The Identification of Deviance and its Impact on Retention in a Multiplayer Game

Kenneth Shores
GroupLens Research
University of Minnesota,
Minneapolis, MN 55455
shores@cs.umn.edu

Yilin He
GroupLens Research
University of Minnesota,
Minneapolis, MN 55455
hexxx305@umn.edu

Kristina L. Swanenburg
Department of Psychology
University of Pittsburgh,
Pittsburgh, PA 15213
kls94@pitt.edu

Robert Kraut
Human-Computer Interaction Institute
Carnegie Mellon University,
Pittsburgh, PA 15213
robert.kraut@cs.cmu.edu

John Riedl
GroupLens Research
University of Minnesota,
Minneapolis, MN 55455
riedl@cs.umn.edu

ABSTRACT
Deviant behavior in online social systems is a difficult problem to address. Consequences of deviance include driving off users and tarnishing the system's public image. We present an examination of these concepts in a popular online game, League of Legends. Using a large collection of game records and player-given feedback, we develop a metric, *toxicity index*, to identify deviant players. We then look at the effects of interacting with deviant players, including effects on retention. We find that toxic players have several significant predictive patterns, such as playing in more competitive game modes and playing with friends. We also show that toxic players drive away new players, but that experienced players are more resilient to deviant behavior. Based on our findings, we suggest methods to better identify and counteract the negative effects of deviance.

Author Keywords
deviance; cooperation; retention; online games; norms; League of Legends;

ACM Classification Keywords
H.5.3. Collaborative Computing

INTRODUCTION
In this paper, we examine the impact of deviance on retention in the online multiplayer game League of Legends. Deviance can be defined as an aberration from normative behavior. Norms are shared ways of feeling, thinking and behaving [15]. Deviation from norms can lead to rejection of the deviant [27] and to decreases in a group's cohesion [31] and performance [11].

The phenomenon of deviance has been explored in the past in both offline (see [14] for a review) and online settings. Research suggests that deviance is quite common in virtual groups [7][9]. Greater anonymity in a virtual environment has been linked to increased deviance [8][24], and at the same time may lead to an enhanced awareness of group-specific norms [17]. If group-specific norms encourage disruption, then an individual could adhere to their group's norms by deviating from community norms. Designers of online systems are motivated to decrease deviance because deviance is assumed to have a negative impact on user retention[2][20][7]. User retention is important online because virtual groups, from guilds in online games to Wikipedia editors, suffer from high turnover [10][23].

Since retention is so important online, and deviance poses a threat to that retention, system designers have attempted to decrease deviance in a variety of ways. Many forums utilize surveillance techniques to decrease deviance. Moderators review posts to their forums and may alter or delete inappropriate posts or even remove members who repeatedly violate the forum's rules. Some designers have developed more creative solutions to deal with a deviant user, including "hellbanning" the user by making all of that user's actions invisible to others [2]. As a result, the user is seemingly ignored by the entire community. Other techniques to dissuade deviance include allowing other users to rate specific artifacts created by their peers and to report violations when they occur. A popular extension of user-rated comments is to automatically hide comments receiving mostly negative votes. In many online games, including League of Legends, players can send reports about disruptive players to moderators who review the report and dole out punishments if necessary.

Another technique that is commonly used to decrease deviance is reputation systems. Reputation systems decrease

CSCW'14, February 15–19, 2014, Baltimore, Maryland, USA.

http://dx.doi.org/10.1145/2531602.2531724

the ephemeral consequences of poor online interaction by creating a reputation that follows a user or player. This is in contrast to applying ratings to individual comments, inventories or other system-specific artifacts. These reputations are based on ratings by other users. Reputation systems have proven effective online; positive reputations on Ebay have been linked to improved performance [26]. The threat of the "shadow of the future" is used by reputation systems. Ratings persist from one interaction to the next, threatening to harm future interactions [25]. These systems simultaneously punish deviant users and allow other users to avoid them.

Riot, the developer of League of Legends, has put substantial effort into studying player deviance, and has focused on a type of deviant behavior it labels as "toxic". Riot defines "toxicity" as any behavior that negatively impacts other players' experiences [30]. Toxic behavior is a subset of deviant behavior and is often considered "un-sportsmanlike". It includes a large range of behaviors, such as sending offensive messages or intentionally helping the opposing team. However, specific definitions of toxic behavior may be situational to subsections of the community.

In an effort to reduce toxic behavior, Riot has implemented a system called the Tribunal, which allows players to judge player behavioral reports of suspected toxic players [29]. Riot has also experimented with exposing players to a variety of messages while a match is loading. By varying message text and color, Riot has found that certain combinations decreased reports of deviance by up to 10% [20]. This demonstrates that subtle design decisions and interventions can be effective in influencing deviant behaviors.

Deviant behavior in League of Legends is complex and embedded in community norms. Some segments of the user population may have different expectations around use of profanity or giving commands. It is important to note that the reporting system used by Riot indicates behavior that is considered deviant - if cursing is acceptable, a player will not report their teammate for it. Because the Tribunal has players review the behavior of other players, a reported player is being flagged to be evaluated by peers according to community norms. Riot provides suggested guidelines for evaluating behavior [28], but determining acceptable behavior is ultimately situational.

Present Study

In order to address problems related to deviance, we first need to detect and identify deviant behavior. We can then explore the effects of this behavior on users. One specific concern to social system designers is user retention, and measuring the effect of deviance on retention is an important part of assessing it as a threat.

