Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Identifying Bullies with a Computer Game

Juan F. Mancilla-Caceres, Wen Pu, Eyal Amir
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801
{mancill1,wenpu1,eyal}@illinois.edu

Dorothy Espelage
Department of Educational Psychology
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801
espelage@illinois.edu

Abstract

Current computer involvement in adolescent social networks (youth between the ages of 11 and 17) provides new opportunities to study group dynamics, interactions amongst peers, and individual preferences. Nevertheless, most of the research in this area focuses on efficiently retrieving information that is explicit in large social networks (e.g., properties of the graph structure), but not on how to use the dynamics of the virtual social network to discover latent characteristics of the real-world social network. In this paper, we present the analysis of a game designed to take advantage of the familiarity of adolescents with online social networks, and describe how the data generated by the game can be used to identify bullies in 5th grade classrooms. We present a probabilistic model of the game and using the in-game interactions of the players (i.e., content of chat messages) infer their social role within their classroom (either a *bully* or *non-bully*). The evaluation of our model is done by using previously collected data from psychological surveys on the same 5th grade population and by comparing the performance of the new model with off-the-shelf classifiers.

1 Introduction

The behaviors of youth friendship networks are usually studied through long surveys and questionnaires that are inefficient to administer to large school populations, and that may impose a considerable fatigue on the participants. In the context of bullying, research focuses on identifying both individual and contextual factors that contribute to involvement as a *bully* or *non-bully* (i.e., victim, and/or bystander) among children and adolescent groups. Consistently, peer group membership has emerged as a central factor that determines whether an adolescent becomes involved in risky behaviors (Bronfenbrenner 1977; Richard and Jessor 1992) such as bullying. To protect adolescents who are surrounded by risky behaviors, efficient ways of analyzing adolescents' relationship networks are needed.

In this paper, we present the design of a game built to observe and collect data about social networks within grade-school classrooms. Also, using the data of the game, we create a model using bayesian networks (Pearl 1988) where the interaction among players (in this case, the content of chat messages) is determined by the role that each pair of interacting players has in the classroom's social network. We found that the roles of the participants in the classroom are observable through the interactions within the game and that the data obtained through them provides enough information to identify potential bullies in the classroom.

The game is intended to emulate the circumstances of natural interactions amongst participants. Its features include the presence of limited resources (i.e., points and coins), a collaborative task, and a competitive task. By restricting the channels of communication to text messages, we have a non-intrusive way to monitor and analyze the interactions of participants with their peers. From this, we are interested in capturing cues about the role of the players in the classroom that is not explicit in their game behavior, i.e., we are not interested in recognizing a specific type of text message but to be able to infer, by observing the participants' interactions, whether their classroom role is *bully* or *non-bully*. The fact that the players of the game belong to the same classroom, and all players identities are known, helps ensuring that players will not change their real-world role, as there still exists a sense of accountability. Currently, the game provides the following information: teammates nominations, text messages amongst the players (which can be either public or private), and point transactions.

In a previous study (Mancilla-Caceres et al. 2012), we showed that bullies and non-bullies play the game quantitatively different, in terms of the amount of messages sent and received through private channels, the content of those messages, and the nominations made at the beginning of the

game. In total, ninety seven 5th graders from six different classrooms and two different Midwestern middle-schools were surveyed using state-of-the-art psychological questionnaires and were labeled as either *bullies* or *non-bullies* by a field expert.

2 Game Design

Traditionally, surveys are used to collect information about peer relations in classrooms. The administration of these surveys is usually costly and burdensome. Instead, we designed an online multiplayer trivia game (played by 5th graders belonging to the same classroom) used to obtain meaningful observations about peer relations. The game is played as follows:

- **Stage 1.** At the beginning of the game, each player nominates other players whom they would (or would not) like to have as members of their team. This information is interpreted as a sign of possible friendship but it is not used to create the teams, which are previously created using the data from the survey (ensuring that in every team there is at least one player classified as a bully).
- **Stage 2.** After the nomination stage, players on the same team collaborate to answer a set of 5 trivia questions about general knowledge by choosing one of four possible answers. The team must ensure that all its members submit the same (correct) answer or no player receives their reward (in the form of coins).
- **Stage 3.** After the collaborative stage, the game turns adversarial. The team must now answer another set of 5 trivia questions, but in this case all members must select a different answer with the caveat that one of the answers is marked as wrong. Unless every member submits a different answer, and one member selects the wrong one, nobody gets the reward (the player who selects the wrong answer gets nothing).

