A Graph Neural Network Approach for Choosing Robot Addressees in Group Human-Robot Interactions

Sarah Gillet
KTH, Royal Institute of Technology
Sweden
sgillet@kth.se

Iolanda Leite
KTH, Royal Institute of Technology
Sweden
iolanda@kth.se

Marynel Vázquez
Yale University
United States
marynel.vazquez@yale.edu

Abstract: Interacting in groups is common in our everyday lives. Social robots have been shown to effectively support these group interactions. However, we still lack knowledge on how we could build autonomous socially intelligent robots that could support varying groups with varying needs in realistic interactions. In this paper, we take the first step toward exploring autonomous and adaptive robots for groups by exploring how we can model groups. The goal of the model is to allow for the effective cloning of demonstrated behavior. We propose, to model groups as graphs and to use Graph Neural Networks (GNN) to model the behavior policy. We compare this modeling approach with a sequential neural network with a long feature vector as an input. In a dataset in which teenagers demonstrate how a social robot could support their small group education sessions, we show that GNNs not only outperform the sequential neural networks but also use fewer trainable parameters.

Keywords: Socially Intelligent Robots, Behavioral Cloning, Graph Neural Networks, Groups

1 Introduction

Humans interact in groups in various everyday situations. The interactions between group members in groups can thereby be characterized by several key processes which we refer to as group dynamics. Since positive group dynamics have been found to lead to more motivation, higher-performing teams [1] and generalized trust [2], recent efforts have explored how robots shape interactions among human group members and affect a group's dynamics [3, 4, 5, 6, 7, 8].

In real-world environments, a social robot might need to interact with groups of varying sizes and display complex behaviors to be effective. In addition, one key observation of this work is that group interactions offer the robot multiple possibilities of 'who' to address, for example, with a specific question: 'What do you think?' could be directed to one of the individual group members or the whole group. Prior work has shown promising results in understanding different verbal and non-verbal strategies to shape interaction dynamics. For that, most recent efforts focused on either understanding 'what' the robot should do to support group interactions, e.g. making vulnerable statements [9] or using gaze [10]. Other works used for example the least active participants to decide 'who to address [11, 4, 12]. However, these works are focusing on evaluating rule-based

Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

heuristics rather than how a robot could learn complex social behaviors that allow for autonomy in groups with a variety of needs and characteristics.

In this work, we propose representing groups as graphs when developing socially intelligent robots that need to decide 'who' they should address in groups of potentially varying sizes. Fundamental to our approach is the use of Graph Neural Networks (GNN). The GNN thereby uses the structure of the group interaction inherent to the graph. This way, each group member (nodes) and the interaction between group members (edges) can be modeled explicitly. The robot then uses the GNN to reason with the help of the structure of the interaction to decide 'who' to address for shaping the interaction within the group. We argue that this explicit modeling of the group interaction is key to successfully capturing the dynamics of the group and is important for deciding 'who' should be addressed, for example, to encourage them to tell their opinion.

We explore the use of GNNs for socially intelligent robots by using imitation learning, i.e. behavioral cloning. The used dataset provides interactions in which teenagers demonstrated how they wanted a robot to support their small-group education sessions. Teenagers were interacting in groups of three with the robot controlled by a fourth teenager. The robot acted as a facilitator and was, as such, not part of solving the task within the educational session. The robot's task instead was to support the group interaction to allow for a 'better' experience. It was left open to the teenagers as experts about teenagers' group interactions to decide what 'better' meant to them.

The particular dataset used allows us to compare the modeling of the group as a graph to a standard form of modeling interaction - through one single feature vector. In sum, our main contributions are: (1) proposing to use graphs to model group interactions when developing socially intelligent robot behaviors for shaping interactions in groups (2) demonstrating the effectiveness of modeling groups as graphs as opposed to a single vector in a behavioral cloning task using Graph Neural Networks (3) advancing the knowledge of creating socially intelligent robots for situated group interactions through imitation learning.

2 Related Work

2.1 Robots in Groups

The study of robots and groups has gained importance within the field of HRI [13], including how people perceive robots in groups and how they influence and facilitate group dynamics [3]. In particular, robots have been shown to improve situations of conflict [7, 14] and emotional support [15] and foster the expression of vulnerability [9] or perception of cohesion [6]. Further, prior work has been interested in studying how robots could support the process of inclusion among adults [5], and children [4, 16] or shape participation behavior [17, 11, 12].

