Few Labels are Enough! Semi-supervised Graph Learning for Social Interaction

Nicola Corbellini[1] Jhony H. Giraldo[2] Giovanna Varni[3] Gualtiero Volpe[1]
[1] Casa Paganini, InfoMus, DIBRIS, University of Genoa, Italy
[2] LTCI, Télécom Paris - Institut Polytechnique de Paris, France
[3] Department of Information Engineering and Computer Science, University of Trento, Italy
nicola.corbellini@edu.unige.it

Abstract

Endowing machines with social intelligence is a fundamental goal of artificial social intelligence. Dealing with human-centered phenomena requires, however, a considerable amount of manually annotated data, making data annotation a costly and challenging task that hinders the training of supervised learning algorithms. In this study, we apply an approach grounded on Graph Convolutional Network (GCN) to alleviate the annotation burden. As a test bed, we select emergent states analysis with specific reference to the team potency. At first, we build the POTENCY dataset by fusing three datasets on social interaction. Next, we compute a set of multimodal features characterizing the social behavior of the team members and the team as one. Finally, we feed the POTENCY dataset to a semi-supervised GCN, trained on a binary node classification task, with variable amounts of labels. We show that GCN can assign team potency labels to an unlabeled team in the dataset by using only a few labeled examples (i.e., 10% of data), with performances comparable to or higher than those of two baseline algorithms carrying out the same task in a fully supervised way.

1. Introduction

Artificial social intelligence aims to equip machines with the ability to analyze and interpret social phenomena [47]. Among these phenomena, *emergent states* are dynamic constructs that arise from team actions and interactions [42]. Examples of emergent states include cohesion, team potency, and the transactive memory system. Studying emergent states has particular relevance for artificial social intelligence since they can characterize the affective, motivational, behavioral, or cognitive state of a team [38]. Such investigation presents unique challenges due to the multiparty and nuanced nature of team interactions [27].

Prior research has explored computational approaches for automated analysis of emergent states (*e.g.*, [36, 26, 46, 13]). These approaches mostly rely on fully supervised learning paradigms, demanding a large number of labeled examples. Annotating social interaction data is however, a challenging, time-consuming, and costly task that requires expert raters and precise coding schemes, such as the Advanced Interaction Analysis for Teams (Act4Teams) scheme [27]. The labor-intensive task of data annotation is further compounded by the challenge of determining the appropriate number of annotators and demonstrating annotation certainty and reliability [2]. Moreover, data labeling involves making critical choices that can influence the results, such as the method of data unitizing [8].

In this paper, we propose an approach grounded on Graph Convolutional Network (GCN) to alleviate the annotation burden and enable effective analysis of emergent states even with a limited amount of annotated data. Specifically, we leverage relational information among data by modeling them as a graph and employing a GCN in a semi-supervised setting.

We assess the approach by applying it to automated analysis of team potency, *i.e.*, "*the collective believe that a team can be effective*" [24]. This emergent state was selected because of its link with group performance and satisfaction [23, 33]. Endowing a machine with the ability to cope with low potency scenarios could, indeed, allow us to devise socially intelligent machines that can support the team's well-being and have a positive impact on its functioning [25]. More specifically, we measure the performances of binary classification of team potency (*i.e.*, low vs. high potency) with a variable amount of labeled examples (from 1% of the dataset size, up to 100%).

The main contribution of this work is as follows: we show that the GCN-based approach can successfully assign team potency labels by using only a few labeled examples (i.e., 10% of data). The performances of the GCN are comparable to or higher than those of two baseline algorithms carrying out the same task in a fully supervised way.

2. Related Work

The cost of manual annotation remains a long-standing and open problem, especially for computational approaches to human-centered phenomena. Various techniques were proposed to address different aspects of this issue.

Supervised methods using end-to-end deep learning techniques were used to tackle uncertainty and unreliable labels. For instance, Prabhu *et al*. [41] proposed an end-to-end model to address subjectivity in emotion annotations, whilst Wang and colleagues [50] incorporated an agreement-oriented loss function to model label unreliability in their deep learning model. Deep learning methods, however, often require a significant amount of labeled data to avoid overfitting [21], and their resource-intensive nature contributes to environmental pollution [53].

Semi-automated methods such as Active Learning (AL) and semi-supervised algorithms were applied to alleviate the burden of data labeling. AL delegates the annotation procedure to a machine learning method, which selects unlabeled samples based on a *query strategy* and presents them to the annotator for labeling. Various AL approaches were proposed [56, 55, 49]. Zhang and colleagues show potential savings of up to 79.17% of labels for emotion recognition in spoken interactions using only audio data [55]. Effectively combining AL with modern deep learning algorithms remains, however, an open problem [43]. Furthermore, AL techniques can encounter challenges with technically complex data, as designing an effective *query strategy* is non-trivial, and uninformative examples may be selected. Recently, Voß and colleagues [48] tackled multimodal dis-/agreement classification in human-robot interactions and YouTube videos using semi-supervised deep architectures. While their work demonstrates promising results, it still relies on a supervised branch for the final classification, necessitating a significant amount of labeled examples to generalize effectively.

