Investigating the Behavior of Malicious Actors Through the Game of Mafia

Samee Ibraheem
University of California, Berkeley, Berkeley, California, United States

Vael Gates
UC Berkeley, Berkeley, California, United States

John DeNero
UC Berkeley, Berkeley, California, United States

Tom Griffiths
Princeton University, Princeton, New Jersey, United States

Abstract

In deception games, deceivers must find ways to draw in unknowing bystanders, and bystanders must develop strategies for detecting falsehoods. What are the strategies that people use in these roles, and can computer systems also detect these behaviors? We address this question through text-based games of Mafia, wherein players are assigned to deceptive roles (mafia) or roles incentivizing detecting deception (bystanders). We find that participants adopt sophisticated role-based strategies, wherein the mafia, who are outnumbered but know the identities of all players, act carefully to secure the votes of the bystanders by speaking more even as verbose speakers tended to be eliminated. These role-based behaviors were distinct enough that a computational classifier could distinguish between mafia and bystanders with 70.3% accuracy and outperform human players. Understanding the systematic features defining honest and deceptive players advances our ability to automatically detect online deceit and grasp group dynamics in real-world collaboration.