The present study examines deviance and retention in League of Legends. First, we develop a metric based on indications of deviant behavior in our dataset. We look specifically at indications of toxic behavior, as it is the most concerning type of deviance in League of Legends. We examine our metric in several situations, looking for situations where toxic behavior differs. We then perform a regression with this metric and several in-game behavior measures to look for predictors of players leaving the game. We explore the findings of this regression with regard to social and structural factors in the game, and use this to draw conclusions for the design of other games and social systems.

BACKGROUND: LEAGUE OF LEGENDS

League of Legends is a game which is played in many independant matches. In each match, two teams of five players compete on a symmetric map to destroy the opposing team's objectives while defending their own[1]. League of Legends has a large global player base, with about 12 million active players daily [21]. Teams are often made up of strangers who will never play together again. Effective communication and teamwork are important factors in winning a match, and a lack of cooperation often leads to losing a match [28]. This need for cooperation and communication with teammates makes a player's experience highly dependent on social interactions with other players.

League of Legends has several different game modes. Teams are composed of players who choose one of over one hundred characters, and characters can be classified into roles.[2]

In each match, players control a single character. With their character, they must destroy the enemy structures while defending their team's structures. During a match, players unlock abilities and purchase items which make their character and their teammate's characters stronger. Players must use reflexes, knowledge of the game, and cooperation to defeat the enemy team. An average match lasts 31 minutes, but can range from 20 to 60 minutes or more. League of Legends tracks many statistics through each match, including number of kills of enemy characters, number of deaths of each player's character, and number of assists (helping a nearby teammate score a kill). Kills reward gold, and assists reward gold in addition to the gold given for a kill, increasing the team's total reward and encouraging cooperation.

The primary game mode is called "normal" (or "unranked") mode. Normal is the most popular game mode, and in it a player's skill rating is hidden. A second game type is "ranked" mode, where players are given a public skill rating. Winning or losing a match in this mode will alter this rating, which is visible to other players and friends. This mode is often considered more competitive, and fewer players participate. It is also restricted to only allow maximum level players. The final game mode is "co-op vs. AI" - a mode where players fight computer-controlled enemies. Each game mode has different expectations and social norms around it, and many players play primarily a single mode. Co-op vs. AI mode will not be discussed further. It represents a qualitatively different game experience because there are not two teams of human players.

[1]The limited popularity of other game modes excluded them from our analyses. They are not discussed in this work.

[2]Each character in League of Legends has a distinct set of abilities, but differences at the level of individual characters do not affect this work and will not be discussed.

Matchmaking

In League of Legends, players are matched with teammates and enemies based on three factors: experience, player skill, and grouping. The matchmaking system attempts to make "fair" matches for each game, so both teams have an equal probability of winning when the match begins.

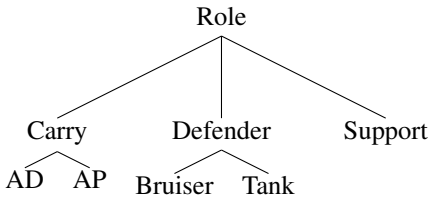
Experience: Players who have a similar amount of experience with the game are matched together. Players earn experience points for each match they play and these points advance a player's level until they reach the maximum level of 30. Players gain levels at a relatively consistent rate per match, so below level 30, player level is a reasonable proxy to "number of matches played." League of Legends has a steep learning curve, and giving players a level to indicate experience helps reduce frustration for newer players.

Player Skill: The matchmaking system also tries to match players with similar skill levels. This matchmaking system is based on Elo, a rating system initially designed for chess and commonly used in online competitive games [12]. Thus players who have won more matches, or matches against harder opponents, will be matched against more skilled opponents in the future.

Grouping: Finally, the matchmaking algorithm attempts to match groups of players who have queued (joined a match) together against other groups who have queued together [32]. While the impact of this is smaller than other factors, the matchmaking algorithm is designed with the assumption that players who queue together will play more effectively than strangers who are grouped together.

Character Roles

Characters are classified into roles based on a character's ability set - each character has four abilities, and these abilities can emphasize damage-dealing, defense or utility. Character roles can be defined as carry, defender and support. We further divide these roles based on community norms, as follows:



Carry is split between 'AP' and 'AD', named for their in-game damage type. Both roles deal lots of damage but have low survivability. 'Tank' and 'Bruiser' are both high-survivability roles. Tanks tend to have abilities that stun or control other characters, while bruisers are durable but also do substantial damage. The 'Support' role specializes in utility and control abilities, such as a 'stun' ability which causes an enemy to be unable to act for a few seconds. Supports often rely on cooperation with teammates playing other roles to best make use of their abilities.

Team composition in League of Legends is quite flexible. Each player's roles are chosen at the start of each match through team discussion. One group may have five carries, while another has four defenders and a support. Game knowledge and tactical decisions both contribute to group composition, and team cooperation is an important factor in winning a match.

Figure 1. A screenshot of Duowan's add-on software. 1) Thumbs-up and thumbs-down scores. 2) Player's favorite characters. 3) Game modes and win statistics. 4) Recent match history.

METHODS

Dataset

Data was collected through a software add-on made by Duowan[3]. The add-on creates a reputation system for League of Legends players on a game server in China. Whenever a player using the add-on played a match, statistics for all players in that match were recorded by the add-on. The dataset contains 2.5 million players playing 18.25 million matches across three months. Players in the data set have played an average of 67.26 matches during the data collection period. The data collected by Duowan and used in this study is public data, and our analysis was performed without intervention, and without handling any private data.

