experimental setup: We conducted a classification experiment to assess the effectiveness of the group level features in assessing group cohesive states. The class label is assigned based on the mean of task cohesion and social cohesion: negative if the group cohesion is less than the mean values and positive if greater than the mean values. Besides using the majority classifier as a baseline, we also compared the performance of using self-reported affect ratings. This is motivated by correlations found between affect and cohesive ratings. We compared three classifiers: logistic regression classifier using the 'sag' solver, a linear SVM and random forest trees. The model parameters were selected with 5 fold cross-validation on the training data by grid search. The final performance was evaluated using the leave-one-out (LOO) cross-validation strategy that takes all the days except one for training and tests on the one day sample being left out.

Table 5. Classification accuracies and F1 values for group task and social cohesion using group level features extracted from the badge data and the baseline method that uses self-reported ESM affect ratings and the Majority Classifier (MC).

Ratio	Category	Accuracy			F1	
		Badge	ESM affect	Baseline MC	Badge	ESM affect
100% data	*Task cohesion*	**74.68%**	69.62%	59.49%	**0.748**	0.694
	Social cohesion	**64.05%**	49.44%	51.69%	**0.640**	0.496
80% data	*Task cohesion*	**80.30%**	62.12%	51.52%	**0.803**	0.621
	Social cohesion	**64.62%**	56.92%	50.77%	**0.647**	0.569

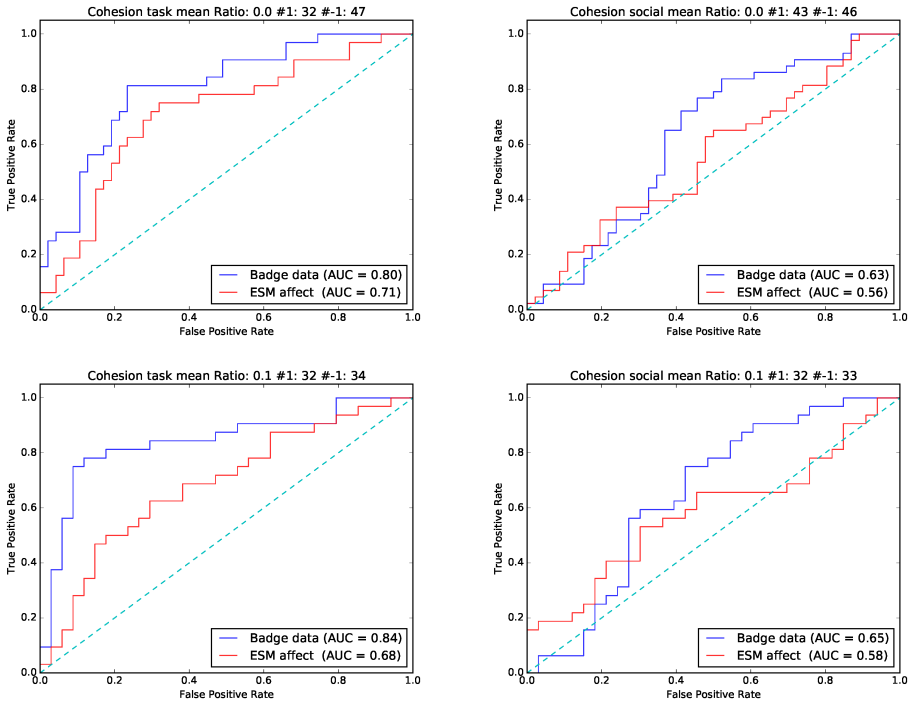

Fig. 7. Classification results of AUC-ROC for group task and social cohesion. Top row: 100% data; bottom row: 80% data

Results: Table 5 summarizes the accuracies and weighted F1 values from two classes for our proposed method and the baseline methods. The logistic regression classifier was used as it achieved the best performance in this task. We presented results of evaluations on using 100% of the samples and removing 20% of the samples around the mean values. Similar to previous works that removed 50% of the samples around the mean [17, 34], removing part of the samples in the middle would make the task easier by eliminating ambiguous data points. Our results show that using the group level features extracted from the badge data outperforms the baseline methods in both group task and social cohesion. If using 80% of samples, the top 40% (positive class) and bottom 40% (negative class) divided by the mean value, there is a huge improvement in both accuracy (over 80%) and F1 values. We plot