In the specific area of analysis of emergent states, previous works mostly explored supervised settings relying either on data taken from already annotated datasets or on data captured and annotated for the purpose. For example, Hung *et al*. classified teams as high or low on cohesion from manually annotated meetings recordings [26]. Maman and colleagues [37] analyzed the temporal dynamics of cohesion by using the GAME-ON dataset [36], which was recorded and manually annotated for the purpose. Lee *et al*. [32] developed a computational model of interpersonal trust through hand-coded nonverbal social behaviors and explored the temporal dynamic of the construct by means of hidden Markov models. Coming to team potency, Castro-Hernandez and colleagues [7] addressed it as both a regression and a classification problem, but their work was narrowly focused on virtual teams and required the time-consuming recording of a dataset of students' interactions. Corbellini *et al*. [13] tackled multimodal team potency classification using traditional machine learning algorithms on publicly available datasets, which were manually annotated for the purpose. All these studies suffered from the annotation burden that required a lot of effort and delayed the achievement of results.

3. Background

In the following, we briefly describe the emergent state we selected to test our approach (*i.e*., team potency) and the class of artificial neural networks we adopted to detect it (*i.e*., Graph Convolutional Networks).

Team potency. Potency is an emergent team phenomenon known for its link with team performance and satisfaction [23, 33, 45]. Many studies show that team potency grounds on participatory [19], supportive [14], and cohesive [31] social interactions among the team members. More in detail, team potency is a motivational construct, meaning that it reflects "*team beliefs relating to the intensity, direction, and effort regulation toward team task accomplishment*" [42]. As an emergent phenomenon, it "*arise[s] from interactions among individuals, [is] shaped by the context over time, and manifest[s] at higher levels of the system*" [30]. Let us consider the case in which one person has low confidence in herself. Still, she could be confident of her team's success, leaning on her teammates [24]. Accordingly, team potency is not related to what the individual thinks; rather, it is a shared belief in the team as one. Finally, team potency is *task-independent* as it is not related to the task the team is involved in. Rather, it is a general idea that the team will perform well in a broad spectrum of circumstances [42].

Graph Convolutional Network (GCN). Let $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ be an undirected, unweighted graph, where $\mathcal{V} = \{1, \dots, N\}$ is the set of N vertices, $\mathcal{E} \subseteq \{(p, q) \mid p, q \in \mathcal{V} \text{ and } p \neq q\}$ is the set of edges between nodes p and q, and $\mathbf{X} \in \mathbb{R}^{N \times F}$ is a feature matrix, where F is the number of node features. The adjacency matrix of the graph is denoted as $\mathbf{A} \in \{0, 1\}^{N \times N}$, where $\mathbf{A}(p, q) = 1 \; \forall \; (p, q) \in \mathcal{E}$, and $\mathbf{A}(p, q) = 0$ otherwise. Moreover, $\mathbf{D} \in \mathbb{R}^{N \times N}$ is a diagonal matrix representing the graph degree such that $\mathbf{D}(p, p) = \sum_{q=1}^{N} \mathbf{A}(p, q) \; \forall \; p = 1, \dots, N$. Thus, a GCN layer is defined as follows [28]:

$$\mathbf{H}^{(l+1)} = \sigma(\tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}^{(l)} \mathbf{W}^{(l)}), \tag{1}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ and $\tilde{\mathbf{D}}$ is its degree matrix. $\mathbf{H}^{(l)}$ is the output of layer l, with $\mathbf{H}^{(0)} = \mathbf{X}$. $\mathbf{W}^{(l)} \in \mathbb{R}^{d_{in} \times d_{out}}$ is the matrix of learnable parameters in layer l, and $\sigma(\cdot)$ is a non-linear activation function. We omit the bias term in (1) for simplicity.