In addition to game statistics, the dataset also contains feedback for each player, given by other players using the add-on. This feedback is either positive ("*Player* is a good player, have a thumbs up.") or negative ("I don't want to play a match with *Player* in the future."), and is displayed in chat for all the players at the end of the match. The message is clearly labeled as coming from the add-on, and the feedback given is stored by the add-on and displayed on a player profile on Duowan's site.

Because the data was only collected from games where at least one user is using the addon, the method of data collection risks missing some matches played. We were able to compute an estimate of the percentage of collected matches, and found that 91.6% of matches played between the first and last sighting of a player were tracked by the add-on. Because newer players are less likely to use the add-on, we suspect that

[3] `www.duowan.com`

most players missed in our dataset were very low level. This is the primary reason we do not look at the retention patterns of very new players in this work.

Toxic Behavior Detection

A survey on the Duowan's forums asked players to choose the reason that best fit why they have given players a thumbs-down [1]. Posted in July of 2012, the survey presented nine responses. However, because the survey was in Mandarin Chinese and used game-specific terminology, responses were translated and clustered by the authors into a smaller set of categories. The majority of the 138 participants (78.23%) stated that they would give a thumbs-down for various deviant behaviors, including verbal abuse (outside of socially acceptable cursing or yelling), refusing to continue to play a match and helping the enemy team win. A minority (16.98%) stated that thumbs-down was given to players who did not perform well in the match. The remaining responses indicated that some users never give thumbs-downs. The categories we consider to be deviance-related are the same as the types of reports that Riot collects regarding player behavior that indicate a violation of either community or group norms. This means that thumbs-downs are mostly given for deviant behavior. Thus we believe that thumbs-downs are a reasonable proxy for measuring deviance.

To identify players in our dataset that exhibited deviant behavior, we developed a metric using the thumbs-up and thumbs-down ratings that a player received. These counts were given on a weekly basis, so we looked at the changes in a player's scores for each week as an indication of deviant behavior. A metric was developed that combined the two scores as a way to compare between players. Called *toxicity index*, the metric was based on the ratio of thumbs-down to thumbs-up during a given week. Because players participated in a varying number of matches, we added a normalization factor (1 for thumbs-down and 4 for thumbs-ups) to the resulting ratio to reduce the likelihood of having extreme scores from a small number of matches. The constants for the normalization factor were tuned by hand, by examining the distribution of resulting toxicity indexes in our sample. We also excluded players who participated in fewer than ten matches during a given week, because a player with a small number of matches would have few thumbs-up and thumbs-downs given, which we felt gave us low confidence for making inferences about their general behavior.

The final formula for player toxicity index was defined as follows:

$$toxicity\ index = \frac{thumbs\text{-}down + 1}{thumbs\text{-}up + 4}$$

We expected that players exhibiting more deviant behavior would have a high ratio of thumbs-down to thumbs-up scores, based on the reasons for giving a thumbs-down in the survey above.

Our dataset had several factors and limitations that influenced the toxicity index. For many of our questions, we sampled players from matches during a particular week. For these players, we calculated their toxicity index in the previous week and used it as a predictor of behavior during the studied week. Because we could not determine specific matches where thumbs-ups or thumbs-downs were given, scores from the previous week were used to prevent a conflation of predictors and outcomes.

Another limitation was that a player's teammates could only give a thumbs-up to the player when they had won a match, and a thumbs-down when they lost. A concern might be "What if deviant players are just *good* at the game?". An Elo-based matchmaking system is unlikely to allow players to maintain a win-ratio that is far from .5, and to address this concern, an analysis was performed assessing player win-ratios. The analysis found an average win:loss ratio of .4961, with a standard deviation of .0633. Given that active players in our dataset have a mean of 30 and a median of 20 matches per week, this is an imbalance of only a few wins or losses for most players.

HYPOTHESES

Broadly, we have several hypotheses that can be split into two research questions:

1. *What factors predict a high toxicity index?*
2. *What factors predict quitting or continuing to play the game?*

Toxicity Index

We had several hypotheses to test with this metric. We hypothesized that the type of game modes played, a player's role in a match, attributes of a player's teammates and overall experience with the game will affect a player's toxicity index. Since competitiveness and aggression are highly correlated [6] and aggression can be considered a type of deviance, we predicted that players in more competitive game modes would act more deviantly. Ranked matches are considered more competitive because by playing them, a player's ranked Elo is prominently displayed and publicly available. Additionally, high-Elo players can earn rewards for their account.

> **Hypothesis 1.1:** Players who primarily play more competitive game modes will have a higher toxicity index than players who play primarily less competitive game modes.

To test this, we generated toxicity indexes for 139,855 ranked players and 201,145 unranked players from matches in our dataset. All players were level 30. We then compared toxicity indexes based on the game mode sampled from.

Similarly, we believed that the role a player chooses in a match is related to the player's level of deviance. The high-damage "carry" role can be viewed as the most competitive role because both "AD" and "AP" characters focus on doing lots of damage to enemy players. The primary goal of all roles is competition with the opposing team, but carries often compete within their own team to be the best damage-dealer. Scoring a high number of kills or dealing a lot of damage is a clear indicator of a successful carry, and a predictor of winning. Support roles can be viewed as the least competitive in this regard. A successful support contributes to winning by

assisting and protecting their teammates, and by adapting to team needs. The focus of supports is to cooperate with their teammates and help them take objectives, and there is little competition within teams to be the "best cooperator". Defenders (including both Tank and Bruiser roles) must split focus between damage-dealing and protecting other members of their team, and thus we predict an intermediate level of competitiveness.

Hypothesis 1.2: Playing characters with mostly damage-dealing abilities will predict the highest toxicity index, and playing characters with cooperative or supportive abilities will predict the lowest toxicity index.