2.2 Imitation Learning in HRI

The use of imitation learning (also often referred to as learning from demonstration in robotics [18, 19]) has been demonstrated to learn robot policies for a variety of human-robot interaction scenarios. In particular, a common approach is to learn robot behaviors from expert human demonstrations, such as in kinesthetic teaching of manipulation skills [20, 21]. Closer to our work, imitation learning was used by Jain et al. [22] on a human-human conversation dataset to predict non-verbal behaviors in a conversation, including back-channelling. Similarly to the work in kinesthetic teaching, we use expert demonstrations that were collected by directly controlling the robot instead of a human-human interaction dataset. However, we focus on modeling social interactions and learning a socially intelligent behavior policy.

2.3 Graph Neural Networks and HRI

Given the importance of Graph Neural Networks (GNN) to our approach, we will first provide background information about one form of GNNs: Message-passing Graph Neural Networks (MPNN). At the end of this Section, we provide an overview on prior work that used GNNs.

Message-passing Graph Neural Networks [23] are composed of one or more *Graph Network blocks* (GN blocks). A GN block takes as input a directed graph G and produces an updated graph G'. Let $G = (\mathbf{u}, \mathbf{V}, \mathbf{E})$, where $\mathbf{u}$ is a global attribute (or feature) for the graph, $\mathbf{V} = \{\mathbf{v}_i\}_{i=1:n}$ are attributes of the graph's nodes, and $\mathbf{E} = \{(\mathbf{e}_k, r_k, s_k)\}_{k=1:m}$ corresponds to the edges. Each $\mathbf{e}_k$ in E is an edge attribute with (r_k, s_k) being the corresponding indices of the receiver and sender nodes. Then, a GN block operates in 3 steps, carrying strong *relational inductive biases* via specific architectural assumptions. First, the edge features are updated using an edge update function ϕ^e, $\mathbf{e'}_k = \phi^e(\mathbf{e}_k, \mathbf{v}_{r_k}, \mathbf{v}_{s_k}, \mathbf{u})$. Second, the node features are updated. For example, for node i, $\mathbf{v'}_i = \phi^v(\bar{\mathbf{e}}'_i, \mathbf{v}_i, \mathbf{u})$ with $\bar{\mathbf{e}}'_i = \rho^{e\to v}(\{(\mathbf{e}'_k, r_k, s_k)\}_{r_k=i, k=1:m})$ being aggregate information from all edges that have the node i as receiver. Third, the global feature $\mathbf{u}$ for the graph is updated as $\mathbf{u}' = \phi^u(\bar{\mathbf{e}}', \bar{\mathbf{v}}', \mathbf{u})$ using all the edges, $\bar{\mathbf{e}}' = \rho^{e\to u}(\{(\mathbf{e}'_k, r_k, s_k)\}_{k=1:m})$, and node information, $\bar{\mathbf{v}}' = \rho^{v\to u}(\{\mathbf{v}'_i\}_{i=1:n})$. The update functions $\phi^e(\cdot), \phi^v(\cdot), \phi^u(\cdot)$ and the aggregate functions $\rho^{e\to v}(\cdot), \rho^{e\to u}, \rho^{v\to u}(\cdot)$ are differentiable functions. Importantly, the aggregate functions are often implemented via symmetric mathematical functions, like element-wise averaging, because nodes and edges in a graph typically lack a natural order.

GNNs have been previously used to model groups in HRI settings. For example, to generate poses that are suitable to be perceived as being part of a group [24] or predict group behavior [25]. The closest to our work is a classification task to predict back-channeling behavior in groups [26]. In this work, the authors compare modeling individuals vs the joined group through graphs and GNNs. To the best of our knowledge, there is no prior work that explores cloning non-physical, purely social behaviors in group interactions through the use of GNNs.

3 Approach

In this section, we describe our method which involves representing groups through graphs and using Graph Neural Networks to represent a socially intelligent behavior policy for the robot. The goal of the behavior policy is to decide 'who' should be addressed by the next action, e.g. when asking 'What do you think?' the decision could be at 'whom' to gaze to ensure the question is targeted at the chosen addressee.