At the end of the game, the player with the highest score receives a prize (in form of a gift card), encouraging players to oppose their fellow teammates by coercing them to pick the wrong answer (unlike in the collaborative stage, in which the team is encouraged to work together to maximize their individual reward). Also, during all stages of the game, players are allowed to make coin transactions and to peek at the answer at the cost of some of their coins. Figure 1 shows a screenshot of the game interface for the nomination stage, and one of the trivia stages.

These rules ensure that the players in the team communicate through the chat interface, and either collaborate or compete (depending on the stage) in order to guarantee the best possible individual output. During the collaborative stage, it is in the best interest of each player to share their knowledge about the question and not to let other players answer incorrectly or peek at the answer (in order to retain points).

During the adversarial stage, the wrong answer is marked explicitly (written in bold letters) in order to make it obvious and to encourage players to discuss who will pick such answer. It is in the best interest of each player not to pick the wrong answer, but also to ensure that someone else in their team picks it. Given the restricted channel of communication, this can only be achieved by negotiating (either aggressively or not) through the chat channel. These coercive strategies may also be used to obtain coins from other players.

As previously mentioned, the winner of the game is the player with the largest amount of coins. The team with the largest cumulative score (i.e., the sum of all the individual rewards) also gets awarded a team prize.

Every time the game is played, the following information is obtained:

1. Nominations: Users' team preferences (friends/rivals nominations and the order in which they are selected).

2. Interactions: Raw messages from users (all chat messages both in public and private channels along with the time at which the message was sent), and Points transactions, e.g., transfer and forfeiture of points (in exchange for information). The content of each message was analyzed by 20 independent raters (each message evaluated by two raters) and classified into 2 categories: Prosocial/Coercive (whether the purpose of the message is to coerce the recipient into doing something not optimal for themselves), and Positive/Negative Affect (whether the sentiment expressed by the sender is either positive or negative).

Survey

We also collected information with a survey to measure aggression and delinquency (Espelage and Holt 2001; Espelage, Mebane, and Adams 2004), which includes different scales to assess various types of bullying behavior (e.g., teasing other students), fighting behavior (e.g., being in a physical fight), participants' attitudes and perceived attitudes among friends towards bullying and the extent to which they would assist a victim, and the extent to which students agree that they are leaders among their peers, make decisions for friends, and force others to do what they ask. Past research demonstrated that the scales converges with peer and self-nomination reports of bullying (Espelage and Holt 2001).

All participants were surveyed with the previously described questionnaires, and an expert psychologist analyzed the surveys and generated labels for each participant as either *bully*, or *non-bully*. The survey data was used to generate the labels that will later be used to train the model and to evaluate the results yielded by the computer game. The ultimate goal is to develop appropriate psychometric measures solely through the data provided by the computer game (i.e., chat messages, in-game transactions, etc.).

Our main hypotheses are that those students who are labeled as *bullies*, will have a need for control and dominance, will engage in coercive tactics directed toward non-friends, and will solicit support for these tactics from friends within and outside their team. These will be tested by observing the emotional tone (or content) of the messages sent in all pairwise interactions of players in the game.