This analysis was applied to the same samples of players selected for Hypothesis 1.1. For each sampled game, the player's toxicity index was added to the appropriate measure based on the role of their selected character, split by game mode. From this, the mean toxicity index was calculated for each role.

Playing with friends should also impact the player's perceived deviance. Deviant behavior is more likely to be reinforced when playing with a friend than with strangers, in turn increasing its frequency. Additionally, friends are likely to imitate each other's deviant behavior through social learning, which would also increase the frequency of deviance [3].

Hypothesis 1.3: Playing with friends will increase the number of thumbs-downs given by non-friends.

We classified players as *friends* whenever they played two or more matches together, on the same team, in the week we measured their toxicity index. Because of the size of the game's player pool, the likelihood of two strangers being incorrectly classified as friends in this system is extremely low. A player's friends could give them a thumbs-up after winning a match, and arguably would do so for different reasons than a stranger, so we were not able to use our toxicity index here. Instead, we looked at the raw number of thumbs-up and thumbs-down votes per match played in a given week, and categorized players by the number of friends they played with in that week. We assumed that friends would not give thumbs-downs to each other, and divided the number of thumbs-downs received by the number of strangers played with during the week. Frequency of thumbs-ups received was calculated to include friends. 10,000 players were sampled for this analysis.

We also expected that a player's long-term experience with the game would impact the player's rate of deviance. Players who have played longer will have been socialized to the community norms [19]. In League of Legends, a player might assume that some behavior is normative because deviant behavior is more memorable than non-deviant behavior [4]. A player may exhibit this behavior in a later group where the behavior is considered deviant. Therefore, variations in group norms and conflict between group and community norms may lead to behavior being regarded as deviant. There are other reasons to suspect that experienced players may act in deviant ways. As players become more comfortable with the game, they may become more easily frustrated by players who do not know effective playing strategies, or with players who use strategies that disagree with their expectations. This increased frustration could lead to more deviant behavior [5]. We expect toxicity index to be a proxy to deviant behavior, therefore:

Hypothesis 1.4 Players who have been playing longer will have a higher toxicity index.

Retention

We were also interested in the relationship between deviance and player retention. We considered two types of retention: short-term and long-term. Being retained in the short-term means continuing to play the current game session. Players can play matches successively for as long as they wish, and a continuous block of matches is considered a session. Long-term retention refers to either permanently quitting the game, or leaving for an extended period. To better quantify these cutoffs, we performed an analysis of time between the end of one match and the beginning of the next. Referring to work by Geiger & Halfaker [13] we plotted time between matches on the bucketed log-scaled plot in Figure 2. For short-term retention we chose a threshold of one hour: any time a player has at least one hour between the end of a match and the beginning of the next, we assume they had ended their session. This cut-off was chosen for several reasons. It is the same session duration used by Geiger and Halfaker in their Wikipedia session analysis, and anecdotally, we as players feel it is a reasonable and safely-inclusive cutoff; Someone might spend fourty-five minutes waiting for friends to finish a concurrant match, but taking more than an hour between matches indicates the player is doing something else. For long term retention we chose a threshold of one week. Any time a player plays no matches for at least one week, we assume they have either left the game permanently or are taking an extended break.

Hypothesis 2.1: Playing with teammates with a high toxicity index will decrease retention.

Though we consider several factors related to retention in our regressions, we predict that playing with players who frequently exhibit deviant behavior will have a negative impact on player retention. As mentioned earlier, deviance is assumed to have a negative impact on retention, and this is a central question for our work. We expect this prediction to apply to both short and long-term retention.

To discover factors that impacted retention, 341,295 players were sampled from our dataset. The toxicity index was calculated for the selected players, as well as for every player they encountered during a chosen week. For each match played, we labeled players for whether it was their last match for both short- and long-term retention; did the player quit the session or did they leave the game for an extended period after that match? A binomial logistic regression was performed for each type of retention on each player-match combination. In addition to in-game performance statistics such as "number of kills" and "did the player win the match?", the player's toxicity index and the average toxicity index of their teammates (excluding themselves) were included in the regression. We

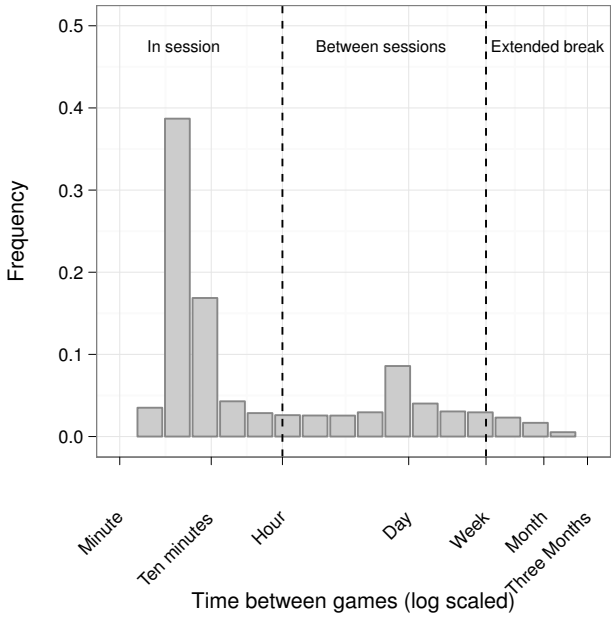


Figure 2. Distribution of time spent between matches.

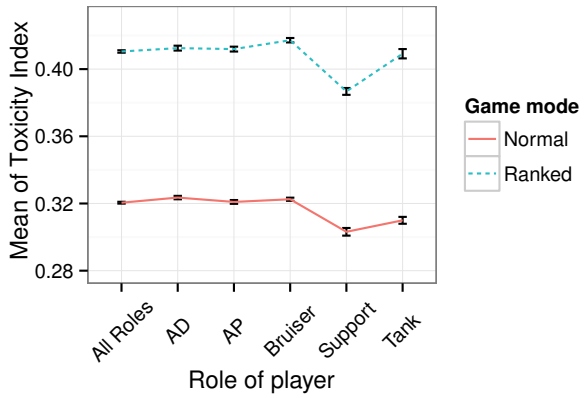


Figure 3. Toxicity index and player role. AP and AD are versions of the "carry" role. Tanks and Bruisers are types "defenders".

also partitioned our regressions by player level. Since level is a proxy to number of matches played, we consider players who are not maximum level to be at a different stage of learning the game and learning community norms. This is supported by several game design decisions, including increased character customization options for high-level players and by limiting ranked mode to only be played by maximum level players.

RESULTS AND DISCUSSION

RQ1: *What factors predict a high toxicity index?*

Hypothesis 1.1:

Players who primarily play more competitive game modes will have a higher toxicity index than players who play primarily less competitive game modes

Players were sampled and divided into into ranked (n=139,855) and unranked (n=201,145) categories. All players were level 30. A plot of the mean toxicity index of the players with standard error bars is shown in Figure 3. Ranked players had an mean toxicity index of 0.41 with a standard deviation of 0.27. Unranked players had an mean toxicity index of 0.32 with a standard deviation of 0.25. These groups are statistically significantly different (t-test $p < 0.001$).

Discussion

As predicted, players in the more competitive ranked matches were associated with higher toxicity indexes than players in normal matches. Thus, even though a more competitive attitude is the norm for players playing the ranked game mode, players were exhibiting more deviant behavior than players in normal mode matches.

Hypothesis 1.2:

Playing characters with mostly damage-dealing abilities will predict the highest toxicity index, and playing characters with cooperative or supportive abilities will predict the lowest toxicity index.

We classified 341,295 level 30 players according to role and game mode. Figure 3 plots the mean and standard error of toxicity index of different roles on the X axis. Both carry roles in Figure 3 were found to have a relatively high toxicity index. Bruisers (one type of defender) also had a high toxicity index for unranked, and the highest mean toxicity index for ranked. This disagreed with our hypothesis that both bruisers and tanks would have lower toxicity than the carry roles. For both ranked and unranked, supports had a low toxicity index. Tanks also had a lower toxicity index, relative to AP, AD and bruisers.

We performed an ANOVA for both the ranked and unranked samples and found significant differences between them: *Ranked* ($F(4, 139850) = 39.31, p < .001$)) *Unranked* ($F(4, 201160) = 23.83, p < .001$)). A Tukey's HSD test was performed between roles for each mode. For ranked, support was statistically significantly different than all other roles ($p < .05$). No other roles had significant differences in ranked mode. For unranked, support was significantly different from AP, AD and bruisers ($p < .05$). The tank role was also significantly different from AP, AD and bruiser ($p < .05$). Tanks and supports were not significantly different from one another in unranked play.

Discussion

Roles were found to vary in average toxicity index. Hypothesis 1.2 predicted that the roles with the most damage-dealing abilities would have the highest toxicity index. The least competitive role, support, was associated with lower toxicity levels than other roles in both ranked and unranked matches. Tanks were found to have significantly lower toxicity indexes than both carry roles in unranked matches, but were not significantly different in ranked matches. Bruiser mean toxicity index was not significantly different from the more damage-dealing roles in either game mode, and in ranked matches bruiser had the highest toxicity index of all roles. Differences between bruisers and tanks went against our classifica-

tion, which expected them to be similar. We suspect that this difference stems from the way they are designed. The roles have different specializations: Bruisers often have abilities oriented towards dealing damage, while tanks generally have abilities that are best used to protect teammates or survive incoming damage. While the grouping of bruisers and tanks led to an incorrect prediction about bruisers, the idea that a character's ability set relates to the player's toxicity index is still reasonable. Importantly, these findings do not support or discourage a causal relationship, where players who would exhibit behavior leading to thumbs-downs are attracted to the more damage-oriented characters and roles.

Hypothesis 1.3:

Playing with friends will increase the number of thumbs-downs given by non-friends.

Players were bucketed according to the average number of friends they played with across all matches for a given week. We expected bias in the number of thumbs-ups for players who frequently played with friends, because we expect friends to be inclined to give each other thumbs-ups. Thus, the rate of thumbs-up is the number of thumbs-ups received, divided by the total number of players encounterd in the sampled time period. The thumbs-down rate is the number of thumbs-downs received divided by the total number of non-friends played with. A large number of players did not play with any friends during the sampled week - these players were put in bucket 0. Otherwise, players were bucketed by the ceiling of the mean number of friends played with across matches.

Figure 4 shows the mean and standard error plots for the rate of thumbs-ups and thumbs-downs by number of friends. We performed an ANOVA and found significant differences ($F(1, 9998) = 33.36, p < .001$), and a Tukey's HSD test was performed to assess differences between buckets. For thumbs-down, playing with 0 friends was significantly different from playing with either 1 or 2 friends ($p < .05$ for both). For thumbs-up, playing with 0 friends was significantly different from all other buckets, and playing with 1 friend was significantly different from playing with 2 or 3 friends ($p < .05$ for all).