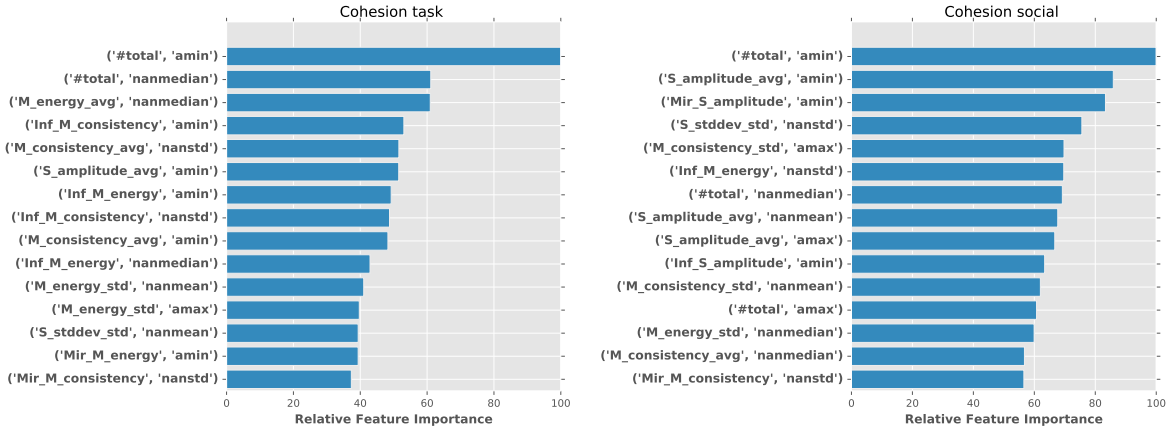

Fig. 8. Top 15 predictive features among the total 85 features for group task and social cohesion. The relative feature importance is normalized by the maximum of the feature coefficient values and multiplied by 100.

the ROC curves for using the badge data and ESM ratings in Figure 7. The badge features achieve an AUC of 0.8 in classifying task cohesion using 100% samples, much higher than random guessing (AUC = 0.5) and using ESM affect (AUC = 0.71). The AUC value increases to 0.84 when using 80% of data, much higher than using ESM affect (AUC = 0.68). The badge features achieve an accuracy of 64% when using 100% of data in classifying social cohesion, much higher than using ESM affect (49.44%) and MC (51.69%). The performance is similar when using 80% of the data. The accuracy results are consistent with the other two evaluation metrics: the F1 values (Table 5) and the AUC values (Figure 7). Overall, these results confirmed that behavior and interaction features extracted from group levels are effective in assessing team cohesion compared to the baseline methods.

We examined the importances of the proposed group level features that contribute to the task and social cohesion. The feature importances are ranked using the coefficient values of the logistic regression with L_2 penalty, which implicitly sets some of the weak features weights to close to zero. Figure 8 illustrate the top 15 predictive features among the total 85 group level features. The relative feature importance is computed by first taking the absolute value of the feature coefficient and then normalized by the maximum of the feature coefficient values multiplied by 100. As shown in Figure 8, the minimum number of *#total* communications among all team members is shown as the most important feature in both task and social cohesion. In particular, the minimum number of *#total* communications plays a significant role compared to other features in predicting task cohesion, while this effect is less apparent in predicting social cohesion.

7 DISCUSSION

7.1 Affect Estimation

When examining the three aspects of affect, we found that the proposed features are most effective in predicting A_2 energetic arousal. Surprisingly, individual activities achieved the best accuracy. When combined with other feature sets, the classification accuracy decreased, and there was little improvement in F1 values. These results suggest that the A_2 energetic arousal level is most related to individual behaviors. The prediction results are consistent with the correlation analysis shown in Figure 6, which shows larger correlations with individual activity features compared to the other two affect dimensions. When examining the performance in detecting A_3 valence, social interaction features (f2f and Dyadic) achieve better accuracy than individual behaviors. The

accuracy and F1 values are best when combining all three feature sets. However, the AUC value is not the best for the combined category, which indicates that the classifier can achieve good performance if only a good threshold is chosen. Our method provides minimum advantages in detecting the A_1 tense arousal compared to the baseline method, although the majority of features were identified to correlate with A_1 (see Figure 6). This is likely due to the reason that the class distributions are highly imbalanced (baseline MC accuracy 82.78%), thus making the classification task challenging. Nevertheless, the f2f features achieve the best F1 values and AUC scores, which suggest high importance in assessing A_1 tense arousal.