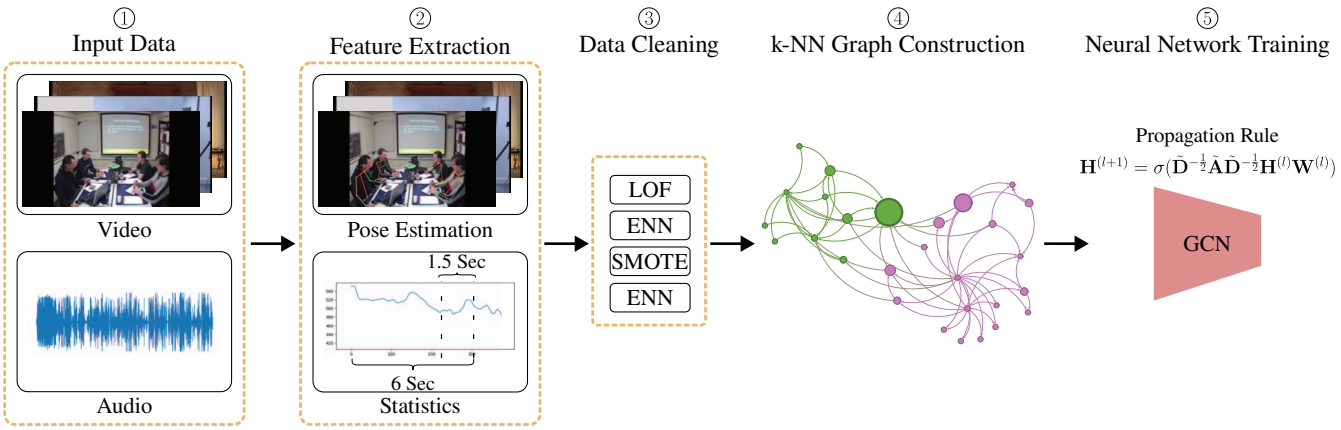

Figure 1. The pipeline of the approach: features are extracted from audio/video recordings and MoCap data when available. Next, a *per-dataset* cleaning is applied to remove outlier (LOF) and labels overlapping (ENN). Then, we augment the datasets (SMOTE) and we combine them to get the POTENCY dataset. A k-NN graph is built to feed a GCN that classifies high vs. low potency nodes.

4. Methodology

The pipeline of the proposed method is illustrated in Figure 1. To begin, a preprocessing step is applied to each dataset, involving the removal of outliers and filtering of source signals. Next, a collection of features is computed to characterize the multimodal social behavior of the team members and of the team as one. The feature set then undergoes a cleaning process, and the resulting data is augmented by using a combination of over and under-sampling techniques. Finally, we build a graph by applying a k-Nearest Neighbors (k-NN) approach and proceed to train a GCN for potency classification.

4.1. Task Formulation

We formulate the team potency binary classification task as a semi-supervised node binary classification problem. Let $G = (\mathcal{V}, \mathcal{E}, \mathbf{X})$ be the graph constructed from the feature matrix $\mathbf{X}$. Similarly, let $\mathcal{V}_l \subseteq \mathcal{V}$ be a subset of nodes with associated labels[1] $y \in \{0, 1\}$ describing the team as being either low or high on team potency. The task is to classify the set of remaining unlabeled nodes $\mathcal{V}_u = \mathcal{V} \setminus \mathcal{V}_l$ in a transductive scenario. This means the algorithm uses the whole feature matrix $\mathbf{X}$ and the set of labeled nodes $\mathcal{V}_l$ to give a label to unlabeled nodes. In other words, the information which is implicit in the unlabeled examples is used to give them a label.

4.2. Dataset

Due to the lack of publicly available datasets on potency, we build the POTENCY dataset by aggregating parts of existing datasets concerning teams, and in which team members (i) are engaged in a specific task, and (ii) are collaborating to achieve a shared goal. Here below, few information is given about each of the selected datasets:

AMI [6] consists of audio-video recordings in a meeting scenario in which participants discuss seated around a table. Two families of meetings are available: (i) *remote controller design task*, in which team members discuss to solve a design task, and (ii) *miscellanea*, in which team members discuss various topics, *e.g.*, fictitious planning of an office move, research, and so on. Among the available data, we use the mixed audio tracks of the team members and the videos' lateral views.

MULTISIMO [29] consists of audio and video recordings[2] of 3-people teams solving a quiz. The participants are seated around a table, two of them are the players and the third one is the game facilitator. We select the mixed audio tracks of the team members and the videos that captured the whole scene.

GAME-ON [36] consists of audio, video, and MoCap recordings of 3-people teams freely moving while playing an escape game. We select video data from the frontal view camera and we mix together the individual audio tracks.

Table 1 reports the technical specifications of every dataset.

Table 1. Technical specifications of the selected datasets.

Dataset	Video	Audio	MoCap
AMI	350 × 280px, 25fps	16 kHz, Mono	N/A
MULTISIMO	1920 × 1080px, 30fps	48 kHz, Stereo	N/A
GAME-ON	1280 × 720px, 50fps	48 kHz, Mono	50 Hz

The POTENCY dataset finally consists of 18 teams having from 3 to 4 team members each. Table 2 summarizes the composition of the dataset.

[1] In semi-supervised learning scenarios $0 < \frac{|\mathcal{V}_l|}{|\mathcal{V}|} << 1$.

[2] The dataset also includes Kinect skeletons that are not retained for analysis because they are noisy in some of the segments of interest

Table 2. The POTENCY dataset composition

Dataset	Teams	Samples	Labels (low vs high)
AMI	6	162	81 vs 81
MULTISIMO	6	138	33 vs 105
GAME-ON	6	153	54 vs 99
POTENCY	18	453	168 vs 285

4.3. Preprocessing

Due to the different technical specifications of the datasets, preprocessing is tailored to each of them.

Audio. We re-sample the audio tracks to 44.1kHz and convert them to monophonic. Hence, we filter them with the Audacity[3] noise-reduction algorithm.

