Understanding Toxic Behavior in Online Games

Haewoon Kwak*
Telefonica Research
Barcelona, Spain
kwak@tid.es

ABSTRACT

With the remarkable advances from isolated console games to massively multi-player online role-playing games, the online gaming world provides yet another place where people interact with each other. Online games have attracted attention from researchers, because i) the purpose of actions is relatively clear, and ii) actions are quantifiable. A wide range of predefined actions for supporting social interaction (e.g., friendship, communication, trade, enmity, aggression, and punishment) reflects either positive or negative connotations among game players, and is unobtrusively recorded by the game servers. These rich electronic footprints have become invaluable assets for the research of social dynamics.

In particular, exploring negative behavior in online games is a key research direction because it directly influences gaming experience and user satisfaction. Even a few negative players can impact many others because of the design of multi-player games. For this reason these players are called *toxic*. The definition of toxic play is not cut and dry. Even if someone follows the game rules, he could be considered toxic. For example, killing one player repetitively is often deemed toxic behavior, although it does not break game rules at all. The vagueness of toxicity makes it hard to understand, detect, and prevent it.

League of Legends (LoL), created by Riot Games with 70 million users as of 2012, offers a new way to understand toxic behavior. Riot Games develops a crowdsourcing framework, *the Tribunal*, to judge whether reported toxic behavior should be punished or not. Volunteered players review user reports and vote for either pardon or punishment. As of March 2013, 105 million votes had been collected in North America and Europe.

We explore toxic playing and reaction based on large-scale data from the Tribunal [1]. We collect and investigate over 10 million user reports on 1.46 million toxic players and corresponding crowdsourced decisions made in the Tribunal. We crawl data from three different regions, North America, Western Europe, and Korea, to take regional differences of user behavior into account. To obtain the comprehensive view of toxic playing and reaction based on huge data collection, we answer following research questions in a bottom-up approach: how individuals react to toxic players, how teams interact with toxic players, how general toxic or non-toxic players behave across the match, and how crowds make a decision on toxic players. We find large-scale empirical support for some notoriously difficult theories to test in the wild, which are bystander effect, ingroup favoritism, black sheep effect, cohesion-performance relationships, and attribution theory. We also discover that regional differences affect the likelihood of being reported and the proportion of being punished of toxic players in the Tribunal.

We then propose a supervised learning approach for predicting crowdsourced decisions on toxic behavior with large-scale labeled data collections [2]. Using the same sparse information available to the reviewers, we trained classifiers to detect the presence, and severity of toxicity. We built several models oriented around in-game performance, reports by victims of toxic behavior, and linguistic features of chat messages. We found that training with high agreement decisions resulted in more accuracy on low agreement decisions and that our classifier was adept in detecting clear cut innocence. Finally, we showed that our classifier is relatively robust across cultural regions; our classifier built from a North American dataset performed adequately on a European dataset.

Ultimately, our work can be used as a foundation for the further study of toxic behavior.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Sociology, Psychology*

Keywords

League of Legends; online video games; toxic behavior; crowdsourcing; machine learning

*This work is collaboration with Seungyeop Han and Jeremy Blackburn.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579690.

1. REFERENCES

[1] H. Kwak and S. Han. "So Many Bad Guys, So Little Time": Understanding Toxic Behavior and Reaction in Team Competition Games. *Under review.*

[2] J. Blackburn and H. Kwak. STFU NOOB! Predicting Crowdsourced Decisions on Toxic Behavior in Online Games. In *WWW*. 2014.