Identifying Bullies with a Computer Game

Eyal Amir\(^1\), Juan F. Mancilla-Caceres\(^2\), Wen Pu\(^2\), and Dorothy Espelage\(^3\)

\(^1\) Faspark (formerly University of Illinois CS),  
\(^2\) Department of Computer Science,  
\(^3\) Department of Educational Psychology  
University of Illinois at Urbana-Champaign

Microsoft Research Machine Learning Summit
Classroom Bullying is not Cyberbullying.

Bullying behavior is not explicit.

Bullies are bi-strategic.
Bayesian Network Model Classifies Real-World Bullies from Computer-Game Behavior

A BN model provides the most balanced results in terms of accuracy and recall. Other classifiers scored low on precision.
Overview

- Game for collecting data about interactions of grade-school students

- Model *game* interactions with two-layer Bayesian Networks to identify bullies in the *real-world* classroom.

- Relational inference for social networks: Two-Layer Networks ERG(P*) model
Real-World Behavior

- 60% of middle school students say that they have been bullied, and 16% of school staff believe that students are bullied.

- A bully is 6 times more likely to be incarcerated by the age of 24.

- Surveys rely on self report, are time-consuming, and not sufficiently detailed.

- Observational studies impose fatigue on the researcher.
Welcome!

Who do you want to form the team with?

<table>
<thead>
<tr>
<th>Yes</th>
<th>No</th>
<th>No Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>No</td>
<td>No Preference</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>No Preference</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>No Preference</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>No Preference</td>
</tr>
</tbody>
</table>
Game – Collaborative Stage

Collaborative Question 1:
What NBA coach recently won his 10th NBA championship?
- Phil Jackson
- Doc Rivers
- Pat Riley
- Greg Popovich

Ensure that all members of your team provide the SAME answer.

Submit Answer
Peek at the answer (-5 coins)
Game – Competitive Stage

Competitive Question 2:

According to one hundred 5th graders, who is the best basketball player in the NBA?

- Chris Paul
- Carlos Boozer
- Kobe Bryant
- Joakim Noah

All members of your team must provide a DIFFERENT answer, and at least one must choose the answer in bold.

Submit Answer

Peak at the answer (-5 coins)
Behavior Observed During Gameplay

Diana to Bob: send me some coins
Bob to Diana: no
Diana to Bob: don't talk to me then
Diana to Emily: send me some coins
Emily to Diana: no
Diana to Emily: why not
Emily to Diana: no
Diana to Emily: i said why
Emily to Diana: i hate you that why
Diana to Emily: f*** you
Diana to Adam: hi
Adam to Diana: bi
Diana to Adam: shut up boy
Adam to Diana: you

Positive Nomination  Bully (according to survey)  Bully/Victim (according to survey)
Negative Nomination  Victim (according to survey)  Same team
Coins Exchanged During Gameplay

Arrow X to Y means X gave more coins to Y than Y to X
Comparing Survey and Game Data

Hypothesis: Participants labeled as *bullies*, use a different strategy in the game than *non-bullies* (i.e., behave quantitatively different).

Method: Set of 2 (Bully/Non-Bully) x 2 (Collaborative/Competitive Stage) ANOVA to find if there is a significant difference in the features extracted from the game.

- No interaction between the independent variables were found.
- Bullies/Non-Bullies send more private messages, peek more at the answer and send more negative nominations.
- Everybody sent and received more messages during the Collaborative stage.
- Content analysis of the message has been done (but it is not included in this analysis).
Results of 2x2 ANOVAs

<table>
<thead>
<tr>
<th>Feature</th>
<th>Non-Bullies</th>
<th>Bullies</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peeked</td>
<td>1.06</td>
<td>1.55</td>
<td>0.023</td>
</tr>
<tr>
<td>Prsent</td>
<td>15.77</td>
<td>23.12</td>
<td>0.037</td>
</tr>
<tr>
<td>NegNomSent</td>
<td>3.31</td>
<td>5.05</td>
<td>0.090</td>
</tr>
</tbody>
</table>

**Peeked**: Number of times peeked at the answer.

**Prsent**: Number of chat messages sent through private channel.

**NegNomSent**: Number of negative nominations sent during Stage 1.
Learning real-world roles from pair-wise interactions.