3.1 Problem formulation

A social robot aiming to shape group dynamics needs to perceive the group and its interaction dynamics and be able to take an action that chooses whom to address to support the interaction. Therefore, we pose the problem as a sequential decision-making problem. At any time-step t, the robot's environment is captured as a state variable $\mathbf{s}_t$. The robot can choose an action a_t which corresponds to choosing the 'who'.

The robot's goal is to then learn a policy $\pi : \mathbf{s}_t \mapsto a_t$ that indicates which action a to take in a given state s to most accurately clone the demonstrated behavior present in the dataset.

Human groups naturally occur in varying sizes. For instance, small groups of 2 or 3 people might be common in the home [27] while robots might naturally take part in bigger groups interactions in public environments [28] or educational settings [29]. Further, people might dynamically join and leave human-robot groups [30, 31]. Therefore, the decision-making process described above needs to be able to handle different group sizes and generalize to group sizes unseen during training. Note that the given dataset does not allow the exploration of varying group sizes. However, the GNNs can be applied to groups of varying sizes due to the nature of their components.

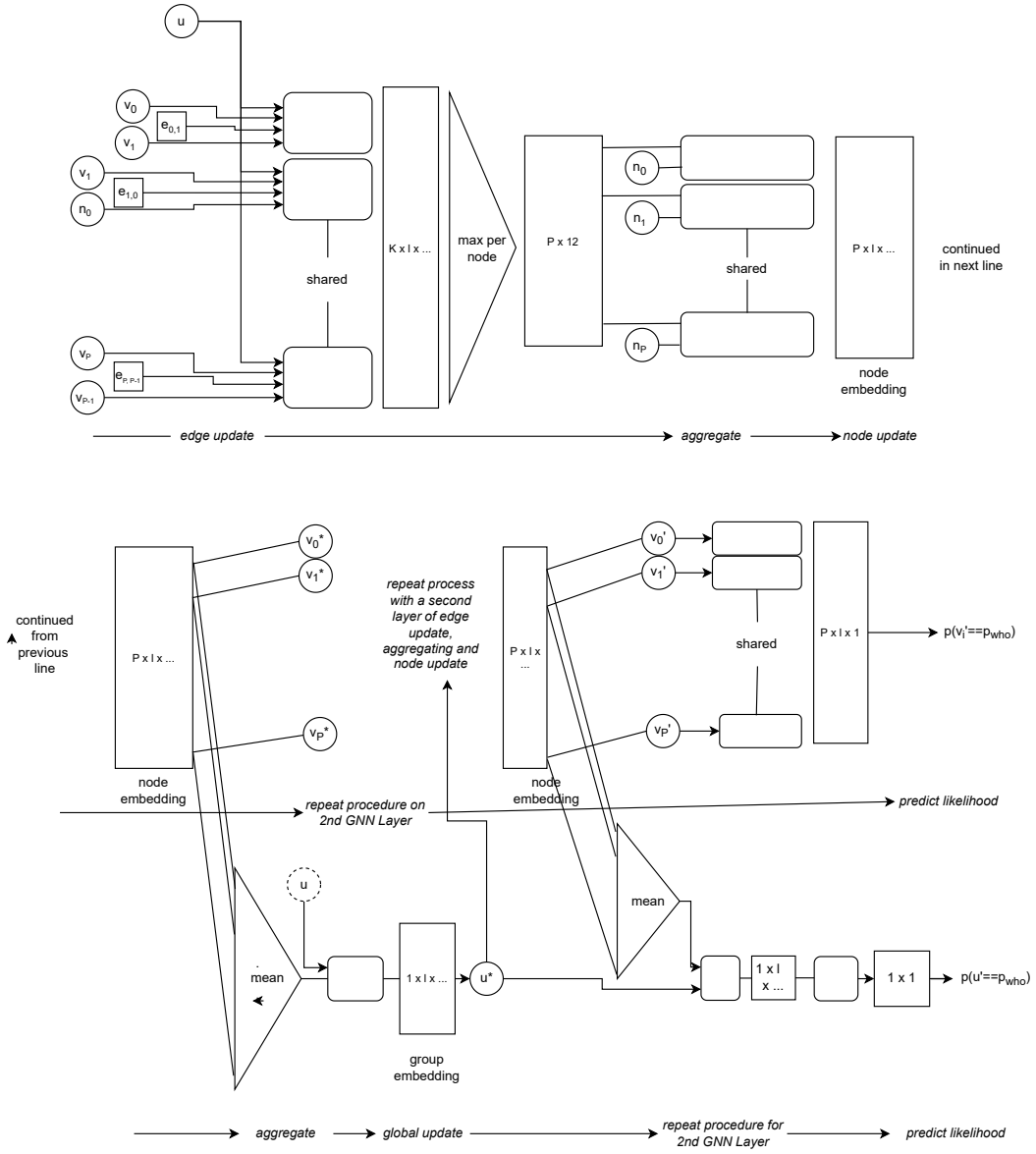


Figure 1: Overview on the network architecture representing the GNN. Round shapes indicate input to the networks. Rounded rectangular shapes represent the neural networks from which the GNN is built. Sharped-edged rectangular shapes represent the outputs of these networks.

3.2 Representations for state and actions

We focus on instances in the dataset in which a 'who' was selected. These decisions on 'who' were, however, not provided continuously but at arbitrary time steps t_d chosen by the expert, the teenager controlling the robot in the group interaction. We model the state s_{t_d} of a group human-robot interaction as a graph $G = (u, V, E)$, where each node $v \in V$ represents one of the interacting human group members. We denote human group members with $p_i \in P$ where p_i represents one human group member. The group member chosen as the addressee is referred to as p_{who}. Edges $e_{ij} \in E$ encode the relation between group members p_i and p_j. General information is captured in the global graph attribute u of G. As the interaction is perceived at 2Hz, the representation at s_{t_d} captures a time series of l time steps of perceived features. Capturing a time series of features allows the behavior policy to reason over the development in a time window of the interaction rather than a single glimpse at one moment in time.

This formulation is applicable to groups of various sizes, i.e., with 2 or more humans.[1]

