A Study on Reducing Malicious Replies on the Internet: An Approach by Game Theory

Sanghun Lee

Abstract—Since the advent of the information era, the Internet has brought various positive effects in everyday life. Nevertheless, recently, problems and side-effects have been noted. Internet witch-trials and spread of pornography are only a few of these problems. In this study, problems and causes of malicious replies on internet boards were analyzed, using the key ideas of game theory. The study provides a mathematical model for the internet reply game to devise three possible plans that could efficiently counteract malicious replies. Furthermore, seven specific measures that comply with one of the three plans were proposed and evaluated according to the importance and utility of each measure using the orthogonal array survey and SPSS conjoint analysis. The conclusion was that the most effective measure would be forbidding unsigned user access to malicious replies. Also notable was that some analytically proposed measures, when implemented, could backfire and encourage malicious replies.

Keywords—Conjoint Analysis, Game Theory, Internet, Malicious Replies, Prisoner’s Dilemma

I. INTRODUCTION

Since the advent of the information era, the Internet has brought such various positive effects as ease and speed in everyday life. Nevertheless, recently, problems and side-effects have been noted. Internet witch-trials, language profanity, and spread of pornography are only a few of these problems. Among these problems, the most conspicuous is the malicious reply problem. In such places of vast internet industry as South Korea, malicious replies on internet boards have brought the social problem of online violence that resulted in suicides of such famous actors as Choi Jin-Sil. Recently, the malicious reply problem has become a substantial threat to the information society. Especially as the access to the Internet has increased, online libel, slander, and sexual harassment, the number of cases increasing year by year. Reports by the Korea Internet Safety Commission, a randomly selected newspaper website had an alarming malicious reply rate of over 50%.

II. ASPECTS OF INTERNET MALICIOUS REPLIES

A. Severity of Malicious Replies

As internet users increasingly spend time in internet boards and forums, the Internet is becoming an outlet of gratuitous attacks and indiscriminate advertisements. In South Korea, where the negative side-effects of internet board culture are recognized as one of the most serious social problems, the practice of posting replies that abuse the power of anonymity is called “malicious reply culture.” Malicious replies are replies on internet boards that contain such problems as witch-trial, language profanity, and pornography advertisement.

In this study, the example of South Korea, the country of high malicious reply rates, will be observed.

In Korea, as malicious replies have recently increased in number, public complaints about the Internet has also increased. Table I shows recently occurred cyber violence, categorized according to the type of social problem. According to the table, from 2004 to 2005, public complaints regarding cyber violence have increased by 8,608 cases, and, especially for those cases of libel and violence, the numbers rose significantly by a factor of 34. Also, according to a research by the Korea Internet Safety Commission, a randomly selected newspaper website had an alarming malicious reply rate of over 50%.

<table>
<thead>
<tr>
<th>Year</th>
<th>Libel</th>
<th>Sexual Harassment</th>
<th>Violence</th>
<th>Prodigal</th>
<th>Disrupting Social Discipline</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>470</td>
<td>27,603</td>
<td>597</td>
<td>41</td>
<td>5,324</td>
<td>34,035</td>
</tr>
<tr>
<td>2005</td>
<td>1,513</td>
<td>29,898</td>
<td>2,667</td>
<td>250</td>
<td>8,315</td>
<td>42,648</td>
</tr>
<tr>
<td>2006.6</td>
<td>828</td>
<td>9,760</td>
<td>1,043</td>
<td>2,958</td>
<td>5,986</td>
<td>20,575</td>
</tr>
</tbody>
</table>

Also, Korea Communications Standards Commission reported that, from 2001 to 2006, it held a total of 24,385 cases of counseling regarding online libel, slander, stalking, and sexual harassment, the number of cases increasing year by year. These statistics show that, although the Internet has become the vital part of life, it has also become the haven of violence.