Movement. We process AMI and MULTISIMO video recordings using OpenPose [5] for body pose estimation. We remove the outliers from the resulting 2D body poses with a Hampel filter [35]. Outliers are defined as those points exceeding 3 times the Median Absolute Deviation in a window of 7 frames for both datasets. Then, a Savitzky-Golay filter [44] is applied to smooth the trajectories. The filter parameters are: (i) window size of 25 frames with a polynomial of order 3 for AMI, and (ii) window size of 30 frames with a polynomial of order 2 for MULTISIMO. GAME-ON already provides cleaned 3D positional data at 50Hz. Thus, we project them on a 2D plane by removing the depth axis.

Unitizing. Teams' recordings are segmented into 15s non-overlapping segments for each data source (*i.e*. audio, video, and movement trajectories). Such a window size was already successfully applied for annotating affective social behaviors [8] drawing on the results of Ambady and colleagues [1]. A subset of the remaining segments is retained for analysis. We remove segments not relevant to the group potency such as transitions between tasks. The process yields a total of 151 segments.

4.4. Annotation

Commonly, team potency is assessed by administering a questionnaire, following a *referent-shift composition model* [9]. According to this model, the focus of the assessment is shifted from the individuals to the team. This means that, prior justification of within-group agreement, each person's rating is averaged among team members to obtain a team-level score. This constraint is necessary to verify that there is consensus among team members and it legitimates scholars to aggregate the individual ratings. To annotate potency, we adopt the 8-items scale developed by Guzzo *et al*. [24]. Specifically, in this study, the 7-point version of the scale (from *To no extent* (1) to *To great extent* (7)) is used [14]. The work of [20] indeed shows that this scale is a reliable tool for potency assessment.

We recruited two annotators (*i.e*., psychologists trained for team analysis) to rate the audio-visual segments. Annotators were instructed about the task by watching sample segments from each of the datasets, that are not included in the annotation procedure. Each annotator viewed the segments in random order and rated them over Guzzo's scale.

To verify the annotators' agreement, we compute the $r^*_{wg(j)}$ index [34]. This index is commonly used to assess interrater agreement for Likert-type responses and to support the averaging of individual ratings to the group level [39]. $r^*_{wg(j)} \in [-1, 1]$, where -1 is maximum dissent and 1 is maximum consensus. Following the best practices outlined in [39], we consider $r^*_{wg(j)} \geq 0.7$ a reliable consensus. Any disagreement was solved by verbal discussion between the annotators.

As reported in Section 4.1, we formulate the problem as a binary classification task (high potency vs. low potency). To get the binary labels, we rearrange the interrater average scores as follows:

$$y = \begin{cases} \text{low, if } score \leq 4 \\ \text{high, if } score > 4, \end{cases} \tag{2}$$

being $score$ the average potency score assigned by the annotators to a segment.

Finally, since team potency is an emergent phenomenon, we cannot expect it to change too often in a short period of time [52]. Consequently, we review the ground truth to clean artifacts resulting from the interrater average. We define as outliers those segments having an opposite label with respect to both the previous and the next 15s segments over time. That is, in case the previous and the next segments have a different label with respect to the middle one, the label of the middle segment is changed accordingly. As a result, we change the label of three segments.

4.5. Feature extraction

Since feature extraction is not the focus of this research, we leverage the state-of-the-art on analysis of team behavior and emergent states (*e.g*., [22, 37, 13, 15]). We select a collection of features that were already successfully used in existing works. Features describe the behavior, in its paralinguistic and movement components, of both individual team members as well as the team as one.

Concerning paralinguistic features, since individual audios are not available for all datasets, we only compute those referring to the team as one. Specifically, we use OpenSmile [17] to compute the functional set of the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [16]. Such a set was previously successfully exploited to classify cohesion dynamics [37] and for predicting team performances [57]. Since GeMAPS loudness can be affected by the

[3]Audacity® software is copyright © 1999-2021 Audacity Team. The name Audacity® is a registered trademark

recording settings of the specific dataset, we normalize its mean and standard deviation with respect to those computed on the whole source audio.

Regarding movement features, we extract two individual features, *i.e.*, Upper Body Energy (UBE) [13, 37] and the distance of the head of each team member from the team barycenter (BarDist) [22, 13, 37]. We also extract one feature concerning the team as one, *i.e.*, the entropy of the displacement of the heads (HeadsEnt) [15]. Computations are performed following the algorithms proposed in the referenced literature. In the case of UBE, since GAME-ON takes place in a scenario in which people are free to move, we account for this by removing the chest energy of each team member. This step allows us to remove the influence of walking from the obtained values.

After extracting the features, we compute their mean, standard deviation, skewness, and kurtosis over 6s sliding windows with 1.5s of overlap. As a result, our feature vector consists of 88 audio and 33 movement features, for a total of 121 behavioral features.