We represent the actions that the robot takes as the available addresses within the interaction. In addition, the robot can target the whole group.

3.3 Behavioral cloning with GNNs to predict addressees

We train the behavior policy using behavioral cloning [32]. That is, we used supervised learning to map observed states $\mathbf{s}_{t_d}$ to actions a_{t_d} given paired input-target data from the dataset. The dataset will be further discussed in Section 4.1 We represent the behavior policy $\pi(a, s)$ through a GNN. Specifically, we are using Message-Passing GNNs as detailed in Section 2.3.

As illustrated in Figure 1, each node v is passed through two GNN layers resulting in the updated node v'. To predict if person p_i represented through updated node v'_i is the chosen p_{who}, we use an additional function representing the likelihood of p_i being the chosen addressee p_{who}:

$$\phi_{pv}(v') = p(v' == p_{who}) \tag{1}$$

Similarly, we predict based on the updated global graph attributed u' the likelihood of the whole group being the addressee with function $\phi_{pu}(u') == p(u' == p_{who})$.

We choose the highest likelihood among the participants p_i represented through v'_i and the whole group represented through u' as the chosen 'who' as action a_{t_d}.

To accommodate the l time steps that represent each state $\mathbf{s}_{t_d}$, functions $\phi_u, \phi_v, \phi_e, \phi_{pv}, \phi_{pu}$ are represented through a Long Short Term Memory (LSTM) [33].

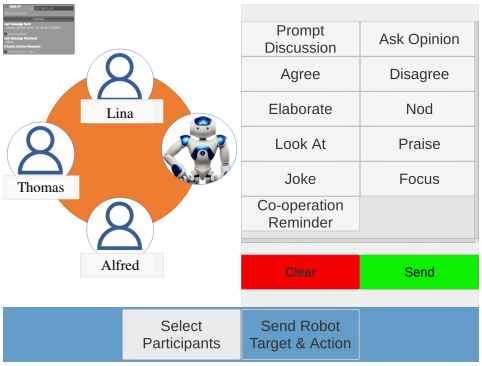
4 Experimental Set-up

4.1 Dataset of group-robot interactions

The goal of the dataset collection was to invite teenagers to demonstrate how a robot could enable 'better' group interactions among a group of teenagers. In this sense, the teenagers were seen as experts that could provide a unique demonstration of robot behaviors. Therefore, the dataset was collected in an interaction among teenagers in which three teenagers worked on a discussion-based task as common in small-group educational settings. The robot acted as a facilitator with the goal of making the group interaction 'better' and was controlled by a fourth teenager. This teenager basically had the task to decide which robot action to take and whom to address with the action to achieve the goal of a 'better' group interaction. A fifth teenager was observing the interaction. The interaction is visualized in Figure 2a. The teenagers demonstrated how the robot should act

[1] Also, our formulation generally assumes that there is a single robot in the interaction – but if there were more, they would be added as nodes to the graph and would be treated as one more interactant by the robot being controlled through $\pi(a, s)$. Evaluating our approach in the latter setup is out of the scope of this paper.