Sanghun Lee is a student in Korean Minjok Leadership Academy, Hoengseong, Gangwon 225823 South Korea (phone: 010-4167-3350; e-mail: rttrt2@naver.com)
TABLE II
ANNUAL SURVEY OF ONLINE VIOLENCE [2]

<table>
<thead>
<tr>
<th>Year</th>
<th>Total</th>
<th>Libel</th>
<th>Sexual Harassment</th>
<th>Stalking</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>1,054</td>
<td>278</td>
<td>204</td>
<td>22</td>
<td>550</td>
</tr>
<tr>
<td>2002</td>
<td>3,616</td>
<td>1,248</td>
<td>224</td>
<td>53</td>
<td>2,091</td>
</tr>
<tr>
<td>2003</td>
<td>4,127</td>
<td>1,916</td>
<td>557</td>
<td>95</td>
<td>1,649</td>
</tr>
<tr>
<td>2004</td>
<td>3,913</td>
<td>2,285</td>
<td>332</td>
<td>81</td>
<td>1,225</td>
</tr>
<tr>
<td>2005</td>
<td>8,406</td>
<td>5,735</td>
<td>889</td>
<td>193</td>
<td>1,589</td>
</tr>
<tr>
<td>2006.6</td>
<td>3,179</td>
<td>2,087</td>
<td>451</td>
<td>62</td>
<td>579</td>
</tr>
<tr>
<td>Total</td>
<td>24,385</td>
<td>13,549</td>
<td>2,647</td>
<td>506</td>
<td>7,683</td>
</tr>
</tbody>
</table>

B. Current Aspect of Malicious Replies
In August to November, 2008, Dong-A Daily and Korea Internet Security Agency analyzed about 320,000 replies of news articles with more than 200 replies. Among these replies, 69,671 replies posted on major articles were analyzed by hand. The result was that malicious replies without constructive opinion were 14.3% of the total replies.

As the analyzed replies were those that remained after the website's reply clean-ups, the actual percentage of malicious replies is speculated to be higher. Also, the most common type of malicious replies was threats and swears, which accounted for 65.3% of the replies categorized as malicious reply. The second most common was spam reply, which accounted for 36.5%.

An interesting problem was the problem of reply oligopoly, as the top 5% most enthusiastic users who replied posted 30.5% of all the replies, and the top 50% posted 44% of all the replies. As, according to the research, average of 2.5% netizens post replies, calculation gives that 0.12% of users make 1/3 of all replies.

Through malicious replies, many victims suffer from such assaults as libel and sexual harassment. Consequently, regulatory policies are becoming more restrictive, obstructing active internet environment. Many users who post malicious replies show the tendency to attract other's attention to raise their hits. These users abuse the anonymity of the Internet to post violent replies, lead witch-trials, or fabricate public opinion. The social problem of malicious replies is becoming a substantial concern.

C. Netizen’s Competition for Hits
On internet boards, the netizens' interaction can be seen as a game. In the situation where the number of users who normally visit a board is fairly constant (and therefore restricted), as some netizens use malicious replies to induce readers (or hits), the readers of other replies decreases. Thus, there is a game for hits in which the normal users of the board participate. Game theory can be used to set and analyze the model for this game.

Game theory is a branch of mathematics that analyzes the favorable decisions of a game player in a competitive game that holds many players[4]. More specifically, game theory studies the ways that a game player can maximize his payoff in the premise that players do not know each other's strategy. In such complicated societal situations as the internet reply problem, game theory can be used to model and analyze the situation mathematically.

III. CONSTRUCTION OF THE MALICIOUS REPLY GAME MODEL
A. Diminishing Marginal Utility of Malicious Reply
Malicious replies have diminishing marginal utility, as can be seen in Table II. This means that the rate of increase of the malicious reply function decreases as time passes. This is because, as the number of malicious replies increase, the number of hits that each malicious reply receives decreases.

On the other hand, for healthy replies, marginal utility increases. Malicious replies tend to repel readers, but healthy replies tend to attract more readers to visit the board. Thus, equilibrium forms between the marginal utilities of malicious and healthy replies.

B. Application of Game Theory
Internet boards can be seen as a game in which each user competes for his own profit. Focusing on a particular user who would post a malicious reply, there is more probability the user could attract hits when there are fewer existing malicious replies on the board. Consequently, a user's best decision is to hope that other users do not post malicious replies and write an attention-grabbing malicious reply himself. This means that
cooperation with other players negatively effects the player. This type of game is called the multi-player non-cooperative game. Table III is one speculation that could model the malicious reply game in current internet boards. This type of game is specifically called the ‘prisoner's dilemma,’ where the rational choice is to post a malicious reply. Cooperating with other users and posting a healthy reply can result in each player getting 20 hits, but rational choice leads each user to post a malicious reply, resulting in 10 hits for each user. Thus, the optimal result (Pareto optimum) is not achieved, and the Nash equilibrium (Pareto sub-optimum) becomes the equilibrium state.

**TABLE III\(^{1}\)**

<table>
<thead>
<tr>
<th></th>
<th>Malicious Reply</th>
<th>Healthy Reply</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malicious Reply</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Healthy Reply</td>
<td>5</td>
<td>20</td>
</tr>
</tbody>
</table>

Let us generalize the case to the case where more than three players participate in the reply game. Assuming that a player decides to post a malicious reply, his expected payoff with respect to the other \(x\) number of players who post malicious replies can be modeled by the function:

\[
f(x) = \frac{T - r(p - x)}{x} - c\]

\(T = \text{(Total Number of Readers)}\)
\(r = \text{(Average Number of Replies per Person)}\)
\(p = \text{(Number of Reply Game Players)}\)
\(c = \text{(Cost of Posting Replies)}\)

Also, randomly assuming, for the sake of a rough understanding, that there are 10 number of game players, 100 (established 80, increased 20) number of readers increased by malicious replies, 8 average number of replies per person, and 5 unit of cost of posting replies \(f(x)\) gives Table IV as the number of hits with respect to the number of game players. Here, as the number of malicious replies increase, the number of hits per malicious reply decreases. On the other hand, for healthy replies, cooperation between players brings the increase of hits. As the number of players who cooperate increases, the effect of the cooperation increases, and it is possible to model the expected payoff posting healthy replies as an exponential function. For number of players who choose to post healthy replies, the expected number of hits is modeled by:

\[
s(x) = r x^k\]

\(k = \text{(Cooperative exponent)}\)

Also, assuming that a player decides to post a healthy reply, his expected payoff with respect to the other \(x\) number of players who post healthy replies can be modeled by the function:

\[
g(x) = \frac{s(x)}{x} - c\]

Again, assuming that \(k = 1.5\), the expected payoff is calculated as shown in Table V. Here, as the number of players who choose to post healthy replies increase, the expected payoff increases, and, in cases, the rational choice is to post a healthy reply. Thus, for healthy replies, this means that cooperation with other players positively effects the player. This type of game is called the multi-player cooperative game.

Nevertheless, healthy reply is not the rational choice. If all players cooperate to post healthy replies, expected payoff per person is large (20.3 in the assumed case). If one player betrays and posts a malicious reply, however, his expected payoff is 23, and the rational choice of each player becomes posting a malicious reply. As a result, all players post malicious replies, forming the Nash equilibrium.

**C. Analysis of Players’ Decision Making Process**

On a specific internet board in a specific time, let the number of connected users \(N\), the total number of hits on a reply \(T\), the number of users who choose to post malicious replies \(x\), and non-participants who only read replies \(y\). The function of number of hits is then denoted \(s (N-x-y)\).
The expected payoff for a person who chooses to post a malicious reply is as follows:
\[
\frac{T - s(N - x - y)}{x}
\]
(1)
Also, if the probability \( p \) that a player chooses malicious reply follows the binomial distribution, the probability \( P_x \) that there are \( x \) players who choose malicious reply is
\[
P_x = C_x^N p^x (1 - p)^{N-x}
\]
(2)
On the other hand, if probability \( q \) that a user just reads replies also follow the binomial distribution, the probability \( P_y \) that the number of readers \( y \) is
\[
P_y = C_N^{N-x} q^y (1 - q)^{N-x-y}
\]
(3)
Using (1), (2), and (3), the average expected payoff can be calculated. The expected payoff of a player who chooses to post a malicious reply is
\[
\sum_{x=0}^{N} \left[ C_x^N p^x (1 - p)^{N-x} \sum_{y=0}^{N-x} (C_N^{N-x} q^y (1 - q)^{N-x-y} (T - s(N - x - y))) \frac{1}{x} \right]
\]
(4)
The expected payoff of a player of who chooses to post a healthy reply is
\[
\sum_{x=0}^{N} \left[ C_x^N p^x (1 - p)^{N-x} \sum_{y=0}^{N-x} (C_N^{N-x} q^y (1 - q)^{N-x-y} (T - s(N - x - y))) \frac{1}{x} \right]
\]
(5)
Thus, the player must compare (4) and (5) to make a rational choice. If the value of (5) is bigger than that of (4), the player’s rational choice is to post a healthy reply. This means that if other variables are controlled, the increase in the total number of hits results in the decrease in the number of healthy replies. In other words, the more active an internet board is, the more numerous malicious replies are. This means that restricting hits can prevent malicious replies. This measure, however, collides with the issue of active internet environment to which the world’s internet technology is aiming.

1. \( \frac{\partial \Pi}{\partial T} = -\frac{1}{x} < 0 \)

This inequality implies that, when other variables are controlled, the increase in the total number of hits results in the decrease in the number of healthy replies. In other words, the more active an internet board is, the more numerous malicious replies are. This means that restricting hits can prevent malicious replies. This measure, however, collides with the issue of active internet environment to which the world’s internet technology is aiming.

2. \( \frac{\partial \Pi}{\partial (N-x-y)} = -\frac{s(N-x-y)}{(N-x-y)^2} + \frac{s'(N-x-y)}{x} < 0 \)

Here, as \( s(N-x-y) \) is a strictly decreasing function, and \( s'(N-x-y)<0 \). This means that the increase of healthy replies will result in the decrease in the expected payoff of choosing healthy reply and a comparative increase of malicious replies. Hence, to prevent malicious replies, there is a need to support healthy replies so that \( s'(N-x-y) \) is increased.

3. \( \frac{\partial \Pi}{\partial s(N-x-y)} = \frac{1}{N-x-y} + \frac{1}{x} > 0 \)

This inequality implies that, under variables controlled, the increase in cooperation-induced hits results in the increase of healthy replies. Hence, supporting healthy reply cooperation and increasing the derivative of \( s(N-x-y) \) can theoretically reduce malicious replies.

IV. DEVISING OF MEASURES AND EVALUATION
A. Devising of Measures against Malicious Replies
For an internet community as a whole, the best public interest is to reach the Pareto optimum, or the condition where
all the participants maximize their expected payoff. Nonetheless, the equilibrium state of current internet communities, as seen in Table VI, is not the Pareto optimum, but rather region I. This equilibrium state is the result of the prisoner's dilemma form of game involving internet replies. Hence, to reduce malicious replies and reach the Pareto optimum, three plans that break the current deadlock can be suggested. Furthermore, these plans, combined with the mathematical analysis of the game model, can be specified into seven detailed measures against malicious replies.

Plan 1 Encouraging and advertising cooperative healthy reply ([I]→[III] shift in Table VI)

Plan 2 Encouraging hits for healthy reply ([I]→[II] shift in Table VI)

Plan 3 Restricting hits for malicious reply ([III]→[III] shift in Table VI)

To objectively evaluate the efficiency of the above proposed measures, there is a need for proper evaluation weights. A commonly used method of analyzing such plans and measures as above is the conjoint analysis. Conjoint analysis, frequently used in marketing and product management, uses data such as preference rank to find the attributes and attribute values that customers consider important[7].

To reduce the malicious replies, attributes about replies that most internet users consider important must be analyzed so that the importance of the attributes can be compared. Hence, the use of conjoint analysis is reasonable for the evaluation of the devised measures against malicious replies.

Conjoint analysis involves the process of selecting reasonable attributes, attribute levels, and survey forms. In this study, there are \(2^7 = 128\) possible combinations of selected measures (which can also be viewed as the number of subsets of the set of the seven measures).

Generally, a conjoint analysis survey asks the subject to rank the given combinations. The full method, however, is inapplicable when the number of the combinations is too large. This is the case in this study, as there are 127 combinations. Hence, a more convenient method using the orthogonal design will be implemented. The orthogonal design method, used frequently in conjoint analysis, is usually in the form of orthogonal array, shown in Table VII[6]. With seven measures, the orthogonal array reduces the 127 combinations into 8 combinations. These 8 combinations can be used in a survey for conjoint analysis.

### Table VI

<table>
<thead>
<tr>
<th>Malicious Reply</th>
<th>Healthy Reply</th>
</tr>
</thead>
<tbody>
<tr>
<td>[II] C</td>
<td>[III] D</td>
</tr>
</tbody>
</table>

### Table VII

<table>
<thead>
<tr>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Card Type</th>
<th>1 2 3 4 5 6 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Card 1</td>
<td>1 1 1 1 1 1 1</td>
</tr>
<tr>
<td>Card 2</td>
<td>1 1 1 2 2 2 2</td>
</tr>
<tr>
<td>Card 3</td>
<td>1 2 2 1 1 2 2</td>
</tr>
<tr>
<td>Card 4</td>
<td>1 2 2 1 1 2 2</td>
</tr>
<tr>
<td>Card 5</td>
<td>2 1 2 1 2 1 2</td>
</tr>
<tr>
<td>Card 6</td>
<td>2 1 2 1 2 1 2</td>
</tr>
<tr>
<td>Card 7</td>
<td>2 2 1 1 2 1 2</td>
</tr>
<tr>
<td>Card 8</td>
<td>2 2 1 1 2 1 2</td>
</tr>
</tbody>
</table>
TABLE VIII
ORTHOGONAL ARRAY FOR THE CONJOINT ANALYSIS OF THE 7 MEASURES

<table>
<thead>
<tr>
<th>Ms. 1</th>
<th>Ms. 2</th>
<th>Ms. 3</th>
<th>Ms. 4</th>
<th>Ms. 5</th>
<th>Ms. 6</th>
<th>Ms. 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>○</td>
<td>×</td>
<td>×</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>×</td>
<td>○</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>○</td>
</tr>
<tr>
<td>×</td>
<td>×</td>
<td>×</td>
<td>○</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

For an objective evaluation of the efficiency of the proposed measures, a survey about the measures was devised and was taken by high school students who are familiar with the internet environment and internet reply culture.

Date: December, 2010
Subject: 30 Korean high school students
Method: paper survey in a classroom setting; survey sheet and cards provided in the Appendix

C. Data Analysis and Discussion

The statistical analysis process of the survey results was facilitated by the conjoint analysis tool of IBM's SPSS 12.0. The result of the conjoint analysis, including the importance and utility of each measure, is shown on Table IX. The utility of each measure describes the satisfaction that users feel about the measure, and the importance denotes the percent ratio of the utility of the measure in the case in which all measures are implemented.

TABLE IX
ATTRIBUTES AND LEVELS OF THE SURVEY CARD

<table>
<thead>
<tr>
<th>Measure</th>
<th>Utility</th>
<th>Importance(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2917</td>
<td>15.98</td>
</tr>
<tr>
<td>2</td>
<td>-0.4083</td>
<td>17.79</td>
</tr>
<tr>
<td>3</td>
<td>0.3500</td>
<td>11.57</td>
</tr>
<tr>
<td>4</td>
<td>-0.2083</td>
<td>12.69</td>
</tr>
<tr>
<td>5</td>
<td>0.2833</td>
<td>10.23</td>
</tr>
<tr>
<td>6</td>
<td>0.1833</td>
<td>16.15</td>
</tr>
<tr>
<td>7</td>
<td>0.0417</td>
<td>15.60</td>
</tr>
</tbody>
</table>

The conjoint analysis showed that measure 2 was considered the most important (17.8). In summary, excluding measures 2 and 4, the conjoint analysis showed that the measures were considered efficient in the following order: measure 6 > measure 1 > measure 7 > measure 3 > measure 5.

V. Conclusion

The Internet must respect each user's interest, but it must place public benefit as the first in priority. Nonetheless, the selfish motives of internet users are encouraging the increase of malicious replies on internet boards, resulting in the decrease in the public benefit of the Internet. As individual users will be interested in the increase of hits, it is highly probable that they will choose to post malicious replies. On the whole, the users' decisions will not get the optimal results.

In this study, this social situation was analyzed through a game model, and long-term solutions were proposed. In the long term, it is more beneficial to reduce the anonymity of internet boards or diversify the ways to induce hits other than the content of replies, although this could decrease the short-term number of hits and number of users.

Nonetheless, the above theory, evaluated by survey, still requires a gradual verification in the real internet environment before the actual implementation of the proposed countermeasures to malicious replies.

APPENDIX

A. Conjoint Analysis Results (SPSS 12.0)
ACKNOWLEDGMENT

Park Yong Sung, teacher of Korean Minjok Leadership Academy Department of Mathematics, has been a great help in writing the thesis. Also, the author thanks his father, who has given him more than just advice, and his mother, who has stayed up until late in night to give him a dish of oranges, and his brother, who has always given him a good laugh.

REFERENCES