4.6. Feature Set Cleaning and Augmentation

From the feature set, we filter out outliers *per dataset* using the Local Outlier Factor (LOF) algorithm [4] with a neighborhood equal to 20. We remove a total of 25 samples (1.85% AMI, 5% MULTISIMO, and 9.8% GAME-ON). Furthermore, to prevent overfitting, we augment the feature set *per dataset*. We first apply the Edited Nearest Neighbors (ENN) [51] method to clean regions of the feature set where a dense overlapping of discordant labels occurs. The number of neighbors for the search is set to three. Next, the Synthetic Minority Over-sampling Technique (SMOTE) ([10]) method is adopted for over-sampling the feature set. We use the default number of neighbors equal to five. Finally, ENN with a neighborhood equal to seven is run to remove possible noise due to the over-sampling, as suggested in [3].

4.7. Graph Construction

To build the graphs, we exploit the geometrical information in the POTENCY dataset. Specifically, let $\mathbf{X} \in \mathbb{R}^{N \times F}$ be the features of the augmented dataset defined in Section 4.6, being N the number of samples and F the size of the feature vector (*i.e.*, 121 as shown in Section 4.5). Each node in $\mathcal{V}$ is thus an F-dimensional vector. We use a k-Nearest Neighbours (k-NN) method with $k = 129$ to construct the graph, *i.e.*, we connect each node to its k nearest neighbors. We assign a weight equal to 1 to these k-NN connections. Finally, we force the graph to be undirected.

5. Experiments

To show the effectiveness of our approach, we compare the GCN algorithm against two baselines: a Label Propagation Classifier (*LPA)* [58], and the Variational Splines of Pesenson (*V-Splines*) [40]. Training is performed with an increasing percentage of labels, *i.e.*, 1%, 5%, 10%, 25%, 50%, 75%, and 100%. 35 different seeds are used for each algorithm.

We evaluate the algorithms in a Leave-One-Team-Out (LOTO) setting. Namely, for each algorithm (i) we iteratively mask the samples of one team so that all such samples are unlabeled, (ii) we then perform training using a random subset of nodes sampled from the remaining nodes, and (iii) we finally compute a metric to evaluate the performances of the algorithms by comparing the labels assigned to the samples of the team that was left out with the ground-truth (*i.e.*, the labels provided by the manual annotation). Steps (ii) and (iii) are repeated for all percentages of labeled samples listed above. As a metric, we choose the F1 score.

The overall effect of the algorithms and label percentage on the performance metric is assessed with a two-way repeated measures permutation ANOVA [54]. Thus, we investigate (i) the main effect of the algorithms on the performance at each label percentage, and (ii) the main effect of label percentages on the performance of each algorithm. In both cases, if statistical significance is detected, we perform pairwise permutation t-test post-hoc[12] and False Discovery Rate (FDR) correction.

5.1. Implementation Details

We train GCN using Pytorch Geometric 2.3 [18] on an Nvidia Geforce Rtx 3090 GPU. We adopt the same architecture described in [28]. Specifically, we define the GCN architecture with 2 graph convolutional layers with 32 hidden units and the ReLU activation function. We use the Adam optimizer with a learning rate of 0.01 to minimize the binary cross-entropy loss with an L2 regularization factor of 10^{-4}. We further regularize using a Dropout layer with 50% of dropout probability. GCN is trained for 200 epochs.

6. Results and Discussion

Table 3 shows the obtained F1 scores vs. the label percentages for GCN and the two baselines. F1 scores are reported for the low and high potency classes. The average F1 score is reported as well, and also displayed in Figure 2.

The two-way analysis of the average F1 scores shows that both factors (*i.e.*, the algorithms and the percentage of labels) have a significant effect on the performance metric ($p < .001$ for both). The post-hoc analyses for the main effect of the algorithm on the performance at each percentage of available labels confirm that GCN significantly outperforms the baseline algorithms for amounts of labels lower than 75%. In detail, GCN performs better than the LPA classifier for amounts of labels lower than 75% ($1\% : p < .001$, $5\% : p < .001$, $10\% : p < .001$, $25\% : p < .001$, $50\% : p < .001$); no significant difference

Table 3. Per-class and average (±std) F1 scores for an increasing amount of labels and for the three algorithms.