Discussion

Our analysis shows that playing with friends increases the number of thumbs-downs a player recieves, especially for players who on average play with 1 or 2 friends. A corollary to this is that playing with friends is associated with higher levels of deviance. Players may become socialized to display deviant behaviors through positive reinforcement from friends. A player may experience a conflict between community norms and the norms of their group of friends - behavior that deviates from community norms may be encouraged by friends.

Hypothesis 1.4:

Players who have been playing longer will have a higher toxicity index.

We also examined toxicity indexes based on the total number of matches played. We computed the mean toxicity index for players, bucketed by total number of games played. However, we found counfounds for addressing this hypothesis with our data. Further analysis indicates that a lack of data about low level players causes our normalization ratio to bias the mean toxicity index. Because total thumbs-up and thumbs-down measures will increase over time, mean toxicity index will tend to move from a mean dictated by our base toxicity index ratio to a mean that represents the community at max level. This reaffirms the notion that the dataset's collection methods discourage analysis involving low-level players. In other explorations we control for this by selecting currently active players, rather than assessing measures of whole user histories.

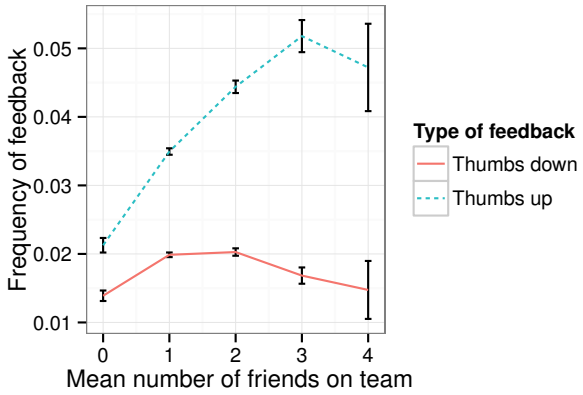


Figure 4. Average number of friends on a team vs frequency of feedback signals. Frequency is the rate of given thumbs-up or thumbs-down per encountered teammate.

Discussion

While our dataset is not sufficient to explore this question, we believe that the factors suggeting the hypothesis are still reasonable. There are several potential game-situations where confidence in game-understanding may encourage deviant behavior.

One common situation is when a player believes their team can no longer win the match, leading them to quit by exiting the game. This leaves their four teammates at a severe disadvantage. While their teammates may or may not agree with this prediction, community norms dictate they should play it out or end the game by team-vote. This is a scenario where a player's experience influences their confidence about a situation, and is used as a justification for their deviant behavior. We would encourage future work to explore this concept.

Retention

Hypothesis 2.1:

Playing with teammates with a high toxicity index will decrease retention.

A binomial logistic regression was performed to look at factors that predict short-term and long-term retention. This regression was split into four categories in two dimensions: short-term vs long-term, and level<30 vs level=30.

	Short Term Retention						Long Term Retention					
	Level < 30			Level = 30			Level < 30			Level = 30		
	β	SE	P	β	SE	P	β	SE	P	β	SE	P
(Intercept)	.873	.010	$< .001$	.960	.006	$< .001$	5.502	.081	$< .001$	6.900	.088	$< .001$
Match Length	-.231	.010	$< .001$	-.194	.006	$< .001$	-.150	.059	.011	-.074	.071	.298
Win Current Match?	.082	.010	$< .001$	-.016	.006	.004	.113	.062	.067	.005	.073	.944
Average Teammate Toxicity Index	.011	.042	0.790	.032	.010	$< .001$	-.351	.178	.049	-.206	.112	.066
Player Toxicity Index	.028	.046	0.545	.086	.016	$< .001$	.640	.269	.017	.370	.261	.156
Player Level	.078	.025	.002				.394	.127	.002			
Friends in Match	.117	.010	$< .001$	.120	.006	$< .001$	.069	.069	.003	.398	.101	$< .001$
Player Level : Teammate Toxicity Index	-.005	.050	0.923				.618	.258	.016			
Player Level : Players Toxicity Index	.032	.048	0.498				-.276	.320	.388			
Player Toxicity Index : Teammate Toxicity Index	-.009	.028	0.738	-0.039	.018	.029	-.189	.217	.385	.263	.296	.374

Table 1. Four logistic regressions corresponding to short term retention and long term retention. Positive β indicates a prediction of increased likelihood of retention. Inputs are scaled. Variables omitted from certain regressions are left blank.

Short Term, Level < 30

Match length was a significant negative predictor of short term retention ($\beta = -0.231, p < .001$) meaning that longer matches predicted a reduced likelihood of continuing the session. Winning the current match was a significant positive predictor of continuing the session ($\beta = .082, p < .001$). Level was a significant predictor as well - higher level players were more likely to continue a session ($\beta = 0.078, p < .002$). Number of friends in the current match was the final significant predictor ($\beta = 0.117, p < .001$). Playing with friends increased the likelihood of continuing a session. The toxicity index of the player or the player's teammates was not a significant predictor of short-term retention behavior for players below level 30.

Short Term, Level = 30

Match length was also a significant predictor of discontinuing the session for level 30 players ($\beta = -0.194, p < .001$). Unlike players under level 30, winning the match predicted a decreased likelihood of continuing the session ($\beta = -0.016, p < .004$). A high average toxicity index of teammates was a strong predictor of continuing the session ($\beta = .032, p < .001$). Players with higher toxicity indexes were also more likely to continue a session ($\beta = .086, p < .001$). Similar to players under level 30, playing with friends was a strong predictor of continuing a session ($\beta = 0.120, p < .001$). The positive prediction of player-toxicity index and teammate-toxicity index is moderated when both factors are high ($\beta = -0.039, p < .029$), showing that an all-around high toxicity index tends to predict that a player will end their session.

Long Term, Level < 30

Match length was a significant predictor of players below level 30 taking an extended break from (or quitting) the game ($\beta = -.150, p < .011$). A high toxicity index of teammates was also a stronger predictor for driving these players to quit ($\beta = -.351, p < .049$). However, a player with a higher toxicity index was more likely to continue playing ($\beta = .640, p < .017$), and players with a higher level (who had played more matches over-all) were also less likely to leave ($\beta = .394, p < .002$). Playing with friends had a smaller impact, but increased long-term retention ($\beta = .069, p < .003$). A player's level and their teammate's toxicity had an interaction effect which suggests that higher level players were more resilient to toxic teammates ($\beta = .618, p < .016$).

Long Term, Level = 30

Number of friends was the only significant predictor of long-term retention at level 30. ($\beta = 0.397, p < .001$). Playing with friends increases the likelihood of continuing to play over long periods, and playing with more friends increases it further. Surprisingly, toxicity index had no significant impact on retention once a player reached level 30.

Discussion

Hypothesis 2.1 predicted that playing with teammates with a high toxicity index will decrease both short term and long term retention. We found that toxicity index had a variety of effects on retention.

For players who had reached level 30, a higher average toxicity index of their teammates increased the likelihood of continuing their session. And a player with a higher toxicity also has an increased likelihood of continuing their session. However, when both a player's toxicity index and their teammate's average index is high, players tend to end their session. There are several possible interpretations for this finding, and it warrants further study. We suspect that deviant players may enjoy deviant behavior, or at least are not deterred by any immediate consequences, but that a clash between a high toxicity index player and teammates with a high toxicity index may be frustrating enough to motivate taking a break.

For long term retention, the story is a bit different. Players who are below level 30 and have a high toxicity index are more likely to continue to play League of Legends. Also, playing with higher-than-average toxicity index teammates predicts long-term departure: frustrating teammates can drive off players who are learning the game. However, at level 30 players are more resistant to teammates with a high toxicity index, and are not driven away from the game for long periods by them.

Why are players below level 30 likely to be driven off when players at max level are not? One possibility is a selection effect against players with low tolerance for toxic behavior. In this situation, players must either develop tolerance for toxic behavior or be driven off. While the regression indicates that players with low toxicity indexes are more vulnerable to being driven off (supporting the notion of a selection effect), it is unclear that toxicity is learned by the surviving players. Future work could explore this dynamic.

Beyond toxicity index, the regression considers several other factors which show predictive power. Playing with friends is a consistent predictor of continuing to play the game, in both the short-term and the long-term for players at all levels of experience. Even though playing with friends also corresponds to a higher toxicity index, having friends to play with has a clear impact on retaining players. Furthermore, for players at maximum level, playing with friends was the only significant predictor of long-term retention - having friends to play with is more likely to keep an experienced player from leaving the game.

Another factor effecting both short-term and long-term retention was match length. Not surprisingly, longer matches led players to end their sessions. But longer matches were also associated with a greater chance of players under level 30 quitting the game for good. Less experienced players may have qualitatively different game experiences as they are not just playing but also learning basics of gameplay and norms around gameplay, and designers should consider this when they design to acclimate new users. In League of Legends, leaving in the middle of a match is discouraged both by community norms and by in-game warning messages, and players who are unexpectedly kept in a long match may be upset if staying causes (for example) schedule conflicts outside of the game.

CONCLUSIONS AND DESIGN IMPLICATIONS

In this paper we examined the nature and effects of a concerning type of deviant "toxic" behavior in League of Legends, a multiplayer online game. We developed a metric called the *toxicity index* for measuring this type of deviant behavior based on peer evaluations of other players, and we use it to examine several predictors of toxic behavior. We also used this metric to explore the commonly held assumption that players who exhibit this deviant behavior reduce the retention rates of other players in the community.

Our findings support the notion that interactions with toxic players decrease the retention rates of new players. Many games have similar social design elements to League of Legends. Competitive first-person shooters often highlight team-based gameplay [16], and player-versus-player options in MMORPGs are frequently team-based [22]. As such, our findings regarding player retention have many implications for designers of other systems.

Designers should take an active role in teaching and encouraging positive community norms, especially when interacting with newer players. For example, designers can propose a set of community norms in order to actively discourage common behaviors that make the game less enjoyable for others. The League of Legends "Summoner's Code" [28] is on implementation of this approach.

Our study examined the effects of a signal for deviant behavior in an online game which encourages both competitive and cooperative behavior. Our findings suggest that emphasizing the cooperative aspects of the system instead of the competitive aspects may be one technique for decreasing deviant behavior. However, some users may be attracted to the system because of its competitive aspects. It is important that interventions are designed to address deviant behavior directly, rather than by simply decreasing competitiveness. This is a problematic balance for designers because actions that are perceived to be deviant by users can influence their retention.

We also saw a consistent link between playing with friends and retention. Playing with friends was the most significant predictor of improved long term retention for players who had reached the maximum level in game. Designers should take note of referral programs like League of Legend's refer-a-friend system, and should incentivise and facilitate in-game activities with friends. This is consistent with findings of social motivation in other domains, such as the finding that relationships in blogging networks predicted better retention rates when those relationships were actively maintained [18]. Though League of Legends has a "friends list", our classification of friendships only consisted of users who played games with one another. Future work should investigate more closely the effects of different kinds of relationships on retention in a competitive space.

Limitations

In this work, variables were not experimentally manipulated and thus causal inferences are limited. Future studies should attempt to experimentally manipulate the factors found to be significantly related to deviance and retention in order to clarify causative directions.

ACKNOWLEDGMENTS

The authors would like to thank Duowan for generously providing us with data.

Additionally, we would thank Riot for making a game that is fascinating to study and fun to play.

Finally, John Riedl passed away between the writing and the publication of this paper. We would like to dedicate this work in his memory. An avid player, insightful mentor and beloved friend, he helped make this work possible.

REFERENCES

1. In what kind of situation would you thumbs-down your teammate? http://bbs.duowan.com/thread-27997215-1-1.html.
2. Atwood, J. Suspension, ban or hellban? http://www.codinghorror.com/blog/2011/06/suspension-ban-or-hellban.html.
3. Bandura, A., and McClelland, D. C. *Social learning theory*. General Learning Press, New York, 1977.
4. Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. Bad is stronger than good. *Review of general psychology 5*, 4 (2001), 323.
5. Berkowitz, L. Frustration-aggression hypothesis: Examination and reformulation. *Psychological bulletin 106*, 1 (1989), 5973.
6. Buss, A. H., and Perry, M. The aggression questionnaire. *Journal of personality and social psychology 63*, 3 (1992), 452459.

7. Davis, J. P. The experience of bad behavior in online social spaces: A survey of online users. *Social Computing Group, Microsoft Research* (2002).
8. Douglas, K. M., and McGarty, C. Identifiability and self-presentation: Computer-mediated communication and intergroup interaction. *British journal of social psychology 40*, 3 (2001), 399416.
9. Dubrovsky, V. J., Kiesler, S., and Sethna, B. N. The equalization phenomenon: Status effects in computer-mediated and face-to-face decision-making groups. *Human-Computer Interaction 6*, 2 (1991), 119146.
10. Ducheneaut, N., Yee, N., Nickell, E., and Moore, R. J. The life and death of online gaming communities: a look at guilds in world of warcraft. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (2007), 839848.
11. Dunlop, P. D., and Lee, K. Workplace deviance, organizational citizenship behavior, and business unit performance: The bad apples do spoil the whole barrel. *Journal of Organizational Behavior 25*, 1 (2004), 6780.
12. Elo, A. E. *The rating of chessplayers, past and present*, vol. 3. Batsford London, 1978.
13. Geiger, R. S., and Halfaker, A. Using edit sessions to measure participation in wikipedia. In *Proceedings of the 2013 conference on Computer supported cooperative work* (2013), 861870.
14. Hogg, M. A., Fielding, K. S., and Darley, J. Fringe dwellers: Processes of deviance and marginalization in groups. *The social psychology of inclusion and exclusion* (2005), 191210.
15. Hogg, M. A., and Reid, S. A. Social identity, self-categorization, and the communication of group norms. *Communication Theory 16*, 1 (2006), 730.
16. Jansz, J., and Tanis, M. Appeal of playing online first person shooter games. *CyberPsychology & Behavior 10*, 1 (2007), 133–6.
17. Lea, M., and Spears, R. Computer-mediated communication, de-individuation and group decision-making. *International Journal of Man-Machine Studies 34*, 2 (1991), 283301.
18. Lento, T., Welser, H. T., Gu, L., and Smith, M. The ties that blog: Examining the relationship between social ties and continued participation in the wallop weblogging system. In *3rd Annual Workshop on the Weblogging Ecosystem* (2006), 12.
19. Levine, J. M., and Moreland, R. L. Group socialization: Theory and research. *European review of social psychology 5*, 1 (1994), 305336.
20. Lin, J. The science behind shaping behavior in online games. http://gdcvault.com/play/1017940/The-Science-Behind-Shaping-Player.
21. MacManus, C. League of legends the world's 'most played video game'. http://news.cnet.com/8301-17938_105-57531578-1/league-of-legends-the-worlds-most-played-video-game/.
22. Nardi, B., and Harris, J. Strangers and friends: collaborative play in world of warcraft. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, CSCW '06, ACM (New York, NY, USA, 2006), 149–158.
23. Panciera, K., Halfaker, A., and Terveen, L. Wikipedians are born, not made: a study of power editors on wikipedia. In *Proceedings of the ACM 2009 international conference on Supporting group work* (2009), 5160.
24. Resnick, P. The social cost of cheap pseudonyms. *Journal of Economics & Management Strategy 10*, 2 (2001), 173199.
25. Resnick, P., Kuwabara, K., Zeckhauser, R., and Friedman, E. Reputation systems. *Communications of the ACM 43*, 12 (2000), 4548.
26. Resnick, P., and Zeckhauser, R. Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system. *Advances in applied microeconomics 11* (2002), 127157.
27. Richard, F. D., Bond, C. F., and Stokes-Zoota, J. J. One hundred years of social psychology quantitatively described. *Review of General Psychology 7*, 4 (2003), 331363.
28. Riot Games. The summoner's code. http://na.leagueoflegends.com/articles/The_Summoners_Code .
29. Riot Games. The tribunal. http://na.leagueoflegends.com/tribunal/en/faq/.
30. Skleres, G., K. T. C., and Pavlas, D., L. J. Improving player behavior in league of legends. http://east.paxsite.com/schedule/panel/improving-player-behavior-in-league-of-legends.
31. Wellen, J. M., and Neale, M. Deviance, self-typicality, and group cohesion the corrosive effects of the bad apples on the barrel. *Small Group Research 37*, 2 (2006), 165186.
32. Zileas. LOL matchmaking explained. http://na.leagueoflegends.com/board/showthread.php?p=122813 .