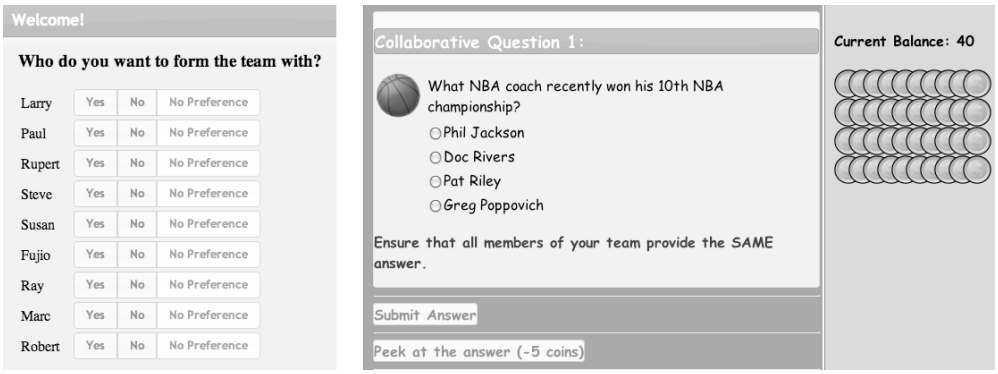

Figure 1: Screenshot of the game. Left image shows the interface to nominate desired teammates. Right image shows a sample question and interface to select an answer.

3 Modeling the Game

Some assumptions are necessary in order to properly model the game. First, the chat messages in the game are unrestricted and written in natural language, which means that the messages must be classified into a small set in order to be able to create appropriate probability distributions. Currently, we are avoiding the need for advanced NLP techniques by having the messages classified by a team of 20 raters (each message was classified by at least two raters) into 2 binary categories: Prosocial/Coercive messages, and Positive/Negative messages.

We considered all pairwise interactions among players independently, i.e., for each pair of players that interacted during gameplay we gathered: the amount of messages sent by one of the players in the pair (the *sender*) to the other (the *recipient*), the nomination given by the *sender* to the *recipient* (as either positive, negative, or don't care), and whether or not they belonged to the same team.

Figure 2 shows a fragment of the proposed bayesian network for one of the classrooms. Notice that each player may interact through a private channel with any of the students of his/her class. The messages sent through this channel may either constitute a prosocial message (e.g., helpful, agreeable, polite, etc.) or coercive (e.g., rude, aggressive, etc.), and may express positive affect (e.g., happy, humorous, etc.) or negative affect (e.g., bored, sad, etc.). Our model assumes that the roles of the *sender* and the *recipient* determine whether the message is either prosocial or coercive, and expresses positive or negative affect.

The roles of the players also determine whether one nominated the other to be on his team (presumably, a victim would not nominate his bully to be in his team, and bullies might nominate each other). The fact that two players are on the same team may also help in understanding the observed interactions. For example, a bully and a non-bully on the same team may cooperate more (have more prosocial messages than coercive) in order to win the game, whereas a bully and a victim on different teams may show a less friendly interaction. In summary, our model considers four observable features for each pair of interacting players: **prosocial/coercive** messages, **positive/negative** messages,

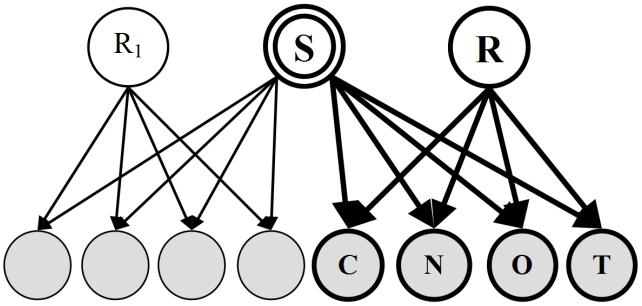

Figure 2: Example of the bayesian network used to model the game's interactions. For each player S, we will consider all the pairwise interactions with the rest of the classroom. For every pair of players, the *sender* S and the *recipient* R, we observe whether the interaction was Prosocial or Coercive (C), Positive or Negative (N), whether S nominated R (O) and whether they belong to the same team (T).

nominations, and **team** membership. In this model, we do not consider points transactions between players.

This model implies that, to obtain the probability of a player being a bully, we need to observe the interactions of the specific participant with all other participants. We build the training set as follows:

1. For each pair of participants, we identify one as the *sender* of messages, and the other as the *recipient* (this implies that the training set also contains a pair where the former *recipient* is now the *sender*, and viceversa).
2. We compare the number of prosocial and coercive messages sent by the *sender* and create a binary variable C which takes the value of 1 if the number of coercive messages is greater of equal to the number of prosocial messages sent, and the value of 0 if otherwise.
3. In a similar way, we create a binary variable N that takes the value of 1 whenever the number of negative messages sent is greater or equal to the number of positive messages.

S	R	C	N	O	T
non-bully	*bully*	1	1	1	0
non-bully	*non-bully*	0	1	1	0
bully	*non-bully*	0	1	-1	0
bully	*bully*	0	0	0	1

Table 1: Examples of training data. Each row represents a training point defined by the interaction of a *sender* (S) and a *recipient* (R) that was either prosocial or coercive (C), positive or negative (N), with a specific nomination (O), and either belonging to the same or to a different team (T).

4. The other two variables are O, if the *sender* nominated positively, negatively, or did not care about having the *recipient* on the same team. And T if the *sender* and the *recipient* belong to the same team or not.

There are five binary variables: *sender* which can be either bully or non-bully, *recipient* which can also be either bully or non-bully, C, N, and T; and a variable with three possible values O for a total of 96 possible values. Given the size of the complete social network, determined by the number of pairwise interactions (a total of 597) and the possible values of the variables, it is possible to find the exact probability of any given *sender* to be a bully given the observation of the four features (and marginalizing over the *recipient* whose value is unknown). Table 1 shows some examples of how a training point looks.

Handling Imbalance in the Data

One of the challenges to address is the intrinsic imbalance in the data; there tends to be a larger number of non-bullies than bullies in every classroom. In our dataset, the average number of bullies per class is 2, whereas the average size of a classroom is 15 students. That gives us a prior probability of 0.12 of being a bully and, across all classrooms, the entropy of the dataset is 2.474. This is a challenge because most off-the-shelf classifiers have an implicit decision threshold that assumes that both positive and negative classes are balanced, i.e., there is a 0.5 probability of a random example being a member of the positive class (in our case, of being a bully). This implies in our dataset that any observed interaction has a greater probability of coming from the interaction of two non-bullies (the majority class) than from any other possible combination (bully- non-bully, or bully-bully). This is exacerbated by the fact that interactions between bullies and non-bullies are not explicit, i.e., the way bullies interact with non-bullies is not necessarily salient. For example, finding a bully is not as easy as finding the kid who says (or writes) more swear words, or who explicitly asks for money from another student. Therefore, interactions between the two most common roles (non-bullies) are more likely to be observed.

The data mining literature provides several ways to deal with this phenomenon (sometimes referred to as novelty detection). Among those suggestions are oversampling the minority class (Noto, Saier, and Elkan 2008), changing the decision threshold whenever possible (Maloof 2003) or redefining the negative and positive class to overcome the imbalance (Elkan and Noto 2008). In this work, we opted for changing the decision threshold due to the intuition that, in our case, it is appropriate to identify as *bully* those participants who have the highest probability of being a bully, regardless of the actual absolute value of the probability.

4 Experiments and Evaluation

Even if a particular player was labeled as a bully using the survey, it does not imply that he will bully all the classmates with whom he/she interacts. For example, the fact that a bully may be coercive towards his victim, does not rule out the possibility of himself behaving nice and cooperative with the rest of people he interacts with. Therefore, we are interested in calculating, for each training point (i.e., each pair of *sender-recipient* variables) the probability P_i^S of the *sender* S being a bully given his interaction with *recipient* i and the observed values of the rest of the variables.

$$P_i^S = P\left(S|C,N,O,T\right) = \frac{\sum\limits_r P\left(S,R_i=r,C,N,O.T\right)}{\sum\limits_s\sum\limits_r P\left(S,R,C,N,O,T\right)} \tag{1}$$

This training procedure will generate n different probabilities for one specific *sender*, where n is the number of pairwise interactions of the *sender*. Because our survey data does not include information about pairwise interactions, we need to consolidate the n probabilities and obtain only one per player. It is our intuition that non-bullies will have a low probability of being a bully across all the interactions they have, whereas a bully will have at least one interaction with high probability. Therefore, the probability of a *sender* S being a bully ($\bar{P}^S$) is the average over all the probabilities across all the *sender*'s interactions (See equation 2). It is expected that non-bullies will have a lower $\bar{P}^S$ than bullies.

$$\bar{P}^S = \frac{1}{n}\sum_i P_i^S \tag{2}$$

If we imagine the real-world scenario of using our game as a tool to identify bullies in a previously unseen classroom, we can see that the best method for evaluation is *leave-one-classroom-out*, i.e., train on five of the available classrooms and evaluate on the sixth one. This is done using as test one classroom at a time, in a cross-fold validation fashion.

After training the bayesian network and averaging all the probabilities for each *sender*, we look for the optimal decision threshold *thr* to label a particular player as a bully. This was done by looking for the threshold that, on the training set, gave the best possible combination of accuracy (*acc*), recall (*rec*), and precision (*pre*).

$$thr = \underset{thr\in(0,1)}{\text{argmax}}\ \left(acc\left(thr\right)\times rec\left(thr\right)\times pre\left(thr\right)\right) \tag{3}$$

where the measures of performance are defined as follows:

$$acc(thr) = \frac{tp_{thr} + tn_{thr}}{tp_{thr} + fn_{thr} + tn_{thr} + fp_{thr}} \quad (4)$$

$$rec(thr) = \frac{tp_{thr}}{tp_{thr} + fn_{thr}} \quad (5)$$

$$pre(thr) = \frac{tp_{thr}}{tp_{thr} + fp_{thr}} \quad (6)$$

where tp_{thr}, fn_{thr}, tn_{thr}, and fp_{thr} stand for true positive, false negatives, true negatives, and false positives (while using *thr* as threshold), respectively. To obtain these values we need to count how many *senders* have a probability $\bar{P}^S$ higher (or lower) than *thr* and were labeled using the surveys as either bullies or non-bullies.

That is, accuracy is the number of correct label assignments for both the positive and the negative class (*bully*, *non-bully*, respectively). Recall is the number of correctly identified bullies, from all the available bullies. And, precision is the the number of correct label assignments from all the assignments done. All three measures of performance were used to keep in mind all possible scenarios in which our method might be used, and to overcome the shortcomings of each of these measures.

Accuracy, although probably the most common measure of performance, might be unreliable in cases where the data is imbalanced (as is ours) because a classifier might optimize this value by simply choosing to label all instances with the label of the majority, in our case, this would automatically obtain an accuracy of 0.88 which is the ratio of non-bullies in the dataset. Recall is also a very common measurement of performance, the problem in our case is that, because we are choosing the decision threshold during training, the optimization could always select the lowest possible threshold and label all players as bullies, which would ensure that all available bullies are correctly labeled, even though it is generating the maximum number of false positives. On the other hand, optimizing precision may lead to the problem of minimizing the number of participants labeled as bullies, in order to minimize the chance of having false positives (i.e., we could potentially have an unlimited number of false negatives). Because it is desirable that our system is capable of identifying all possible bullies in a classroom, and not only the ones who are *clearly* bullies, the best decision threshold *thr*, is the one who maximizes all of the three performance measures.

Comparison with Other Classifiers

Once a threshold is found during training, the test consists on obtaining the probability of a specific *sender* being a bully for each interacting pair in the test set, averaging the probabilities for each participant to obtain $\bar{P}^S$, and comparing this probability with the threshold. This allows us to obtain, for the current classroom, the values of accuracy, recall, and precision.

This training and testing procedure was performed using the model defined by the bayesian network described above and five other off-the-shelf classifiers. These classifiers were ν-SVM (Schölkopf et al. 2000) (as implemented by LIBSVM (Chang and Lin 2011) using a radial kernel), One-Class SVM (Schlkopf et al. 2001) (also, as implemented by LIBSVM), Logistic Regression, Random Trees and Naive Bayes Classifier (the last three as implemented on the WEKA Data Mining platform (Hall et al. 2009) with standard parameters). The results of these experiments are shown in Figure 3. It shows the average performance of each of the methods, and the standard error as error bars.

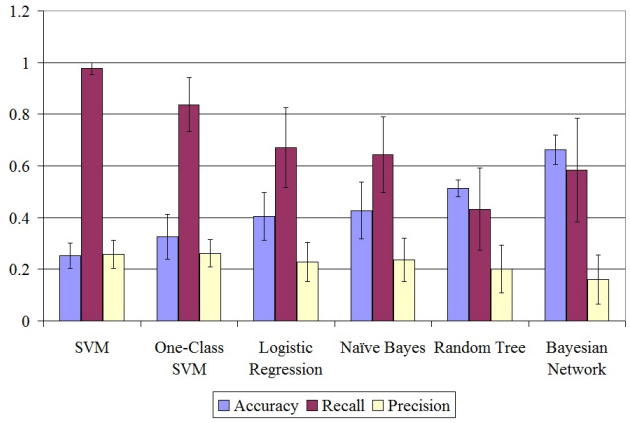

Figure 3: Comparison between off-the-shelf classifiers and our bayesian network, which accomplishes a good trade-off between accuracy and recall, and having significantly higher accuracy than all other methods (except Naive Bayes).

In order to compare the performance of each of the different classifiers (on the three performance measures), a series of t-tests were performed to look for significant differences. The first thing to notice is that all methods perform similarly with respect to precision (i.e., there was no statistically significant difference when comparing the precision between any two methods), and this value is low. This means that all methods produce a relatively large number of false positives in comparison to the number of true positives found. Improving on this value will be the focus of future work (i.e., finding a model or classifier which is able to better distinguish amongst non-bullies and bullies given the data from the game). Therefore, we need to rely only on accuracy and recall to decide which classifier performs best. The method with the highest recall is SVM (significantly better recall than Random Trees), and, as suspected, the cost of having an almost perfect recall is that of having a very low accuracy (it is actually the classifier with the smallest accuracy across all methods, i.e., SVM caused the threshold to be very low and therefore erroneously claiming that most, if not all, of the participants are bullies). In general, as methods improve on accuracy, the recall drops, except when using the bayesian network described in this paper.

Statistically speaking, the bayesian network has the highest accuracy, which is significantly larger than the accuracy of all other methods, except that of Naive Bayes. In the case of recall, none of the methods (not even SVM) generates a statistically significantly larger recall than the bayesian net-

work model. This is due to the large variance of the recall on this method. (Actually, the only statistically significant difference in recall is given by SVM over Random Trees and One-Class SVM over Random Trees).

Given these results, the only two methods that did not performed statistically worse than any other were the bayesian network and the naive bayes approach. Interestingly, they both seem to complement each other in the sense that naive bayes emphasizes recall over accuracy, whereas the bayesian network approach slightly emphasizes accuracy over recall, neither of them being significantly different from the other. Another thing to notice in this comparison is that, although the average recall obtained with the bayesian network is high, the variance is also large, which means that it may be unreliable (i.e., for some classrooms, the recall might be very bad), whereas the naive bayes tends to be more consistent across classrooms.

Also, in the real life scenario where the game can be used by teachers to identify students with a high probability of being bullies, recall might be more beneficial than accuracy in the sense that it would not be too costly to pay special attention to some false positives (which might be easily identifiable by the teacher as non-bullies).

Another important issue to address during this evaluation is the relatively low absolute values of both accuracy and recall, which in the case of our method (the bayesian network) are only slightly larger than 0.5. The most likely explanation for this is that either the model or the game is unable to effectively capture and identify the interaction patterns of all bullies in the game. Most likely, we are only capturing the patterns in the behavior of half the bullies. Although alarming at first, we have come to realize that psychological research (Seigne et al. 2007; Monks, Smith, and Swettenham 2005; Naubuzoka 2009) on bullying has found that there are two main types of bullies: those which have high executive functions (i.e., high capacity of negotiating such that they manage to get their way with both victims and teachers) and those with low executive functions (i.e., those with low capability of negotiating, get frustrated fast and disengage from socially accepted interactions). It is clear that the game would more easily detect those with high executive functions who manage to use the chat interface to manipulate their way into winning the game, whereas those with low executive functions will simply disengage from the game and stop contributing messages in the chat interface, making it extremely difficult for our game and model to identify them.

5 Related Work

With respect to methods of inference in social networks, in (Groh and Hauffa 2011) the authors showed a method for using the content of messages such as the emotional intensity, and other NLP-based sentiment classification methods to analyze email messages and to characterize social relations. (Paradesi and Shih 2011) showed the possibility of using data publicly available from different online social networks to make inferences about health habits using keyword co-occurrences. These two examples of recent research show the interest in the use of social networks to recognize important behavior in the real world (i.e., outside the social network and not explicit in the network) using the content of the interactions of the users in the social network. In our case, the online social network reflects the real world social structure because all the players belong to the same classroom, which allows us to infer real world behaviors. On the other hand, this might not be possible to do in current publicly available online social networks, because there is no evidence that classroom roles are reflected in such networks.

Regarding the use of automatic methods to detect dangerous behavior in online communities, previous research has focused on cyber-bullying and not on physical bullying as we do. For example, (Lieberman, Dinakar, and Jones 2011) focuses exclusively on observable behavior (e.g., finding insulting or racist messages on status updates to online social networks), whereas our approach is aimed at finding the latent behavior associated with children's roles within the classroom, which may or may not be explicit in the game.

Game-based methods for data collection have been successfully used in the past to generate and collect data from social networks. Some examples are Collabio (Bernstein et al. 2009), designed to generate social tags amongst friends in Facebook, and the Turing Game (Mancilla-Caceres and Amir 2011), which encourages Facebook users to evaluate commonsense knowledge through a set of carefully designed rules and stages. Also, previous studies have shown that it is possible to observe real world behavior (e.g., bystander effect) in virtual game environments (Kozlov and Johansen 2010). The game described in this paper differs from previous games in the sense that the desired outcome of the game is not the specific actions the players take on the game (i.e., the classification of the image or text by the player) but the interactions amongst the players. This is a novel approach to gather data because, although the design of the game is important to encourage participation and engagement, the actual game task is not crucial. Also, most of the other games for data collection can be played by a single person, but our game requires teams that belong to a pre-existing social network, which allows us to gather information about the existing social group. This kind of game is especially useful for data-collection in the social sciences because the games provide a natural data-intensive, non-intrusive interaction with participants. The game is applied here to learn about social communications, manipulation methods, and reactions to natural occurrences of friendliness or aggression.

6 Conclusions and Future Work

The results of these experiments, although allowing for some improvement, show great promise regarding the capability of the game of detecting a certain type of bully, while causing less fatigue to the participants (5th graders that played the game expressed a lot of enjoyment regarding their playing experience). Our system is capable of reporting to a teacher the identity of those children with high probability of being bullies, hopefully giving them the opportunity for early detection and intervention.

In future work, it is very important to focus on improving the precision of our results. This can be done in several

ways: The first option is to change the model such that it includes more information that is currently available from the game (e.g., point transactions, conversations in the chat public channel, final score in the game, etc.). Second, it would be appropriate to develop a better way to handle the imbalance in the data, for example, combining the power of max-margin methods (such as SVM) with the simplicity of naive bayes may prove useful as well as including background knowledge to allow our system to identify not only bullies with high executive functions but also those with low executive functions.

7 Acknowledgments

This work was supported in part by NSF IIS grant 09-17123 titled *Scaling Up Inference in Dynamic Systems with Logical Structure*, NSF IIS grant 09-68552 titled *SoCS: Analyzing Partially Observable Computer-Adolescent Networks*, and by a grant from the Defense Advanced Research Project Agency (DARPA) as part of the Machine Reading Program under Air Force Research Laboratory (AFRL) prime contract no. FA8750-09-C-0181. Any opinions, findings, and conclusion or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

References

Bernstein, M.; Tan, D.; Smith, G.; Czerwinski, M.; and Horvitz, E. 2009. Collabio: a game for annotating people within social networks. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, UIST '09, 97–100. New York, NY, USA: ACM.

Bronfenbrenner, U. 1977. Toward and experimental ecology of human development. *American Psychologist* 32(7):513–531.

Chang, C.-C., and Lin, C.-J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27.

Elkan, C., and Noto, K. 2008. Learning classifiers from only positive and unlabeled data. In *Proceedings of the KDD'2008*, 213–220.

Espelage, D., and Holt, M. 2001. Bullying and victimization during early adolescence: Peer influences and psychosocial correlates. *Journal of Emotional Abuse* 2:123–142.

Espelage, D.; Mebane, S.; and Adams, R. 2004. Empathy, caring, and bullying: Toward an understanding of complex associations. In *Bullying in American Schools: A social-ecological perspective on prevention and intervention*, 37–61. New Jersey, NJ, USA: Lawrence Erlbaum Associates.

Groh, G., and Hauffa, J. 2011. Characterizing social relations via nlp-based sentiment analysis. *International AAAI Conference on Weblogs and Social Media*.

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The weka data mining software: an update. *SIGKDD Explor. Newsl.* 11:10–18.

Kozlov, M., and Johansen, M. 2010. Real behavior in virtual environments: psychology experiments in a simple virtual-reality paradigm using video games. *Cyberpsychology, behavior, and social networking.* 13:711–714.

Lieberman, H.; Dinakar, K.; and Jones, B. 2011. Let's gang up on cyberbullying. *IEEE Computer* 93–96.

Maloof, M. A. 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML 2003 Workshop on learning from imbalanced data sets II*, volume 21, 1263–1284.

Mancilla-Caceres, J. F., and Amir, E. 2011. Evaluating commonsense knowledge with a computer game. In Campos, P. F.; Graham, T. C. N.; Jorge, J. A.; Nunes, N. J.; Palanque, P. A.; and Winckler, M., eds., *INTERACT (1)*, volume 6946 of *Lecture Notes in Computer Science*, 348–355. Springer.

Mancilla-Caceres, J. F.; Pu, W.; Amir, E.; and Espelage, D. 2012. A computer-in-the-loop approach for detecting bullies in the classroom. In Yang, S. J.; Greenberg, A. M.; and Endsley, M. R., eds., *SBP*, volume 7227 of *Lecture Notes in Computer Science*, 139–146. Springer.

Monks, C. P.; Smith, P. K.; and Swettenham, J. 2005. Psychological correlates of peer victimization in preschool: social cognitive skills, executive function and attachment profiles. *Aggressive Behavior* 31(6):571–588.

Naubuzoka, D. 2009. Teacher ratings and peer nominations of bullying and other behaviour of children with and without learning difficulties. *Educational Psychology* 23(3):307–321.

Noto, K.; Saier, M.; and Elkan, C. 2008. Learning to find relevant biological articles without negative training examples. In *Ai 2008: Advances in Artificial Intelligence*, volume 5360, 202–213.

Paradesi, S., and Shih, F. 2011. Globalidentifier: Unexpected personal social content with data on the web. *International AAAI Conference on Weblogs and Social Media*.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Richard, and Jessor. 1992. Risk behavior in adolescence: A psychosocial framework for understanding and action. *Developmental Review* 12(4):374 – 390.

Schölkopf, B.; Smola, A. J.; Williamson, R. C.; and Bartlett, P. L. 2000. New support vector algorithms. *Neural Comput.* 12:1207–1245.

Schlkopf, B.; Platt, J.; Shawe-Taylor, J.; Smola, A.; and Williamson, R. 2001. Estimating the support of a high-dimensional distribution. *Neural Computation* 13:1443–1471.

Seigne, E.; Coyne, I.; Randall, P.; and Parker, J. 2007. Personality traits of bullies as a contributory factor in workplace bullying: An exploratory study. *International Journal of Organization Theory and Behavior* 10(1):118–132.