Class	Model	1%	5%	10%	25%	50%	75%	100%
$F1_{low}$	LPA	$0.29_{\pm 0.09}$	$0.32_{\pm 0.07}$	$0.40_{\pm 0.02}$	$0.42_{\pm 0.02}$	$0.48_{\pm 0.03}$	$0.53_{\pm 0.02}$	$0.55_{\pm 0.00}$
	V-Splines	$0.53_{\pm 0.06}$	$0.54_{\pm 0.04}$	$0.54_{\pm 0.04}$	$0.55_{\pm 0.04}$	$0.55_{\pm 0.03}$	$0.55_{\pm 0.02}$	$0.56_{\pm 0.00}$
	GCN	$\mathbf{0.57}_{\pm 0.08}$	$\mathbf{0.61}_{\pm 0.06}$	$\mathbf{0.63}_{\pm 0.03}$	$\mathbf{0.66}_{\pm 0.02}$	$\mathbf{0.66}_{\pm 0.02}$	$\mathbf{0.66}_{\pm 0.02}$	$\mathbf{0.65}_{\pm 0.01}$
$F1_{high}$	LPA	$\mathbf{0.65}_{\pm 0.11}$	$\mathbf{0.82}_{\pm 0.03}$	$\mathbf{0.82}_{\pm 0.01}$	$\mathbf{0.78}_{\pm 0.01}$	$\mathbf{0.76}_{\pm 0.01}$	$\mathbf{0.75}_{\pm 0.00}$	$\mathbf{0.73}_{\pm 0.00}$
	V-Splines	$\mathbf{0.68}_{\pm 0.04}$	$0.68_{\pm 0.02}$	$0.67_{\pm 0.02}$	$0.66_{\pm 0.02}$	$0.65_{\pm 0.02}$	$0.64_{\pm 0.01}$	$0.64_{\pm 0.00}$
	GCN	$\mathbf{0.65}_{\pm 0.06}$	$0.64_{\pm 0.04}$	$0.63_{\pm 0.03}$	$0.62_{\pm 0.02}$	$0.62_{\pm 0.02}$	$0.62_{\pm 0.01}$	$0.61_{\pm 0.01}$
$F1_{avg}$	LPA	$0.47_{\pm 0.05}$	$0.57_{\pm 0.04}$	$0.61_{\pm 0.01}$	$0.60_{\pm 0.01}$	$0.62_{\pm 0.02}$	$\mathbf{0.64}_{\pm 0.01}$	$\mathbf{0.64}_{\pm 0.00}$
	V-Splines	$\mathbf{0.60}_{\pm 0.04}$	$0.61_{\pm 0.02}$	$0.61_{\pm 0.03}$	$0.60_{\pm 0.02}$	$0.60_{\pm 0.02}$	$0.60_{\pm 0.01}$	$0.60_{\pm 0.00}$
	GCN	$\mathbf{0.61}_{\pm 0.04}$	$\mathbf{0.62}_{\pm 0.03}$	$\mathbf{0.63}_{\pm 0.02}$	$\mathbf{0.64}_{\pm 0.02}$	$\mathbf{0.64}_{\pm 0.01}$	$\mathbf{0.64}_{\pm 0.01}$	$\mathbf{0.63}_{\pm 0.01}$

The significant results are shown in **bold**. Multiple bold values for the same percentage of labeled examples mean that the difference between such values is not statistically significant.

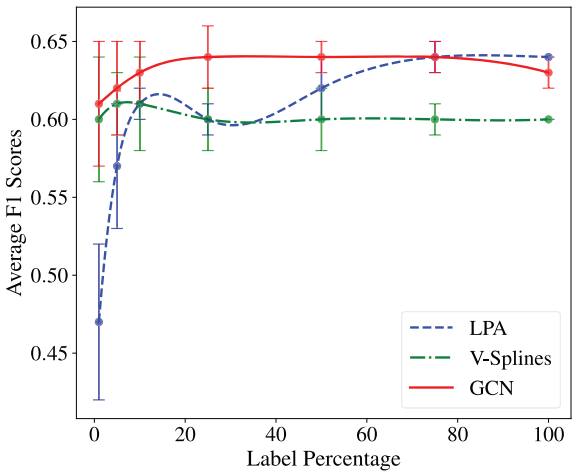

Figure 2. Average F1 scores for an increasing amount of labels and for the three algorithms.

is found for percentage of available labels higher than 75%. GCN outperforms V-Spline for all the tested amounts of labels (5% : $p = .0352$, 10% : $p < .001$, 25% : $p < .001$, 50% : $p < .001$, 75% : $p < .001$, 100% : $p < .001$). The post-hoc analyses for the main effect of the percentages of available labels on the performances of each algorithm reveal that for GCN there is no statistical difference between the average F1 score obtained with the 10% of labels and those obtained with higher amounts of labels (all p values ≥ 0.05). For LPA, there exists a significant difference between the average F1 scores for all percentages. No statistically significant difference is observed for V-Spline. In summary, the analysis on the average F1 scores confirms that an amount of labels as low as 10% is enough for GCN to get the same performances reached with 100% of labels by both GCN and LPA and to get better performances than V-Spline with all the available labels.

About the F1 scores for each class (see Table 3), all the algorithms reach F1 scores higher than 0.6 on the high potency class with any amount of labels, whereas scores are closer or even lower than 0.5 for the low potency class. This is in line with the results reported in [13] suggesting that classifying high potency is easier. Following the results of the analysis of the average F1-scores, we observe specifically that GCN outperforms both LPA and V-Spline for the low-potency class (both $p < .001$), whereas LPA outperforms both GCN and V-Spline for the high-potency class (both $p < .001$). Nevertheless, in an artificial social intelligence scenario, detecting low levels of potency is of particular interest to design proper strategies of intervention to positively impact the team functioning [25] and support humans' effort [11]. Therefore, the significantly better performances of GCN with respect to those of the baselines in the average F1-score and in the F1-score for the low potency class make it more effective than the baselines for giving labels to unlabeled team potency samples based on a small number of labels.