(a) The three teenagers without caps around the table work on a discussion-based task. The Nao robot acts as a facilitator and is controlled by the teenager with cap on the left. The teenager with the cap on the right acts as an observer.



(b) Tablet interface used to control the robot. Teenagers designed the actions in the original study.

Figure 2: Illustration of the data collection set up (a) and user interface to control the robot in (b). Images used with permission of [34].

to support their group interactions by choosing the actions the robot should take on a tablet. The interface used on the tablet is provided in Figure 2b. The dataset was kindly provided by Gillet et al. [34] who conducted the original study and collected the dataset.

Size of the dataset The dataset contains 48 interactions of 15 minutes. During these 48 interactions, 16 teenagers rotated to either work on the group task or control the robot. Within these 48 interactions, 2654 instances demonstrate the selection of an addressee. We use 10% of the dataset for testing and 10% for validation. The majority of actions are directed toward the whole group. In the training set, we up-sample the remaining actions to achieve a more equal distribution of addressees during training.

Actions demonstrated by the teenagers To control the robot, teenagers chose an action type on the right side of the tablet illustrated in Figure 2b, e.g. 'Prompt discussion' or 'Elaborate', and an addressee on the left side of the tablet before clicking the 'Send' button. The whole group was addressed if no group member was chosen as the addressee. In this work, we focus on learning 'who' should be addressed and leave the question of how to clone the action type selection to future work.

4.2 Perception of the state space

The dataset is comprised of audio data captured through lapel microphones worn by each individual interacting with the robot. From the audio data stream, we extracted 37 features for each human group member and three general features describing the group interaction. In particular, our state includes 13-dimensional mel-frequency cepstrum coefficients (MFCC) and 4-dimensional prosody features extracted from individual audio signals. The MFCC features are computed every 25ms with a sliding hamming window of 40ms. In addition, we compute speech intensity through yin-energy and pitch through the fundamental frequency as well as the first derivative of these features. Statistical quantities are applied to feature vectors over time to describe speech over the past second. Specifically, we compute each feature's mean and standard deviation, resulting in a 34-item feature vector. In addition to these low-level audio features, we extracted three high-level audio features capturing: the loudness of the speech in dB, the accumulated relative amount of speech compared to all speech in the group, and whether the human group member is currently speaking.

The interaction between the human group members is captured in edges between all human group members. In this work, no features are captured on the edges. The edges are solely used to connect the individual nodes.

Table 1: Hyperparameter and model variations used to train the GNN. All network architectures used within the GNN were LSTMs. The parameters indicate the output size of the LSTM.

1st GNN Layer		ϕ_u	2nd GNN Layer		l
ϕ_e	ϕ_v		ϕ_e	ϕ_v	
16, 8	8,4	12,8,6	8,4	4,2	20,30

Table 2: Hyperparameter and model variations used to train the sequential neural network. The network architectures used on the first layer were LSTMs. For the second layer, we explored using an LSTM or a dense linear layer.

Layer 1	Layer 2	l
LSTM: 52,26,12	LSTM: 26,13,6, Linear: 26,13,6	10, 20, 30

The global level features are built by a measure of unevenness in speaking amounts, the time since the last action was taken, and the type of action that the teenager had chosen on the tablet in addition to the addressee. The measure of unevenness was originally proposed by [11] and is defined by uneven $= \sum_i (\text{sp}_r^i - \frac{1}{|P|})$, where sp_r^i represents the relative speech amount of participant p_i and $|P|$ the total number of participants.

4.3 Baseline

As a baseline, we explore the modeling of the group in a single feature vector. The features are concatenated in a fixed order, resulting in a 114-item feature vector. A neural network is fitted to the input vector to generate the selected 'who.'

4.4 Metrics for evaluation

We use macro accuracy to ensure that the accuracy for all four possible addresses is considered independent of the number of samples present in the dataset. This metric is specifically important for the test dataset since we only up-sample the training dataset and the test set remains with an imbalance between choosing the individual group members and the whole group as an addressee. In all cases, we used the F1-score representing the harmonic mean between precision and recall to evaluate our models.

5 Experimental Results

We compare the proposed method modeling $\pi : \mathbf{s}_t \mapsto a_t$ through a GNN to a sequential neural network. For both approaches, we performed an extensive grid search to find appropriate model sizes but also hyperparameter settings. Tables 1 and 2 show the explored variations. We used a fixed seed of 42 from the start of development through training of the networks.

The best-performing model according to the macro average of the F1-score was a GNN with the output size of ϕ_e as 16, ϕ_v as 8, ϕ_u as 6, and ϕ_e as 4, and ϕ_v as 4 on the second layer. The considered number of time steps l was 30 steps. The result for the top five models of each model variant are given in Table 3. The five best-performing GNN models outperform all network architectures based on the single input vector. In addition, all models using GNN use fewer trainable parameters, on average the GNNs have only 21.6% compared to the size of the sequential network.

6 Limitations

Behavior cloning is a very limited form of imitation learning. Future work might therefore explore if the benefits of representing groups as graphs and the use of GNN are also effective for other imitation learning methods. Further, the baseline algorithm suffers from using ordered inputs to the model. Potentially, using all permutations of order instead of only one could increase the amount of training

Table 3: Overview on the F1-scores and number of trainable parameters for the five best performing models per modeling approach. The italic marked numbers correspond to the lowest score among all classes.

		F1-score				
per class/model	Parameters	0	1	2	3	Macro avg.
$\phi_e^1 := 16$, $\phi_v^1 := 8$, $\phi_u := 6$,$\phi_e^2 := 4$, $\phi_v^2 := 4$, $l = 30$	10232	*0.379*	0.465	0.455	0.578	**0.467**
$\phi_e^1 := 8$, $\phi_v^1 := 4$, $\phi_u := 6$,$\phi_e^2 := 4$, $\phi_v^2 := 4$, $l = 20$	5272	*0.358*	0.397	0.462	0.571	0.447
$\phi_e^1 := 8$, $\phi_v^1 := 8$, $\phi_u := 12$,$\phi_e^2 := 8$, $\phi_v^2 := 4$, $l = 20$	9384	*0.341*	0.397	0.467	0.636	0.46
$\phi_e^1 := 8$, $\phi_v^1 := 8$, $\phi_u := 8$,$\phi_e^2 := 8$, $\phi_v^2 := 2$, $l = 20$	7584	*0.322*	0.41	0.385	0.654	0.442
$\phi_e^1 := 8$, $\phi_v^1 := 16$, $\phi_u := 8$,$\phi_e^2 := 8$, $\phi_v^2 := 2$, $l = 20$	11392	*0.33*	0.398	0.389	0.611	0.432
LSTM:=52,LSTM=26, $l = 20$	43372	0.328	0.346	0.441	0.647	0.44
LSTM:=52,LSTM=26, $l = 30$	43372	0.335	0.371	*0.308*	0.624	0.409
LSTM:=52,LSTM=26, $l = 10$	43372	*0.279*	0.349	0.428	0.639	0.424
LSTM:=52,Dense=26, $l = 10$	36430	*0.342*	0.389	0.361	0.567	0.415
LSTM:=52,Dense=26, $l = 30$	36430	*0.260*	0.430	0.319	0.586	0.399
Chance level		0.25	0.25	0.25	0.25	0.25

data and, thereby, the outcome for the baseline. However, care must be taken that permutations of the same data point are not spread over train, test, and validation set. Graph Neural Networks are order independent and, therefore, are not subject to the problem of ordering input data.

The trained models achieve a macro-accuracy of a maximum of 0.467. This can still be considered low. To further improve accuracy, a variety of steps could be taken: 1) The dataset only contains audio data. However, prior work has shown that combining audio and video data can benefit learning social behaviors. Future work should explore if the combination of audio and video can improve the accuracy of a similar dataset. 2) Even though teenagers interacted with the robot 48 times for 15 minutes, the dataset can still be considered small. Future work should find methods to collect more data more efficiently or to further improve the effective use of the available data. 3) The teenagers can be considered experts for the chosen task of creating a facilitator for 'better' group interactions. However, different teenagers might have interpreted the task differently or changed their approach to the task over time. Therefore, the dataset might not be fully consistent which could explain the reached accuracy. Future work could consider providing training or involving intense group discussions so that the groups could reach a consensus about what they want their robot to do.

7 Conclusion

In this paper, we explored the use of Graph Neural Networks to model the behavior of socially intelligent robots that need to decide 'who' they are addressing when supporting group interactions. We build upon a dataset in which teenagers demonstrated social behaviors that can support small-group education settings and use imitation learning to clone the behavior demonstrated. We compare the modeling of the behavior policy through a Graph Neural Network with a neural network that is based on one long feature vector instead of a graph representation as input. Our results show that the GNN-based approach outperforms the standard neural network and consists of fewer trainable parameters. Given the sparsity of demonstration of social robot behaviors in human-robot interactions, GNNs seem suitable not only in terms of their performance but also their network size to model social robot behavior that supports human-human interactions.

Acknowledgments

This work was partially funded by the Swedish Research Council (no. 2017-05189), the Swedish Foundation for Strategic Research (FFL18-0199), the Jacobs Foundation (no. 2017 1261 06), the Vinnova Competence Center for Trustworthy Edge Computing Systems and Applications at KTH, and the National Science Foundation (IIS-2143109). We would also like to thank Katie Winkle and Giulia Belgiovine who were part of the original study and Maria Teresa Parreira for helping with the initial analysis and transformation of the dataset. Lastly, we would like to thank Signeuls Stiftelse for funding Sarah Gillet's travel to CORL 2022.

References

[1] B. A. Nijstad. *Group performance*. Psychology Press, 2009.

[2] W. van den Bos, E. A. Crone, R. Meuwese, and B. Güroğlu. Social network cohesion in school classes promotes prosocial behavior. *PLOS ONE*, 13(4):e0194656, Apr. 2018. ISSN 1932-6203. doi:10.1371/journal.pone.0194656.

[3] S. Sebo, B. Stoll, B. Scassellati, and M. F. Jung. Robots in Groups and Teams: A Literature Review. *Proc. ACM Hum.-Comput*, 4(October):37, 2020. URL https://doi.org/10.1145/3415247.

[4] S. Gillet, W. van den Bos, and I. Leite. A social robot mediator to foster collaboration and inclusion among children. In *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, July 2020. doi:10.15607/RSS.2020.XVI.103.

[5] S. Strohkorb Sebo, L. L. Dong, N. Chang, and B. Scassellati. Strategies for the Inclusion of Human Members within Human-Robot Teams. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 309–317, New York, NY, USA, 3 2020. ACM. ISBN 9781450367462. doi:10.1145/3319502.3374808. URL https://dl.acm.org/doi/10.1145/3319502.3374808.

[6] S. Strohkorb, E. Fukuto, N. Warren, C. Taylor, B. Berry, and B. Scassellati. Improving human-human collaboration between children with a social robot. *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, pages 551–556, 2016. ISSN 1664-302X. doi:10.1109/ROMAN.2016.7745172.

[7] M. F. Jung, N. Martelaro, and P. J. Hinds. Using Robots to Moderate Team Conflict: The Case of Repairing Violations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, page 229–236, Portland, Oregon, USA, 2015. Association for Computing Machinery. ISBN 9781450328838. doi:10.1145/2696454.2696460. URL http://dl.acm.org/citation.cfm?doid=2701973.2702094.

[8] E. S. Short, K. Swift-Spong, H. Shim, K. M. Wisniewski, D. K. Zak, S. Wu, E. Zelinski, and M. J. Mataric. Understanding social interactions with socially assistive robotics in intergenerational family groups. *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 236–241, 2017. ISSN 1944-9445. doi:10.1109/ROMAN.2017.8172308. URL http://ieeexplore.ieee.org/document/8172308/.

[9] S. Strohkorb Sebo, M. Traeger, M. F. Jung, and B. Scassellati. The Ripple Effects of Vulnerability: The Effects of a Robot's Vulnerable Behavior on Trust in Human-Robot Teams. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*, (February):178–186, 2018. ISSN 21672148. doi:10.1145/3171221.3171275. URL http://dl.acm.org/citation.cfm?doid=3171221.3171275.

[10] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 61–68. ACM, 2009.

[11] H. Tennent, S. Shen, and M. Jung. Micbot: a peripheral robotic object to shape conversational dynamics and team performance. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 133–142. IEEE, 2019.

[12] S. Gillet, R. Cumbal, A. Pereira, J. Lopes, O. Engwall, and I. Leite. Robot Gaze Can Mediate Participation Imbalance in Groups with Different Skill Levels. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, 3 2021. ACM. ISBN 9781450382892. doi:10.1145/3434073.3444670.

[13] E. Schneiders, E. Cheon, J. Kjeldskov, M. Rehm, and M. B. Skov. Non-dyadic interaction: A literature review of 15 years of human-robot interaction conference publications. *J. Hum.-Robot Interact.*, 11(2), 2022. doi:10.1145/3488242.

[14] S. Shen, P. Slovak, and M. F. Jung. "Stop. I See a Conflict Happening.". In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 69–77, New York, NY, USA, 2 2018. ACM. ISBN 9781450349536. doi:10.1145/3171221.3171248. URL https://dl.acm.org/doi/10.1145/3171221.3171248.

[15] H. Erel, D. Trayman, C. Levy, A. Manor, M. Mikulincer, and O. Zuckerman. Enhancing emotional support: The effect of a robotic object on human–human support quality. *International Journal of Social Robotics*, pages 1–20, 2021.

[16] S. Tuncer, S. Gillet, and I. Leite. Robot-mediated inclusive processes in groups of children: From gaze aversion to mutual smiling gaze. *Frontiers in Robotics and AI*, 9, 2022. ISSN 2296-9144. doi:10.3389/frobt.2022.729146.

[17] V. Charisi, L. Merino, M. Escobar, F. Caballero, R. Gomez, and E. Gómez. The effects of robot cognitive reliability and social positioning on child-robot team dynamics. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

[18] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[19] A. Billard and D. Grollman. Robot learning by demonstration. *Scholarpedia*, 8(12):3824, 2013.

[20] B. Akgun, M. Cakmak, J. W. Yoo, and A. L. Thomaz. Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 391–398, 2012.

[21] C. L. Mueller and B. Hayes. Safe and robust robot learning from demonstration through conceptual constraints. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 588–590, 2020.

[22] V. Jain, M. Leekha, R. R. Shah, and J. Shukla. Exploring semi-supervised learning for predicting listener backchannels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2021.

[23] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[24] M. Vázquez, A. Lew, E. Gorevoy, and J. Connolly. Pose generation for social robots in conversational group formations. *Frontiers in Robotics and AI*, page 341, 2022.

[25] F. Yang, W. Yin, T. Inamura, M. Björkman, and C. Peters. Group behavior recognition using attention-and graph-based neural networks. In *ECAI 2020*, pages 1626–1633. IOS Press, 2020.

[26] G. Sharma, K. Stefanov, A. Dhall, and J. Cai. Graph-based group modelling for backchannel detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7190–7194, 2022.

[27] B. Scassellati, L. Boccanfuso, C.-M. Huang, M. Mademtzi, M. Qin, N. Salomons, P. Ventola, and F. Shic. Improving social skills in children with asd using a long-term, in-home social robot. *Science Robotics*, 3(21):eaat7544, 2018.

[28] L. Moshkina, S. Trickett, and J. G. Trafton. Social engagement in public places: a tale of one robot. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 382–389, 2014.

[29] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human–Computer Interaction*, 19(1-2):61–84, 2004.

[30] M. Vázquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson. Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 42–52. IEEE, 2017.

[31] D. Bohus and E. Horvitz. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*, pages 2–9, 2014.

[32] D. Pomerleau. An autonomous land vehicle in a neural network. *Advances in Neural Information Processing Systems; Morgan Kaufmann Publishers Inc.: Burlington, MA, USA*, 1998.

[33] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[34] S. Gillet, K. Winkle, G. Belgiovine, and I. Leite. Ice-breakers, turn-takers and fun-makers: Exploring robots for groups with teenagers. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1474–1481, 2022. doi:10.1109/RO-MAN53752.2022.9900644.