Previous studies have only reported the correlations between digital traces from mobile phone usage and depression states [2, 3, 16, 29, 47]. Our work provides both correlation analysis and machine learning prediction results on three dimensions of affective states. We observed that valence and tense arousal are more challenging dimensions to predict compared to the energetic arousal. The correlation between the valence variable (sad-pleased) and the tense arousal (calm-nervous) variable in the raw data is -0.78, which indicates these two dimensions measure similar constructs in our data. To improve the predictions of these two dimensions that are related to human internal feelings, we suggest using physiological signals to gather additional information, for instance, heart rate and skin conductance (i.e., perspiration). On the other hand, the energetic arousal referring to fatigue level (tired-awake) is closely related to physical activities, as shown in our results that the individual activity feature set achieved the best accuracy.

Our results confirm that combining both social interaction and individual behavior features benefits the detection of personal affective states. Our results also suggest that social interactions can infer affective states. These results contribute novel insights to existing findings in affect detection. For example, Moturu *et al.* [31] used Bluetooth co-location as a proxy for social contacts and carried out their study with a large community of 54 participants in a residential complex. They did not find a link between daily sociability and mood. Our results show that the link between the amount of daily social interactions and affect existed in a small team. One possible explanation for the difference is that the team members in our study remained within a confined space for a longer duration, which compelled frequent social contacts. Another possible explanation is that Bluetooth co-location information is too coarse to infer precise social interactions. The infrared detections of the sensors used in our study measure fine grained face-to-face communications, thus a more accurate proxy for representing social interactions.

7.2 Group Cohesion Estimation

Table 5 shows that exploiting the behavioral features outperforms both the majority classifier baseline, as well as the ESM affect baseline. This suggests that the group perception of both social and task cohesion is influenced by the mirroring features extracted from the Sociometric data. Surprisingly, when we make the task easier by removing 20% of the data, estimation of task cohesion improves while there is no difference for social cohesion. This is likely to be caused by the narrow distribution of samples for the team's social cohesion compared to the task cohesion. When considering the strong correlation between the group aggregated ESM affect ratings and cohesion, we see that our proposed method outperforms the more competitive baseline classifier.

We can explore the relation between the group aggregated ESM affect ratings and social and task cohesion ratings in more detail in Figure 7. The top row shows the results with 100% of the data while the bottom row shows the results for 80% data. In terms of task cohesion, the higher AUC of the badge extracted features show the importance of the badge extracted features in the case of our data. Interestingly, when considering the easier task with 80% of the data, the difference between using the ESM affect and the badge data increases. This suggests that the group-level badge features are indeed more indicative of the perception of group cohesion compared to the affect ratings alone. For the case of social cohesion, the analysis is more nuanced as the ROC when using the badge data does not consistently outperform the ESM affect data.

The observation that the social cohesion is easier to predict than task cohesion is surprising, as past studies have shown that social cohesion tends to be easier to detect [17, 34]. One possible explanation of this could be related to the cohesion of the team itself. Unlike the teams in prior work, in our data, the members have worked together before the mission and many steps have been taken to maximize team performance and cohesion. This could mean that team members are more motivated by the task and therefore exhibit more coordinated social behavior when team tasks are going well compared to the social cohesion.

As shown in Figure 8, the minimum of face-to-face communications among team members turns out to be the most important feature. This can be caused by anomaly behaviors that individual members lost motivations or opportunities to engage with other members, which suggest an alarm for intervention to maintain team cohesion. We also observe the top 15 features in estimating task cohesion consist of dyadic interaction patterns from movements only. This indicates that quantifying non-verbal patterns is crucial in assessing task cohesion in scenarios where tasks require complex coordinations compared to prior works in short-term meeting setups [17, 34]. On the other hand, the top 15 features in estimating social cohesion are linked to a range of different factors. Our results further show that using affect states collected from the ESM data in the team can infer team process such as cohesion. Our results suggest that the variations and consistency of affective states at the group level play an important role in assessing team states, in contrast to the majority of previous works focusing on individual affect [1, 29, 47]. In particular, the effect is larger in social cohesion compared to task cohesion. One possible reason could be that positive mood increase people's ability to be more open to collaboration and communication, while persistent negative mood can deteriorate interactions between team members [19]. Task cohesion, on the other hand, appears to be closely attributed to the coordination of operations in specific practice [39] and is less easy to be predicted by team members' affect states.

7.3 Small Groups in Short-term versus Longitudinal

One major difference in estimating short-term and longitudinal group cohesion is the availability of sensing modalities. High resolution video-audio data can be easily recorded in short-term settings (i.e., meetings or workshops) where the environments are instrumented with cameras and microphones. Research to understand team process commonly examines explicit behaviors such as facial expressions, gestures, postures and verbal communications [13, 17, 33, 34]. However, continuously measuring team processes for a long duration with these modalities becomes impossible. Wearable social sensors, such as the Sociometric badges used in this study, provide the opportunity to measure spontaneous behaviors among team members in longitudinal settings. Our results show that estimating task cohesion achieves better performance in our longitudinal settings compared to existing works in short-term meeting setups [17, 34]. One of the main reasons for this effect is that we captured long-term face-to-face communications which might correspond to not only explicit conversations but reflect a more general level of social contexts of individuals within the team. Another factor can be that we included dyadic interaction features that quantify the non-verbal dyadic interaction patterns. For example, the movement mimicry patterns are found to be among the top predictive features. This result demonstrates the usefulness of using wearable social sensors to measure team activities that involve non-verbal coordinations.

The link between affect and social cohesion suggests designing interventions on the individual affect level can be a possible way to improve team process. If one member is experiencing a negative mood, other members could use social closeness to mitigate his/her mood. Task cohesion appears to be closely related to the nature of team activities so that appropriate feedback on tasks can be provided. For instance, professional training in communication and coordination can be provided when task cohesion is low. In short-term small group settings, such as meetings or brainstorm workshops, verbal activities are good indicators of task cohesion. Hence, feedback on turn-taking patterns could act as a group-level intervention [21]. It still remains an open challenge to estimate cohesion for team activities which involve little verbal communications. For instance, in a hospital,

the interactions between the doctors and assistants are extremely critical for successful surgeries. These tasks consist of a large amount of implicit non-verbal behaviors and coordinations among team members. Measuring non-verbal behavior using wearable sensors is one promising way to capture and quantify these implicit signals.

7.4 Limitations and Generalization of the Results

The simulation in which this data was collected is specifically designed for the findings to generalize to deep space flight crews, particularly for missions to Mars. The habitat was developed given the known plans from NASA for a mission to Mars and what that would entail, and as such has received funding from the agency to study such teams. We do expect the general findings to generalize to other teams that may find themselves in high stakes settings for extended periods of time, during which they must rely heavily on one another. These directly include scientific teams sent to Antarctica and mountain climbing teams. Some principles that emerge from this research are also likely to apply to military teams, such as special forces groups, but more research will be needed given the kinds of situations that such groups are likely to encounter.

This work verified behavior markers inspired from existing social psychology research. Our ultimate goal is to develop a system that allows the team to function automatically when critical situations happen. This is particularly important in scenarios when no instant external intervention is available. The sensing system should be as un-intrusive as possible, so that they do not disrupt ongoing work flow of the team members. Especially in high-stake industry, such as the military, aviation and health care, every second counts and can make a huge difference. Moreover, teams are the fundamental components nested in organizations. The success of large organizations including the space industry, corporations and academic institutions all depend on performing teams [39]. Although our method is applied to assess affect and cohesion, the behavior features are holistic and can be applied to understand how social interactions and individual behavior patterns influence other team outcomes, such as performance, creativity, productivity and employee satisfaction.

8 CONCLUSION

We used wearable sensors to analyze how social interactions and individual behaviors have an effect on personal affect and group cohesion in a small team. Our study found that individuals' affective states are not only correlated to their own activities but also linked to social interactions within the team. By combing both individual and social interaction features, we achieved better classification performance in detecting personal affect states compared to the baseline methods. Our results also demonstrated that behavior features extracted from individuals' wearable data can be aggregated to group level and are effective in assessing group cohesion. Our work provided empirical evidence to predict individual affect and group cohesion with wearable sensors. Open challenges still exist; we need to design sensors that measure team constructs for their particular applications. Finally, our results provide some indicators of link between individual affect and team cohesion to mitigate team breakdown.

9 ACKNOWLEDGEMENT

Data used in the preparation of this manuscript were provided by S. W. J. Kozlowski, S. Biswas, & C.- H. Chang. Their research, which generated the data, was supported by the National Aeronautics and Space Administration (NASA; NNX13AM77G, S.W.J. Kozlowski, Principal Investigator). Any opinions, findings, conclusions and recommendations expressed are those of the authors and do not necessarily reflect the views of NASA

REFERENCES

[1] Joost Asselbergs, Jeroen Ruwaard, Michal Ejdys, Niels Schrader, Marit Sijbrandij, and Heleen Riper. 2016. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: an explorative study. *Journal of medical Internet research* 18, 3 (2016).

[2] Anja Bachmann, Christoph Klebsattel, Matthias Budde, Till Riedel, Michael Beigl, Markus Reichert, Philip Santangelo, and Ulrich Ebner-Priemer. 2015. How to Use Smartphones for Less Obtrusive Ambulatory Mood Assessment and Mood Recognition. In *Adjunct*

Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct). ACM, New York, NY, USA, 693–702. https://doi.org/10.1145/2800835.2804394

[3] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex (Sandy) Pentland. 2014. Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. In *Proceedings of the 22Nd ACM International Conference on Multimedia (MM '14)*. ACM, New York, NY, USA, 477–486. https://doi.org/10.1145/2647868.2654933

[4] Tjeerd W Boonstra, Mark E Larsen, Samuel Townsend, and Helen Christensen. 2017. Validation of a smartphone app to map social networks of proximity. *arXiv preprint arXiv:1706.08777* (2017).

[5] Shizhe Chen and Qin Jin. 2015. Multi-modal Dimensional Emotion Recognition Using Recurrent Neural Networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (AVEC '15)*. ACM, New York, NY, USA, 49–56. https://doi.org/10.1145/2808196.2811638

[6] Tanzeem Choudhury and Alex (Sandy) Pentland. 2003. Sensing and Modeling Human Networks Using the Sociometer. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers (ISWC '03)*. IEEE Computer Society, Washington, DC, USA, 216. http://dl.acm.org/citation.cfm?id=946249.946901

[7] Hamdi Dibeklioğlu, Zakia Hammal, and Jeffrey Cohn. 2018. Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding. *IEEE Journal of Biomedical and Health Informatics* PP, 99 (2018), 1–1. https://doi.org/10.1109/JBHI.2017.2676878

[8] Trinh Minh Do and Daniel Gatica-Perez. 2011. GroupUs: Smartphone Proximity Data and Human Interaction Type Mining. In *Proceedings of the 2011 15th Annual International Symposium on Wearable Computers*. 21–28. https://doi.org/10.1109/ISWC.2011.28

[9] Trinh Minh Do and Daniel Gatica-Perez. 2013. Human Interaction Discovery in Smartphone Proximity Networks. *Personal Ubiquitous Comput.* 17, 3 (March 2013), 413–431. https://doi.org/10.1007/s00779-011-0489-7

[10] Nathan Eagle and Alex (Sandy) Pentland. 2005. Social Serendipity: Mobilizing Social Software. *IEEE Pervasive Computing* 4, 2 (April 2005), 28–34. https://doi.org/10.1109/MPRV.2005.37

[11] Nathan Eagle and Alex (Sandy) Pentland. 2006. Reality Mining: Sensing Complex Social Systems. *Personal Ubiquitous Comput.* 10, 4 (March 2006), 255–268. https://doi.org/10.1007/s00779-005-0046-3

[12] Jens Edlund, Mattias Heldner, and Julia Hirschberg. 2009. Pause and gap length in face-to-face interaction. In *INTERSPEECH*. 2779–2782. https://doi.org/10.7916/D82F7WT9

[13] Daniel Gatica-Perez. 2009. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing* 27, 12 (2009), 1775 – 1787. https://doi.org/10.1016/j.imavis.2009.01.004 Visual and multimodal analysis of human spontaneous behaviour:.

[14] Phillip Gee, Timothy Ballard, Gillian Yeo, and Andrew Neal. [n. d.]. *Chapter 5 Measuring Affect Over Time: The Momentary Affect Scale*. 141–173. https://doi.org/10.1108/S1746-9791(2012)0000008010

[15] Peter A Gloor, Francesca Grippa, Johannes Putzke, Casper Lassenius, Hauke Fuehres, Kai Fischbach, and Detlef Schoder. 2012. Measuring social capital in creative teams through sociometric sensors. *International Journal of Organisational Design and Engineering* 2, 4 (2012), 380–401.

[16] Javier Hernandez, Pablo Paredes, Asta Roseway, and Mary Czerwinski. 2014. Under Pressure: Sensing Stress of Computer Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 51–60. https://doi.org/10.1145/2556288.2557165

[17] Hayley Hung and Daniel Gatica-Perez. 2010. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *IEEE Transactions on Multimedia* 12, 6 (Oct 2010), 563–575. https://doi.org/10.1109/TMM.2010.2055233

[18] Marco Iacoboni. 2009. *Mirroring people: The new science of how we connect with others*. Farrar, Straus and Giroux.

[19] Peter J Jordan, Sandra A Lawrence, and Ashlea C Troth. 2006. The impact of negative mood on team performance. *Journal of Management & Organization* 12, 2 (2006), 131–145.

[20] Janice R. Kelly and Sigal G. Barsade. 2001. Mood and Emotions in Small Groups and Work Teams. *Organizational Behavior and Human Decision Processes* 86, 1 (2001), 99 – 130. https://doi.org/10.1006/obhd.2001.2974

[21] Taemie Kim, Agnes Chang, Lindsey Holland, and Alex Sandy Pentland. 2008. Meeting Mediator: Enhancing Group Collaborationusing Sociometric Feedback. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)*. ACM, New York, NY, USA, 457–466. https://doi.org/10.1145/1460563.1460636

[22] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A Database for Emotion Analysis ;Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (Jan 2012), 18–31. https://doi.org/10.1109/T-AFFC.2011.15

[23] Steve W.J. Kozlowski, S. Biswas, and CH Chang. August 2013 to December 2017. *Measuring, monitoring, and regulating teamwork for long duration missions. National Aeronautics and Space Administration (NNX13AM77G)*. Technical Report.

[24] Steve W.J. Kozlowski, CH Chang, and S. Biswas. January 2017. Measuring team functioning via multiple methods. *NASA Human Research Program Investigators' Workshop* (January 2017).

[25] Steve W.J. Kozlowski, Georgia T Chao, CH Chang, and R Fernandez. 2015. Team dynamics: Using "big data" to advance the science of team effectiveness. *Big data at work: The data science revolution and organizational psychology. New York, NY: Routledge Academic* (2015).

[26] Steve W.J. Kozlowski and Georgia T. Chao (in press). 2018. Unpacking Team Process Dynamics and Emergent Phenomena: Challenges, Conceptual Advances, and Innovative Methods. *American Psychologist* (2018).
[27] David Krackhardt, N Nohria, and B Eccles. 2003. The strength of strong ties. *Networks in the knowledge economy* (2003), 82.
[28] B. Lepri, J. Staiano, G. Rigato, K. Kalimeri, A. Finnerty, F. Pianesi, N. Sebe, and A. Pentland. 2012. The SocioMetric Badges Corpus: A Multilevel Behavioral Dataset for Social Behavior in Complex Organizations. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. 623–628. https://doi.org/10.1109/SocialCom-PASSAT.2012.71
[29] Yuanchao Ma, Bin Xu, Yin Bai, Guodong Sun, and Run Zhu. 2012. Daily Mood Assessment Based on Mobile Phone Sensing. In *Proceedings of the 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks (BSN '12)*. IEEE Computer Society, Washington, DC, USA, 142–147. https://doi.org/10.1109/BSN.2012.3
[30] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. Towards Multi-modal Anticipatory Monitoring of Depressive States Through the Analysis of Human-smartphone Interaction. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 1132–1138. https://doi.org/10.1145/2968219.2968299
[31] Sai T Moturu, Inas Khayal, Nadav Aharony, Wei Pan, and Alex (Sandy) Pentland. 2011. Using social sensing to understand the links between sleep, mood, and sociability. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*. IEEE, 208–214.
[32] Philipp Müller, Michael Xuelin Huang, and Andreas Bulling. 2018. Detecting Low Rapport During Natural Interactions in Small Groups from Non-Verbal Behaviour. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 153–164. https://doi.org/10.1145/3172944.3172969
[33] P. M. Müller, S. Amin, P. Verma, M. Andriluka, and A. Bulling. 2015. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 663–669. https://doi.org/10.1109/ACII.2015.7344640
[34] Marjolein C. Nanninga, Yanxia Zhang, Nale Lehmann-Willenbrock, Zoltán Szlávik, and Hayley Hung. 2017. Estimating Verbal Expressions of Task and Social Cohesion in Meetings by Quantifying Paralinguistic Mimicry. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI 2017)*. ACM, New York, NY, USA, 206–215. https://doi.org/10.1145/3136755.3136811
[35] T. Nguyen, D. Phung, S. Gupta, and S. Venkatesh. 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)*. 47–55. https://doi.org/10.1109/PerCom.2013.6526713
[36] Daniel Olguín Olguín and Alex (Sandy) Pentland. 2010. Sensor-based organisational design and engineering. *International Journal of Organisational Design and Engineering* 1, 1-2 (2010), 69–97.
[37] Alex (Sandy) Pentland. 2012. Society's Nervous System: Building Effective Government, Energy, and Public Health Systems. *Computer* 45, 1 (Jan. 2012), 31–38. https://doi.org/10.1109/MC.2011.299
[38] Alex (Sandy) Pentland and Tracy Heibeck. 2010. *Honest signals: how they shape our world.* MIT press.
[39] Eduardo Salas, Rebecca Grossman, Ashley M. Hughes, and Chris W. Coultas. 2015. Measuring Team Cohesion: Observations from the Science. *Human Factors* 57, 3 (2015), 365–374. https://doi.org/10.1177/0018720815578267
[40] Eduardo Salas, Scott I Tannenbaum, Steve WJ Kozlowski, Christopher A Miller, John E Mathieu, and William B Vessey. 2015. Teams in space exploration: A new frontier for the science of team effectiveness. *Current Directions in Psychological Science* 24, 3 (2015), 200–207.
[41] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. *IEEE Transactions on Affective Computing* 3, 1 (Jan 2012), 42–55. https://doi.org/10.1109/T-AFFC.2011.25
[42] Mariëlle Stel and Roos Vonk. 2010. Mimicry in social interaction: Benefits for mimickers, mimickees, and their interaction. *British Journal of Psychology* 101, 2 (2010), 311–323.
[43] Yoshihiko Suhara, Yinzhan Xu, and Alex (Sandy) Pentland. 2017. DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 715–724. https://doi.org/10.1145/3038912.3052676
[44] Priyamvada Tripathi and Winslow Burleson. 2012. Predicting Creativity in the Wild: Experience Sample and Sociometric Modeling of Teams. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 1203–1212. https://doi.org/10.1145/2145204.2145386
[45] Ward van Breda, Johnno Pastor, Mark Hoogendoorn, Jeroen Ruwaard, Joost Asselbergs, and Heleen Riper. 2016. Exploring and comparing machine learning approaches for predicting mood over time. In *Innovation in Medicine and Healthcare 2016*. Springer, 37–47.
[46] Benjamin N Waber, Daniel Olguin Olguin, Taemie Kim, Akshay Mohan, Koji Ara, and Alex (Sandy) Pentland. 2007. Organizational engineering using sociometric badges. *Available at SSRN 1073342* (2007).
[47] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM,

New York, NY, USA, 3–14. https://doi.org/10.1145/2632048.2632054

[48] Lynn Wu, Benjamin N. Waber, Sinan Aral, Erik Brynjolfsson, and Alex (Sandy) Pentland. 2008. Mining Face-to-Face Interaction Networks using Sociometric Badges: Predicting Productivity in an IT Configuration Task. *SSRN eLibrary* (2008). https://doi.org/10.2139/ssrn.1130251

[49] Yanxia Zhang, Jeffrey Olenick, Chu-Hsiang Chang, Steve W. J. Kozlowski, and Hayley Hung. 2018. The I in Team: Mining Personal Social Interaction Routine with Topic Models from Long-Term Team Data. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. ACM, New York, NY, USA, 421–426. https://doi.org/10.1145/3172944.3172997