7. Conclusions

In this study, we presented a method to reduce the cost of manual annotations of social interactions. To assess our method we assigned a low vs. high potency label to unlabeled data exploiting only a small portion of manually annotated data. Hence, we trained a GCN on a binary node classification task in a semi-supervised setting. Results show that using a GCN we can reduce the number of labels up to only the 10% to effectively classify social interactions. The benefits of achieving good performance in such a setting could be multiple, such as reducing the time and cost of collecting many manually annotated data. Moreover, this approach decreases the time and energy consumption with respect to training deep architectures on big datasets. This study presents some limitations too. GCN was trained in a transductive framework meaning that the k-NN graph has to be re-constructed and the model re-trained each time new data is added. Adopting such a framework makes the algorithm costly to maintain in a real-world scenario.

References

[1] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

[2] Ron Artstein. *Inter-annotator Agreement*, pages 297–313. Springer Netherlands, 2017.

[3] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[4] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000.

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[6] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. The ami meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction: Second International Workshop*. Springer, 2006.

[7] Alberto Castro-Hernández, Kathleen Swigger, Fatma Cemile Serce, and Victor Lopez. Classification of group potency levels of software development student teams. *Polibits*, (51):55–62, 2015.

[8] Eleonora Ceccaldi, Nale Lehmann-Willenbrock, Erica Volta, Mohamed Chetouani, Gualtiero Volpe, and Giovanna Varni. How unitizing affects annotation of cohesion. In *8th International Conference on Affective Computing and Intelligent Interaction*, pages 1–7. IEEE, 2019.

[9] David Chan. Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of applied psychology*, 83(2):234, 1998.

[10] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[11] Edward Clarkson, Jason A Day, and James D Foley. An educational digital library for human-centered computing. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, 2006.

[12] Paul R Cohen. *Empirical methods for artificial intelligence*, volume 139. MIT press Cambridge, MA, 1995.

[13] Nicola Corbellini, Eleonora Ceccaldi, Giovanna Varni, and Gualtiero Volpe. An exploratory study on group potency classification from non-verbal social behaviours. In *12th International Workshop on Human Behavior Understanding*, 2022.

[14] Ad De Jong, Ko De Ruyter, and Martin Wetzels. Antecedents and consequences of group potency: A study of self-managing service teams. *Management science*, 51(11):1610–1625, 2005.

[15] Roger D Dias, Lauren R Kennedy-Metz, Steven J Yule, Matthew Gombolay, and Marco A Zenati. Assessing team situational awareness in the operating room via computer vision. In *IEEE Conference on Cognitive and Computational Aspects of Situation Management*, 2022.

[16] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

[17] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[18] Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[19] Nuria Gamero, Ana Zornoza, José M Peiró, and Carmen Picazo. Roles of participation and feedback in group potency. *Psychological reports*, 105(1):293–313, 2009.

[20] CB Gibson, A Randel, and PC Earley. Work team efficacy: An assessment of group confidence estimation methods. *Group and Organization Management: An International Journal*, 25(1):67–97, 2000.

[21] J. H. Giraldo, S. Javed, and T. Bouwmans. Graph moving object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2485–2503, 2022.

[22] Donald Glowinski, Maurizio Mancini, Roddy Cowie, Antonio Camurri, Carlo Chiorri, and Cian Doherty. The movements made by performers in a skilled quartet: a distinctive pattern, and the function that it serves. *Frontiers in psychology*, 4:841, 2013.

[23] Stanley M Gully, Kara A Incalcaterra, Aparna Joshi, and J Matthew Beaubien. A meta-analysis of team-efficacy, potency, and performance: interdependence and level of analysis as moderators of observed relationships. *Journal of applied psychology*, 87(5):819, 2002.

[24] Richard A Guzzo, Paul R Yost, Richard J Campbell, and Gregory P Shea. Potency in groups: Articulating a construct. *British journal of social psychology*, 32(1):87–106, 1993.

[25] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. Human-ai complementarity in hybrid intelligence systems: A structured literature review. *PACIS*, 2021.

[26] Hayley Hung and Daniel Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *IEEE Transactions on Multimedia*, 12(6):563–575, 2010.

[27] Simone Kauffeld, Nale Lehmann-Willenbrock, and Annika L. Meinecke. *The Advanced Interaction Analysis for Teams (act4teams) Coding Scheme*, page 422–431. Cambridge University Press, 2018.

[28] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[29] Maria Koutsombogera and Carl Vogel. Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, 2018.

[30] Steve WJ Kozlowski and Georgia T Chao. Unpacking team process dynamics and emergent phenomena: Challenges, conceptual advances, and innovative methods. *American Psychologist*, 73(4):576, 2018.