- Adam
- Bob
- Charlie
- Diana
- Emily

- Bully
- Non-Bully
Modeling the Interactions

- **Coercive** = 1 if Adam sent more coercive messages than prosocial to Bob.
- **Negative Affect** = 1 if Adam sent more negative affect messages than positive ones to Bob.
- **Nominated** = 1 if Adam nominated Bob to be on the same team, -1 if the nomination was negative, 0 otherwise.
- **Same Team** = 1 if Adam and Bob belong to the same team.
Assigning a Label

Each two-layer Bayesian network returns a probability that \( S \) is a bully.

\[
P_{i}^{S} = P \left( S \mid C, N, O, T \right) = \frac{\sum P \left( S, R_{i} = r, C, N, O, T \right)}{\sum_{r} \sum_{s} P \left( S, R, C, N, O, T \right)}
\]

The final label assigned to \( S \) is found by comparing the average of all the predictions with a threshold found during training.

\[
\bar{P}^{S} = \frac{1}{n} \sum_{i} P_{i}^{S}
\]

If \( \bar{P}^{S} > \text{thr} \) then \( S \) is labeled as a bully.
Lifted Inference for Social Networks

Relational Model: 2-Layer BN

\[ \Pi_{a,b,i} \Pr(f_i(a, b) | Bul(a), Bul(b)) \]

Instead of separate models, do inference in the complete model
Exponential-family Random Graph Models (ERGMs)

- A popular model in social network and statistics literatures
- Use subgraph statistics to describe networks

<table>
<thead>
<tr>
<th>feature</th>
<th>count</th>
<th>density</th>
</tr>
</thead>
<tbody>
<tr>
<td>edge</td>
<td>7</td>
<td>0.333 (7/21)</td>
</tr>
<tr>
<td>triangle</td>
<td>1</td>
<td>0.029 (1/35)</td>
</tr>
<tr>
<td>2-star</td>
<td>11</td>
<td>0.105 (11/105)</td>
</tr>
<tr>
<td>3-star</td>
<td>5</td>
<td>0.036 (5/140)</td>
</tr>
<tr>
<td>4-star</td>
<td>1</td>
<td>0.001 (1/105)</td>
</tr>
</tbody>
</table>
ERGMs

Distribution over order-\(n\) graph \(g \in G\)

\[
p_{\theta}(g) = \frac{1}{Z'(\theta)} \exp(\theta^T \phi(g))
\]

Features (densities for \(r\)-subgraphs)

\[
\phi(g) = \left( \frac{t(g,H_1)}{t(K_n,H_1)}, \frac{t(g,H_2)}{t(K_n,H_2)}, \ldots, \frac{t(g,H_r)}{t(K_n,H_r)} \right)
\]

- \(H_i\)s are feature subgraphs; \(t(g,H_i)\) counts the number of subgraphs in \(g\) that are isomorphic to \(H_i\); \(K_n\) is order-\(n\) complete graph.
An approximation to log-partition function for ERGMs
Summary

- Game data can be used to identify bullies in the real world
- Game-based analysis helps psychologists collect data
- Some bullies can be identified with game features, not all
- Lifted Inference with machine-learning models is promising for social networks and social-role detection
Extracted Features

- Amount of private messages sent and received during the collaborative and competitive stages.

- The number of public messages received.

- The number of coins sent and received.

- Positive nominations sent and received.

- Negative nominations sent and received.

- Reciprocated positive/negative nominations, and unreciprocated nominations.
Comparing Survey and Game Data

Hypothesis: Participants labeled as *bullies*, use a different strategy in the game than *non-bullies* (i.e., behave quantitatively different)

Method: Set of 2 (Bully/Non-Bully) x 2 (Collaborative/Competitive Stage) ANOVA to find if there is a significant difference in the features extracted from the game.

- No interaction between the independent variables were found.
- Bullies/Non-Bullies send more private messages, peek more at the answer and send more negative nominations.
- Everybody sent and received more messages during the Collaborative stage
- Content analysis of the message has been done (but it is not included in this analysis).
We have access to more pair-wise interactions than to classrooms.

6 Classrooms

96 Participants

>3000 interactions

Eyal Amir et.al - Identifying Bullies
Behavior Observed During Gameplay.

- Positive Nomination
- Negative Nomination

Non-Bully
Bully

216 to 214: send me some coins
214 to 216: no
216 to 214: don't talk to me then

216 to 208: send me some coins
208 to 216: no
216 to 208: why not
208 to 216: no
216 to 208: i said why
208 to 216: i hate you that why
216 to 208: f*** you

216 to 211: hi
211 to 216: bi
216 to 211: shut up boy
211 to 216: you