[31] Cynthia Lee, Catherine H Tinsley, and Philip Bobko. An investigation of the antecedents and consequences of group-level confidence. *Journal of Applied Social Psychology*, 32(8):1628–1652, 2002.

[32] Jin Joo Lee, Brad Knox, Jolie Baumann, Cynthia Breazeal, and David DeSteno. Computationally modeling interpersonal trust. *Frontiers in psychology*, 4:56004, 2013.

[33] Jeffery A LePine, Ronald F Piccolo, Christine L Jackson, John E Mathieu, and Jessica R Saul. A meta-analysis of teamwork processes: tests of a multidimensional model and relationships with team effectiveness criteria. *Personnel psychology*, 61(2):273–307, 2008.

[34] Michael K Lindell and Christina J Brandt. Assessing interrater agreement on the job relevance of a test: A comparison of cvi, t, rwg (j), and r* wg (j) indexes. *Journal of applied psychology*, 84(4):640, 1999.

[35] Hancong Liu, Sirish Shah, and Wei Jiang. On-line outlier detection and data cleaning. *Computers & chemical engineering*, 28(9):1635–1647, 2004.

[36] Lucien Maman, Eleonora Ceccaldi, Nale Lehmann-Willenbrock, Laurence Likforman-Sulem, Mohamed Chetouani, Gualtiero Volpe, and Giovanna Varni. Game-on: A multimodal dataset for cohesion and group analysis. *IEEE Access*, 8:124185–124203, 2020.

[37] Lucien Maman, Laurence Likforman-Sulem, Mohamed Chetouani, and Giovanna Varni. Exploiting the interplay between social and task dimensions of cohesion to predict its dynamics leveraging social sciences. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021.

[38] Michelle A Marks, John E Mathieu, and Stephen J Zaccaro. A temporally based framework and taxonomy of team processes. *Academy of management review*, 26(3):356–376, 2001.

[39] Thomas A O'Neill. An overview of interrater agreement on likert scales for researchers and practitioners. *Frontiers in psychology*, 8:777, 2017.

[40] Isaac Pesenson. Variational splines and paley–wiener spaces on combinatorial graphs. *Constructive Approximation*, 29:1–21, 2009.

[41] Navin Raj Prabhu, Nale Lehmann-Willenbrock, and Timo Gerkmann. End-to-end label uncertainty modeling in speech emotion recognition using bayesian neural networks and label distribution learning. *IEEE Transactions on Affective Computing*, (1):1–14, 2023.

[42] Tammy Rapp, Travis Maynard, Monique Domingo, and Elizabeth Klock. Team emergent states: What has emerged in the literature over 20 years. *Small Group Research*, 52(1):68–102, 2021.

[43] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys*, 54(9):1–40, 2021.

[44] Ronald W Schafer. What is a savitzky-golay filter? *IEEE Signal processing magazine*, 28(4):111–117, 2011.

[45] Alexander D Stajkovic, Dongseop Lee, and Anthony J Nyberg. Collective efficacy, group potency, and group performance: meta-analyses of their relationships, and test of a mediation model. *Journal of applied psychology*, 94(3):814, 2009.

[46] Enzo Tartaglione, Beatrice Biancardi, Maurizio Mancini, and Giovanna Varni. A hitchhiker's guide towards transactive memory system modeling in small group interactions. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021.

[47] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[48] Hendric Voß, Heiko Wersing, and Stefan Kopp. Addressing data scarcity in multimodal user state recognition by combining semi-supervised and supervised learning. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, 2021.

[49] Johannes Wagner, Tobias Baur, Yue Zhang, Michel F Valstar, Björn Schuller, and Elisabeth André. Applying cooperative machine learning to speed up the annotation of social signals in large multi-modal corpora. *arXiv preprint arXiv:1802.02565*, 2018.

[50] Chongyang Wang, Yuan Gao, Chenyou Fan, Junjie Hu, Tin Lun Lam, Nicholas Donald Lane, and Nadia Berthouze. Learn2agree: Fitting with multiple annotators without objective ground truth. In *ICLR 2023 Workshop on Trustworthy Machine Learning for Healthcare*, 2023.

[51] Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421, 1972.

[52] Hayden JR Woodley, Matthew JW McLarnon, and Thomas A O'Neill. The emergence of group potency and its implications for team effectiveness. *Frontiers in psychology*, 10:992, 2019.

[53] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. Sustainable AI: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, 4:795–813, 2022.

[54] Alexander Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics*, 2000.

[55] Yue Zhang, Eduardo Coutinho, Zixing Zhang, Caijiao Quan, and Björn Schuller. Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015.

[56] Zixing Zhang, Eduardo Coutinho, Jun Deng, and Björn Schuller. Cooperative learning and its application to emotion recognition from speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):115–126, 2014.

[57] Shun-Chang Zhong, Yun-Shao Lin, Chun-Min Chang, Yi-Ching Liu, and Chi-Chun Lee. Predicting group perfor-

mances using a personality composite-network architecture during collaborative task. In *INTERSPEECH*, 2019.

[58] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation.