A game-theoretic analysis of preventing spam over Internet Telephony via audio CAPTCHA-based authentication

Yannis Soupionis[1,3], Remous-Aris Koutsiamanis[2],

Pavlos Efraimidis[2], Dimitris Gritzalis[1]

1. Information Security and Critical Infrastructure Protection Research Laboratory

Dept. of Informatics, Athens University of Economics & Business (AUEB)

76 Patission Ave., Athens, GR-10434, Greece

{jsoup,dgrit}@aueb.gr

2. Dept. of Electrical and Computer Engineering, Democritus University of Thrace

University Campus, Xanthi, GR-67100, Greece

{akoutsia,pefraimi}@ee.duth.gr

3. Joint Research Center, European Commission

Enrico Fermi 2749, Ispra Varese, 21027, Italy

{yannis.soupionis}@jrc.ec.europa.eu

Corresponding Author: Dimitris Gritzalis, 76 Patission Ave., Athens, GR-10434

Greece, tel.: (+30) 210 8203505, e-mail: dgrit@aueb.gr

Abstract

Spam over Internet Telephony (SPIT) is a potential source of disruption in Voice over IP (VoIP) systems. The use of anti-SPIT mechanisms, such as filters and audio CAPTCHA (Completely Automated Public Turing Test to Tell Computer and Humans Apart) can prevent

unsolicited calls and lead to less unwanted traffic. In this paper, we present a game-theoretic model, in which the game is played between SPIT senders and internet telephony users. The game includes call filters and audio CAPTCHA, so as to classify incoming calls as legitimate or malicious. We show how the resulting model can be used to decide upon the trade-offs present in this problem and help us predict the SPIT sender's behavior. We also highlight the advantages in terms of SPIT call reduction of merely introducing CAPTCHA, and provide experimental verification of our results.

Keywords: Spam Prevention, Spam over Internet Telephony (SPIT), Game Theory, Audio CAPTCHA, Nash Equilibria.

1 Introduction

The explosive growth of the Internet has introduced a wide array of new technological advances and more sophisticated end-user services. One of them is VoIP, which is a developing technology that promises a low-cost, high-quality and availability service of multimedia data transmission. Inevitably though, VoIP "inherited" not only these positive features of Internet services, but also some of their problems [8][20][21][46]. One of them is Spam over Internet Telephony (SPIT) [36][37], which is the expression of Spam in VoIP network environments. SPIT is a challenging issue that IP telephony is expected to be facing in the near future. This is the reason why a) major organizations have already started developing mechanisms to tackle SPIT [13][35], and b) the U.S. Federal Communications Commission has extended the Telephone Consumer Protection Act of 1991 to include automated calls, called robocalls [9]. Moreover, it should be stated that the U.S. Federal Trade Commission has created the "Do Not Call Registry" in order to allow users to reduce the number of telemarketing sales calls received (automated or not) [10]. The active registrations in the "Do Not

Call Registry" were over 217 million on October 30th, 2012 [11].

The SPIT threat for VoIP is the analogue of spam for e-mail. However, due to its characteristics, it may also give the opportunity to malicious users to not only send low- or zero-cost unsolicited instant messages but also to make low- or zero-cost unsolicited calls by using automated software (bots). The malicious user's main purpose could be financial, like presenting advertisements, or to extract/steal a legitimate user's personal information (phishing). A real-life example is the "Rachel" robocall enforcement case, where five companies were shut down, because they made millions of illegal pre-recorded robocalls claiming to be from "Rachel" and "Cardholder Services" while pitching credit card interest rate reduction services [12]. Although the similarity of the SPIT phenomenon to the well-established spam threat is easy to identify, this does not lead to the conclusion that the techniques handling spam are appropriate for handling SPIT as well. While applying the anti-spam techniques can be done quite easily in terms of service configuration, some characteristics of SPIT make the direct application of anti-spam techniques inefficient and ineffective. In particular, telephony and instant messaging services operate in real time while email services are based on a "store and forward" model [18][40]. Therefore, the anti-spam techniques can examine the content/body of the email in order to classify it as spam or not, but this is not possible for VoIP real-time communication services [27].

A serious obstacle when trying to prevent SPIT is identifying VoIP communications which originate from software robots ("bots") in real-time. A typical way to tackle these attacks is the use of a Reverse Turing Test, called CAPTCHA (Completely Automated Public Turing Test to Tell Computer and Humans Apart). Since visual CAPTCHA are hard to apply in VoIP systems, audio CAPTCHA appear to be appropriate for defending against SPIT calls/messages

[15][42][43].

VoIP is a useful technology with significant value for legitimate users, as it enables communication and decreases costs. On the other hand, VoIP spammers can obtain significant financial revenues as the email spam paradigm has shown. Therefore, we have a situation where independent decision makers are engaged in a strategic interaction; the actions taken by SPIT senders may influence the defensive actions taken by the VoIP users and the opposite. The outcome of such scenarios is not only a matter of effective tools like audio CAPTCHA challenges, but also of how independent selfish decision makers will act and react in the presence of such tools. Such settings, where two or more independent decision makers interact, can be studied with concepts and tools from Game Theory. The equilibrium points of the respective game-theoretic model can reveal important attributes of the state(s), in which the system is expected to operate. For example, it will reveal how often the audio CAPTCHA will be used or whether the overall rate of SPIT calls decreases in the presence of audio CAPTCHA. In the presence of selfish users, there are examples where the introduction - always with good intentions - of a tool or an extra option for the users may lead to worse overall system performance. This can happen even with the simple addition of a new tool to an existing system. For example, in [17] scenarios are identified where increasing the number of (selfish) security experts of an information network may lead to reduced overall security of the network; the Braess paradox [4] shows how adding an extra route to a traffic network may lead to worse conditions for selfish drivers.

In this paper, we assume the existence of effective audio CAPTCHA challenges and discuss how the strategic interaction between SPIT senders and VoIP users can be modelled as a two-player game in the presence of such CAPTCHAs. In particular, we propose a game-theoretic model and show how the resulting

model can be used to predict the behavior that the two opponent communities will eventually adopt, how it can guide to fewer SPIT messages and how the use of CAPTCHA assists VoIP users against SPIT. As part of the legitimate user defences against SPIT we also integrated an anti-SPIT filter, which classifies each incoming call/message as legitimate, malicious or "unknown" (when it is not possible to have a confirmed answer). After the filter's incoming call classification, the user may directly accept or reject the call or request a CAPTCHA, depending on the precision and the verdict of the filter.

The rest of the paper is organized as follows: first, in Section 2, we illustrate the related work on relevant game-theoretic models and the cost calculation of spam/SPIT. In Section 3, the game theoretic model is introduced and its parameters and assumptions are presented. In Section 4 we present predictions about the SPIT calls/messages percentage of overall calls by computing the Nash equilibria of the game. In Section 5 we present the results of the experimental verification of the theoretical results. Finally, in Section 6, we end with a number of conclusions and our plans for future work.

2 Related Work

As the SPIT phenomenon is practically still in its infancy, we were not able to find relevant research work focusing on the cost of spam for both the SPIT sender and the user, or on relevant game-theoretical models. Therefore, we present research work based on a close relative of SPIT, i.e., e-mail spam.

2.1 Cost of unsolicited communication

Kim Y., et al. [22] propose a method to measure the disutility experienced by e-mail users who receive spam. Their study employs conjoint analysis of stated preference data to estimate e-mail users' overall inconvenience cost attributable

to spam. The results show the inconvenience-originating cost of spam to be about $0.0026 per spam message.

Kanich C., et al. [19] present a methodology for measuring the conversion rate of spam. They produced nearly half a billion spam e-mails and they identified the number that were successfully delivered, the number that passed through popular anti-spam filters, the number that elicited user visits to the advertised sites, and the number of "sales" and "infections" produced. They managed to calculate that the total revenue of a spam campaign is about $7000 and the cost to produce it is the paycheck of three "good" programmers. Therefore the cost per message is about $0.001. Finally, a report placed the retail price of spam delivery at slightly under $80 per million [47]. This price means that each spam email costs $0.00008, but we stick to the previous paper's cost estimates, as this kind of price is an order of magnitude less than what legitimate commercial mailers charge.

2.2 Game-theoretic models

Androutsopoulos I., et al. [1] present an interesting game-theoretic model for the interaction of spam and ordinary e-mail users and later extend their model in [45] to the case where the users are able to use Human Interaction Proofs (HIP). In the latter work, they focused on the scenario where the users can read messages, delete them without reading them or send HIP. They have provided an extensive theoretical analysis of a game-theoretic model for the problem of spam. As discussed earlier, there are important qualitative differences between SPIT and spam. We generalize the model proposed in [45] to a more complicated problem with more actions to account for additional situations that arise in VoIP, and apply it within a related, but substantially different, application context, namely VoIP. We also experimentally confirm the predictions of the

model.

Parameswaran M. [33] suggests that the spammer can strategize to maximize the amount of spam sent by making inferences from the block-list rules. They introduce a theoretical modeling approach for the spammer's behavior and present a comparison of this behavior with the data that has been collected from block-list organizations. The main issue with this work is that is based on collected data, therefore its outcomes cannot be generalized. Shahroudi A.B. et al. [38] examine how VoIP service providers attempt to control the growing phenomenon of SPIT by creating a game-theoretic model of competition between providers. The model is based on the notion that two different service providers, which try to maximize their profit with different business strategies, are competing on shared resources. Each service provider can select to either detect or prevent SPIT in order to address attacks, with consequences to the overall profit of both providers. The research outcome is that the providers are going to focus on mechanisms which detect SPIT attacks, because even though they are more expensive than preventative mechanisms, it maximizes their profit.

Moreover, a discussion of game theory approaches for detection software can be found in [6]. The proposed model is able to assist firms in the configuration process of detection software and a significant outcome is that false-positive and false-negative errors in detection could affect the value of these systems significantly.

In general, even though there is work on applying game-theoretic tools to problems of security, to the best our knowledge this is the first attempt of a game-theoretic analysis of SPIT and how to counter it with audio CAPTCHA.

3 Suggested game-theoretic model

Generalizing and building upon Androutsopoulos et al. [1][45], we define the SpitGame, a game-theoretic model with two players: the SPIT sender (Player I) and the legitimate VoIP user (Player II). The game is illustrated in Fig. 1. We will describe the game in detail and at the same time give short definitions of the game-theoretic terms and concepts that we encounter. For more details on the game-theoretic terms, the reader may refer to textbooks on Game Theory [29][30][31], or to a recent volume on Algorithmic Game Theory [28].

The SpitGame, as shown in Fig. 1, is an extensive form game with imperfect information. The game is initiated whenever a new call/message is sent towards a user. The SPIT sender (Player I) moves first and is able to interfere with the stream of incoming calls and send a new SPIT call at any point. Thus, the frequency with which SPIT senders initiate a malicious call determines the average ratio of SPIT to legitimate calls in the users' incoming streams. For example, if a SPIT sender initiates a SPIT call every four (4) legitimate calls, then the overall probability/rate of SPIT calls will be $p = 0.2$, which is presented as probability p in Fig. 1. Although in reality SPIT senders are not able to completely control all the incoming calls/messages, or to decide whether or not they will insert a new SPIT message/call, the assumption that the SPIT senders control the ratio between SPIT and legitimate calls is reasonable. A similar assumption has been used in the game-theoretic models for SPAM in [1][45] upon which we generalised.

The SPIT sender chooses to make the incoming call SPIT or to allow it to be a legitimate call. The VoIP user does not learn which choice the SPIT sender has made. That is, the VoIP user is imperfectly informed about the game status and for this reason we model this interaction as an extensive game with imperfect information. However, the VoIP user gets some stochastic information

about the game status from the outcome of an anti-SPIT filter. After the move of the SPIT sender, the call is processed by anti-SPIT filters, which are able to flag the calls they consider SPIT. The use of filters is a common countermeasure (in some cases of Internet service providers, this is mandatorily applied to their users). We have assumed that the filter contains a deterministic first stage and a stochastic second stage. In the first stage, an accurate black/white-list, created from past calls, can accept or discard the call. The second stage is invoked if the black/white-list does not identify the caller. In this stage, the filter attempts to guess the nature of the call from the characteristics of the call (e.g. the time/date, the caller domain, the user agent, etc.). In the model we describe, the filter refers only to the second stage, since the first stage does not have a game theoretic aspect.

In our model, the performance of these filters is fully described by six variables: f_l, h_2, h_1, ϵ_1, ϵ_2 and f_s. More specifically, in the case of legitimate calls, the filter will classify the calls accurately with a probability of f_l, it will consider them unknown with a probability of h_2 and it will misclassify them as SPIT calls with a probability of h_1. In the case of SPIT calls the corresponding legitimate, unknown and SPIT classification probabilities are ϵ_1, ϵ_2 and f_s. For example, consider the case when the filter misclassifies the incoming message. In Fig. 1 the probability of misclassifying a SPIT call as legitimate is depicted as $S \rightarrow L$ and the probability of misclassifying a legitimate call as SPIT is depicted as $L \rightarrow S$. Moreover, the filters may not be able to come to a definite conclusion over the nature of the call. In this situation, the filter classifies the call as "*Unknown*", which is common in VoIP communication systems. Although this may be uncommon in email spam filters, since the messages can be classified based on content and header, VoIP is a real time protocol that does not grant the receiver access to the call contents prior to its acceptance/session

establishment. Therefore, whenever a call arrives from an unknown number, the call may be classified as SPIT or legitimate. Since VoIP communication is synchronous, unlike email spam where email is delivered asynchronously and the marked-as-spam messages can be stored, if the call is rejected then there is no way for the user to retrieve its content/purpose. Since much less information is available than in email spam, the anti-SPIT filter should include the "*Unknown*" verdict, which is dominant when a SPIT call is received, since most SPIT calls are initiated from numbers unknown to the user.

In the context of SpitGame, after the move of the SPIT sender the filter classifies the incoming call. The action of the filter is modelled with an artificial third player; such a player is usually called chance in the game. Player chance has three moves, one for each of the possible outcomes of the filter.

The user is informed about the "move" of the filter but not the move of the SPIT sender. The user should decide his move based on the filter's prior classification. He is able to accept the call, reject it or request an audio CAPTCHA. The user is not aware of the true nature of the calls before he listen to them, so when he sees that his filter has classified a call as legitimate, he does not know whether it was misclassified or not. For example, when a user receives a legitimate filter-classified call it is impossible to distinguish in which node ($L \rightarrow L$ or $S \rightarrow L$) of the game he is. In game-theoretic terms, each of the possible outcomes of the filter defines an information set for the VoIP user. Each such set contains two nodes of the extensive-form game, because there are two nodes in the game which may lead to the particular filter decision. The VoIP user, however, is informed only about the information set and not about the particular node of the set in which the game really is.

Therefore, each user has to select a strategy consisting of what he will do with incoming calls depending only on information sets, i.e., the decisions of his filter;

for example, *Accept* calls classified as *Legitimate*, *Reject* calls classified as *SPIT*, and request audio *CAPTCHA* when calls are classified as *Unknown*. Similarly, we may assume that the overall community of users adopts a strategy, whose probabilities reflect the frequencies with which it adopts actions *Accept*, *Reject*, and *CAPTCHA*. That means that the sum of the probabilities of these three actions is equal to 1 for each game node. For example, when a user receives a new call, which is classified as *Legitimate*, then $P(Accept) + P(Reject) + P(CAPTCHA) = 1$, regardless of whether the message was misclassified or not. Likewise, this happens in the other two cases: *SPIT* and *Unknown*.

Whenever a new session is initiated, the actions which the SPIT sender and legitimate user select lead to a particular cost or utility for each player. For example, if the SPIT sender selects to initiate a *SPIT* call and the user selects to *Accept* the call, then the game ends with a utility of $s_a > 0$ for the SPIT sender and a cost of $-u_s < 0$ for user. In summary, every combination of actions of the two players leads to an outcome of the game, and this outcome determines the amount of utility for each participant, which is shown in Fig. 1 and Table 1. Notice that the utilities for the user and SPIT sender do not depend directly on the filter classification, however, the classification does affect the ratio between legitimate and SPIT calls which the user receives.

The utilities for each player are determined by five parameters:

1. u_l: This is the measure of average utility of accepting a legitimate call.

2. u_s: This is the measure of average disutility of receiving a SPIT call, taking into consideration factors such as the average cost of consumed computational resources, the time needed to answer the phone, and the average time it takes to listen to it, which means a general decrease to user productivity.

3. u_c: This is the measure of average disutility of sending a CAPTCHA

puzzle, taking into consideration the annoyance of a legitimate caller, of whom it is required to solve a CAPTCHA challenge in order to reach the user. This annoyance can directly lead to profit loss if the caller is a potential customer, but also indirectly lead to social issues if the user's acquaintances are reluctant or hesitant to call him.

4. s_a: This is the measure of average utility the SPIT sender obtains from each SPIT call that is accepted, taking into consideration factors such as the percentage of users that order products after listening to the SPIT call, and the advertisement campaigns he may be paid to be part of.

5. s_r: This is the measure of average disutility to the SPIT sender of getting a SPIT rejected, taking into consideration all related costs, including the computational resources to create SPIT, and the effort to create an appropriate bot to execute SPIT attacks.

The parameters express a measure (or absolute value) of utility or disutility; as such $u_l, u_s, u_c, s_a, s_r > 0$ and when appearing in pay-offs their sign denotes whether they express utility (+) or disutility (-).

We assumed that the utility from accepting a legitimate call is exactly the opposite of the cost of rejecting it. This is justified by equating the (dis)utility of the user to the information value of the call being (rejected) accepted. Moreover, the utility of accepting the call may be the information value of the call minus the cost of the consumed computational resources for session establishment, while the cost of rejecting it may be simply the information value. This cost difference is so marginal that it was not taken into consideration.

In order to facilitate the examination and analysis of the model, we have set a few restrictions on the costs:

1. The user's disutility for sending a CAPTCHA ($-u_c$) is smaller than the

user's disutility for missing a legitimate message ($-u_l$). In absolute terms, $u_l > u_c$. This means that when a user initiates a call, the process to answer a CAPTCHA for establishing the call is not cost-forbidden.

2. The user's disutility for sending a CAPTCHA ($-u_c$) is smaller than the user's disutility of accepting a SPIT call ($-u_s$). In absolute terms, $u_s > u_c$. Otherwise, the use of CAPTCHA would have no sense, since it would be better for the user to receive SPIT than request a CAPTCHA.

3. The user's disutility of accepting a SPIT call ($-u_s$) is smaller than the user's disutility for missing a legitimate message ($-u_l$). In absolute terms, $u_l > u_s$. This condition is based on the premise that receiving a SPIT call may be annoying and distracting for the callee, but missing a legitimate call is more important since it may mean loss of business opportunities, damage to a business' image and reputation or disruption of the user's social life.

4. The utility for a SPIT sender to have a SPIT call accepted (s_a) is larger than the cost of having the call rejected ($-s_r$). In absolute terms, $s_a > s_r$. Given that in practice the chance of the SPIT sender making a profit from an accepted call is very low and that the cost of making SPIT calls, due to the way VoIP works, is also very low, it can reasonably be assumed that the utility of having a call accepted needs to be high, at least higher that the disutility of making the call, in order for the SPIT sender to have an incentive to make calls. In general, SPIT calls could be profitable even if $s_a < s_r$, if the chance of making a profit from an accepted call could be assumed to be high enough.

The above mentioned utilities for each player actions and the relevant conditions are described in Table 2.

4 Game-theoretic analysis and Nash equilibrium

In this section, we present a theoretical analysis of the SpitGame. The fundamental solution concept for games is the Nash equilibrium (NE), i.e., a state of the game from which no individual player has an incentive to unilaterally deviate. The Nash equilibrium is the most popular solution concept in game theory and has been used in the analysis of a vast number of scenarios with interacting decision makers coming (the scenarios) from diverse application domains including economics, biology, political science, computer science and other ([26][28][29][30]). There are numerous applications of game theory, the Nash equilibrium concept and its refinements in Computer Security. See for example the recent surveys [24][44] and the references therein.

Overall, the formulation of the Nash equilibrium has had a fundamental and pervasive impact in economics and the social sciences [26] and more recently in Computer Science [28][32]. Of course, from the development of the Nash equilibrium concept, there have also been some critiques of it. Some of the main critiques are that the Nash equilibrium concept makes misleading or ambiguous predictions in certain circumstances, that it may not capture correctly non-credible threats, that in many games there are many NE, and, more recently, that the computation of NE is intractable in the general case [7].

However, despite these critiques, the NE and its refinements are undoubtedly the most successful solution concept in game theory, widely used in theoretical and practical applications of game theory. Moreover, most critiques do not seem to apply to the NE of the SpitGame. Firstly, the SpitGame exhibits a unique NE (except for some boundary cases) as is shown in Theorem 2. Consequently, there is no ambiguity in the prediction of the state of the game. Moreover, the NE of the SpitGame is computable in polynomial time via a closed form equation (see Table 8) and thus, neither the critique concerning the intractability

of general NE applies in this case. As discussed later in this section, the NE of the SpitGame is also Subgame Perfect, which removes the non-credible threat issue of some NE. Finally, the NE solution of the SpitGame does not seem to belong to the cases where the NE leads to counter-intuitive solutions, like for example in the case of the Traveller's Dilemma [3].

There are adaptations and refinements of the NE concept for different game settings and purposes. A variation of the NE for extensive-form games is the Subgame Perfect Equilibrium (SPE), which is more appropriate for games with perfect information. In the SpitGame, when Player II has to decide his action without seeing the action of Player I, that is, Player II is imperfectly informed about the game status. However, Player II has access to the outcome of the filter, which provides stochastic information about the action of Player I. The filter verdicts are shown in Table 3. Each of the filter verdicts defines an information set for Player II, who has to decide his action based on the information set. A natural approach for analysing such a model is to use the concept of behavioral strategies ([26][6], and in particular [45][1]), in which players can randomize independently at each information set. In particular, Player II of the SpitGame will have an independent mixed strategy for each of his information sets. A well known fact in game theory, Kuhn's Theorem, states that in extensive-form games with perfect recall, behavioral and mixed strategies are equivalent. The solution concept that we will use to solve the SpitGame is the Nash equilibrium of the corresponding extensive form game, and we will base our analysis on the behavioral strategies of the players.

The interaction between legitimate VoIP users and SPIT senders is a continuous challenge for both parties. Each player, call receiver or SPIT sender, will have to make his choices repeatedly. Moreover, a legitimate caller might be required to solve audio CAPTCHAs when he calls a VoIP user for the first time.

Such overheads may devalue the VoIP service in the eyes of legitimate callers. One may argue that a repeated game could be used to model this interaction. Even though one cannot (and should not) exclude such or other possible formulations of the SpitGame problem, we believe that the current formulation as a one-shot game is well suited for the problem. Each time there is an interaction between two entities, the interaction will be unique, or at least we are only interested in the unique interactions. The subsequent interactions between the same entities can be trivially solved by the outcome of the first game. Then, the legitimate player would know if the call is SPIT or not. The cost incurred to the legitimate callers for solving audio CAPTCHAS is assumed to be captured by the disutility u_c. Note that legitimate callers are not directly modelled in the current SpitGame model. Alternatively, one may consider other game-theoretic formulations of the same problem, for example as a repeated game and/or a game with strategic legitimate callers being part of the model. We leave such possibilities for future work.

We will start the analysis of the SpitGame with the following straightforward observation that Player I will never use a pure strategy at any Nash equilibrium.

Theorem 1. *The SpitGame has no Nash equilibrium where Player I plays a pure strategy.*

Proof. We will use a proof by contradiction. Assume that Player I chooses a pure strategy, for example $SPIT$. Then the optimal response for Player II would be the pure strategy *Reject*. Then however, Player I would be motivated to change his strategy, i.e., there is no NE if Player I plays $SPIT$. If, on the other hand, Player I chooses the pure strategy *Legitimate* then Player II can respond with *Accept*, which makes the move of Player I suboptimal, i.e., again no NE. □

Assume a NE of the SpitGame. Let $(p, 1-p)$ be the strategy of Player I at

the NE (Table 4) and let (p_i, q_i, r_i) be the strategy of Player II at information set i, for $i = 1, 2, 3$. Thus, at the NE, the strategy of Player I is to submit SPIT calls with rate p, i.e., the probability that a new incoming call will be SPIT is p. From the proof of Theorem 1 we know that at any NE

$$0 < p < 1 \; . \tag{1}$$

Player II has three information sets, one for each of the outcomes of the filter, presented in Table 3.

Since Player II does not know which action Player I has made and the outcome of the filter is stochastic, Player II can base his decision only on conditional probabilities. Assume that a new call has arrived and that the corresponding filter verdict is *SPIT*. Player II is informed that the information set is *SPIT* and has to choose a strategy based on this information only. Let P_{LS} be the conditional probability that the incoming call is *Legitimate* given that the filter has classified it as *SPIT*. Using standard probability theory gives

$$P_{LS} = Prob[L/S] = \frac{(1-p)\; h_1}{(1-p)\; h_1 \; + \; p\; f_s} \; . \tag{2}$$

Similarly, we define and calculate the conditional probabilities for all possible cases.

$$\begin{aligned} P_{LS} &= \frac{(1-p)\; h_1}{(1-p)\; h_1 \; + \; p\; f_s}, \quad P_{SS} = \frac{p\; f_s}{(1-p)\; h_1 \; + \; p\; f_s}, \\ P_{LU} &= \frac{(1-p)\; h_2}{(1-p)\; h_2 \; + \; p\; \epsilon_2}, \quad P_{SU} = \frac{p\; \epsilon_2}{(1-p)\; h_2 \; + \; p\; \epsilon_2}, \\ P_{LL} &= \frac{(1-p)\; f_l}{(1-p)\; f_l \; + \; p\; \epsilon_1}, \quad P_{SL} = \frac{p\; \epsilon_1}{(1-p)\; f_l \; + \; p\; \epsilon_1}. \end{aligned} \tag{3}$$

Using the above conditional probabilities of Equation 3 and the SpitGame model as it is depicted in Fig. 1, the average utility of Player I for each of his pure strategies can be calculated. Firstly, note that the average utility for the

pure strategy of Player I *Legitimate*, i.e., Player I does nothing, is

$$U_{1L} = 0 \ . \tag{4}$$

When Player I submits a SPIT call, then his average utility can be calculated as follows. Given the strategy p of Player I, let $V_L(p)$, $V_U(p)$, and $V_S(p)$ be the probabilities that the filter verdict is *Legitimate*, *Unknown*, and *SPIT* respectively. Also, given the strategy of Player II, let $U_{1SL}(p_1, q_1)$, $U_{1SU}(p_2, q_2)$, and $U_{1SS}(p_3, q_3)$ be the average utility of action *SPIT* of Player I in information set *Legitimate*, *Unknown*, and *SPIT* respectively. Then the average utility of action *SPIT* of Player I is

$$U_{1S} = V_L(p)\ U_{1SL}(p_1, q_1) + V_U(p)\ U_{1SU}(p_2, q_2) + V_S(p)\ U_{1SS}(p_3, q_3), \tag{5}$$

where

$$\begin{aligned} V_L(p) &= (1-p)f_l + p\epsilon_1, \\ V_U(p) &= (1-p)h_2 + p\epsilon_2, \text{ and} \\ V_S(p) &= (1-p)h_1 + pf_s. \end{aligned} \tag{6}$$

After expanding the terms in Equation 5 and doing some algebraic manipulation we obtain that

$$U_{1S} = -s_r + \epsilon_1(s_a + s_r)p_1 + \epsilon_2(s_a + s_r)p_2 + f_s(s_a + s_r)p_3 \ . \tag{7}$$

From Theorem 1 we know that Player I uses a mixed strategy at any NE. Thus, both actions of Player I are played with strictly positive probability at any NE; in other words, both actions of Player I belong to the support of his strategy at any NE. A well known requirement for all actions that belong to the support of a NE strategy, is that each of them must achieve the same average utility.

Otherwise, the user would exclude the strategies with lower average utility from his NE strategy. We know from Equation 4 that U_{1L}, i.e., the (average) utility of action *Legitimate* for Player I, is zero. Thus, the average utility of action *SPIT* of Player I must also be

$$U_{1S} = 0\ . \tag{8}$$

Combining the above equation with Equation 7 gives the following Lemma.

Lemma 1. *At any NE of the SpitGame*

$$\epsilon_1 p_1 + \epsilon_2 p_2 + f_s p_3 = \frac{s_r}{s_a + s_r}\ . \tag{9}$$

We now focus on the utility of Player II. Using again the conditional probabilities of Equation 3 and the SpitGame model (Fig. 1), the average utility of Player II for each of his pure strategies at each of his information sets can be calculated. For example, in information set *Legitimate*, the average utility for Player II for action *Accept* of an incoming call is

$$U_{2LA} = P_{LL}u_l + P_{SL}(-u_s) \ = \frac{f_l(1-p)u_l - \epsilon_1 p u_s}{f_l(1-p) + \epsilon_1 p}. \tag{10}$$

Similarly, we can calculate the expected utilities U_{2LC} and U_{2LR} for actions *CAPTCHA* and *Reject*. In the same way, we calculate U_{2UA}, U_{2UC}, and U_{2UR} for the information set *Unknown*, and U_{2SA}, U_{2SC}, and U_{2SR} for the information set *SPIT*.

$$\begin{array}{lll} U_{2LA} = \frac{f_l(1-p)u_l - \epsilon_1 p u_s}{f_l(1-p)+\epsilon_1 p}, & U_{2LC} = \frac{f_l(1-p)(u_l-u_c)}{f_l(1-p)+\epsilon_1 p}, & U_{2LR} = \frac{-f_l(1-p)u_l}{f_l(1-p)+\epsilon_1 p}, \\ U_{2UA} = \frac{h_2(1-p)u_l - \epsilon_2 p u_s}{h_2(1-p)+\epsilon_2 p}, & U_{2UC} = \frac{h_2(1-p)(u_l-u_c)}{h_2(1-p)+\epsilon_2 p}, & U_{2UR} = \frac{-h_2(1-p)u_l}{h_2(1-p)+\epsilon_2 p}, \\ U_{2SA} = \frac{h_1(1-p)u_l - f_s p u_s}{h_1(1-p)+f_s p}, & U_{2SC} = \frac{h_1(1-p)(u_l-u_c)}{h_1(1-p)+f_s p}, & U_{2SR} = \frac{-h_1(1-p)u_l}{h_1(1-p)+f_s p}. \end{array} \tag{11}$$

Now, using the notation of Tables 4 and 5 for the player strategies, and the average utility for each of the pure strategies of Player II (Equation 11) the average utility of Player II for each information set can be calculated. For example, information set *Legitimate*, the average utility of Player II is

$$U_{2L} = p_1 U_{2LA} + q_1 U_{2LC} + r_1 U_{2LR} \, . \tag{12}$$

Similarly, for information sets *Unknown* and *SPIT* the average utility of Player II is

$$U_{2U} = p_2 U_{2UA} + q_2 U_{2UC} + r_2 U_{2UR} \tag{13}$$

and

$$U_{2S} = p_3 U_{2SA} + q_3 U_{2SC} + r_3 U_{2SR} \, , \tag{14}$$

respectively. Expanding Equations 12, 13, and 14, with the expressions of Equation 11 gives a closed expression for the average utility of Player II at each information set $i = 1, 2, 3$. After some algebraic manipulation, and exploiting the symmetry in the expressions for the three information sets, we obtain that the average utility of Player II in each information set is

$$\frac{A_i p_i + B_i q_i + C_i}{D_i}, \quad \text{for } i = 1, 2, 3. \tag{15}$$

where the coefficients A_i, B_i, C_i and D_i are as defined in Table 6. Note that the coefficients D_i correspond to the probabilities of each information set, as they are defined in Equation 6. The coefficients A_i, B_i, C_i and D_i are functions of the strategy p of Player I and other variables. We focus on p and identify the boundary values c_i and d_i for $i = 1, 2, 3$, presented in Table 7.

4.1 The Nash Equilibrium

We are now ready to determine the NE of the SpitGame. Our analysis will be valid for a wide range of parameter values. The main assumption we make is that

$$\epsilon_1 < \epsilon_2 \ . \tag{16}$$

This is a reasonable assumption which also holds for the empirical parameter values we use in the experiments (Table 3). A further plausible assumption is that the probability that the filter verdict is correct is larger than the probability that the verdict is completely wrong. More precisely,

$$h_1 < f_l \ , \text{ and} \tag{17}$$

$$\epsilon_1 < f_s \ . \tag{18}$$

Additionally, we assume that

$$h_2 < f_l \ . \tag{19}$$

The final assumption, which is also a plausible one, states that

$$u_c < 2 \ u_l \ , \tag{20}$$

that is, the cost for Player I to submit an audio CAPTCHA is less than twice the utility of accepting a legitimate call. Note that a cost u_c larger than 2 u_l would make the use of audio CAPTCHAs pointless. The cost of applying an audio CAPTCHA should actually be much lower than 2 u_l.

At any NE equilibrium, the strategy of Player II, i.e., the values of p_i and q_i, must be such that the values of U_{2L}, U_{2U} and U_{2S} are maximized, for the the given strategy p of Player I. An immediate consequence is that if the some

coefficients A_i or B_i are strictly negative then the corresponding p_i or q_i will have to be null at the NE.

For each i, we will compare the coefficients of each pair of p_i and q_i. We will also compare the coefficients of all p_i with each other. In Table 7 the boundary values of p to satisfy specific equations on the coefficients A_i and B_i are given. For $p = c_1$, the coefficient of p_1 in Equation 15 for $i = 1$ becomes $A_1 = 0$. Note, that if $p > c_1$ then $A_1 < 0$, and if $p < c_1$, then $A_1 > 0$. Similarly, if $p = d_1$ then $A_1 = B_1$, if $p > d_1$, then $A_1 < B_1$, and if $p < d_1$, then $A_1 > B_1$. Similar statements hold for coefficients A_2, B_2, A_3 and B_3.

Some observations about the relations between the boundary values of p are in order. Using Equations 16, 17, 18, and 19 we obtain that

$$c_1 > c_2 \text{ and } c_1 > c_3 \ . \tag{21}$$

Similarly, we obtain

$$d_1 > d_2 \text{ and } d_1 > d_3 \ . \tag{22}$$

Using Equation 20 we immediately obtain that

$$d_i < c_i, \text{ for } i = 1, 2, 3. \tag{23}$$

and

$$B_i > 0, \text{ for } i = 1, 2, 3. \tag{24}$$

We can also make some observations about the strategy of Player II. An immediate consequence of Equation 9 is that

$$p_1 + p_2 + p_3 > 0 \ . \tag{25}$$

Thus, at least one of the A_i must be ≥ 0. This, in turn, implies that $p \leq \min\{c1, c2, c3\} = c_1$. Moreover, from Equation 24 we know that all coefficients B_i are strictly positive. This implies that

$$p_i + q_i = 1 \text{ for } i = 1, 2, 3. \tag{26}$$

In other words, the action *Reject* is not used by Player II at any NE. A careful look at the SpitGame in Fig. 1 reveals that action *Reject* of Player II is weakly dominated by his action *CAPTCHA*. This means, that the utility of action *Reject* is less than or equal and in some cases strictly less than the utility of action *CAPTCHA*. However, this observation alone would not be sufficient to exclude action *Reject* from NE strategies. There are well known examples of games having NE where players use also weakly dominated strategies.

Finally, let σ be

$$\sigma = s_r/(s_a + s_r)\ . \tag{27}$$

4.1.1 Case Analysis

We are now ready to obtain the NE of the SpitGame.

Case 1: $\epsilon_1 \geq \sigma$

Let us first consider the case $\epsilon_1 > \sigma$. From Equation 9 we obtain that $p_1 < 1$. Thus in Equation 12 we have $p_1 > 0$ and $q_1 > 0$. Recall, that values of p_1 and q_1 at a NE have to maximize the utility U_{2L}. The only way the expression U_{2L} is maximized for p_1 and q_1 both strictly positive is if $A_1 = B_1 \geq 0$. To have $A_1 = B_1$, it must hold that $p = d_1$. Moreover, the corresponding value of A_1 and B_1 for $p = d_1$ is strictly positive from Equation 24. Thus, the SpitGame has a single NE equilibrium at $p = d_1$. Moreover, for $p = d_1$, we have $A_2 < B_2$ and $A_3 < B_3$. Consequently, $p_2 = p_3 = 0$, and thus $q_2 = q_3 = 1$. Using Equation 7 we get $p_1 = \frac{\sigma}{\epsilon_1}$.

We will now obtain the same results for the case $\epsilon_1 = \sigma$. First we will show that $p_1 = 1$. Assume, $p_1 < 1$. Then $q_1 > 0 \Rightarrow A_1 = B_1$ and, thus $p = d_1$. Moreover, for $p = d_1$, we have $A_2 < B_2$ and $A_3 < B_3$. Consequently, $p_2 = p_3 = 0$. At the same time, using $p_1 < 1$ in Equation 9 gives $p_2 + p_3 > 0$, a contradiction with the previous result. Thus, in this case $p_1 = 1$. From $p_1 = 1$, we obtain $p_2 = p_3 = 0$, $q_1 = 0$, and $q_2 = q_3 = 1$.

Thus, for the case of $\epsilon_1 \geq \sigma$, the SpitGame has the following unique NE

p	p_1	q_1	p_2	q_2	p_3	q_3
d_1	$\frac{\sigma}{\epsilon_1}$	$1 - p_1$	0	1	0	1

Note that we do not show the values of the r_i for the SpitGame, since their value will always be zero, as discussed earlier.

Case 2: $\epsilon_1 < \sigma$

We have to further distinguish three sub-cases based on the relation of the ratios ϵ_2/h_2 and f_s/h_1.

Case 2.1: $\epsilon_2/h_2 < f_s/h_1$

The inequality $\epsilon_2/h_2 < f_s/h_1$ implies that

$$c_2 > c_3 \text{ and } d_2 > d_3 \; . \tag{28}$$

Case 2.1.1: $\epsilon_1 < \sigma < \epsilon_1 + \epsilon_2$

In this case, if p_1 would be $p_1 < 1$, then (as in the case $\epsilon_1 > \sigma$) we would have $p_2 = p_3 = 0$. However, then Equation 9 would be infeasible. Thus,

$$p_1 = 1 \; , \; q_1 = 0 \; . \tag{29}$$

If $\epsilon_1 = \sigma$, then from Equations 29 and 9, we again conclude that $p_2 = p_3 = 0$. If $\epsilon_1 > \sigma$, then for the same reason it must hold $p_2 + p_3 > 0$, that is, at least

one of p_2 and p_3 must be strictly positive (because else Equation 9 would be infeasible).

If $A_2 > B_2 \Rightarrow p_2 = 1$. This, however, makes Equation 9 on p_1, p_2 and p_3, infeasible. The case $A_2 < B_2$ is also not feasible, because then we would have $p_2 = 0$ and $q_2 = 1$, which would again make Equation 9 infeasible. Consequently, it must hold $A_2 = B_2$ and consequently $p = d_2$.

Thus, the NE for Case 2.1.1 is

p	p_1	q_1	p_2	q_2	p_3	q_3
d_2	1	0	$\frac{\sigma-\epsilon_1}{\epsilon_2}$	$1-p_2$	0	1

Case 2.1.2: $\epsilon_1 + \epsilon_2 \leq \sigma$.

Assume that $A_3 > B_3$. Then, $A_3 > B_3 \Rightarrow p < d_3 \Rightarrow p < d_2 \Rightarrow A_2 > B_2 \Rightarrow p_2 = p_3 = 1$. In this case the strategy of Player II would be always *Accept*, which is not a NE strategy (Player I would simply respond always with *SPIT*). Thus A_3 cannot be smaller than B_3. The case $A_3 < B_3$ is also not possible, because it would imply $p_3 = 0$, which in turn would make Equation 9 infeasible. From the above arguments, we conclude that

$$A_3 = B_3 \ . \tag{30}$$

Thus, in this case, $A_1 > B_1$, $A_2 > B_2$ and $A_3 = B_3$ and consequently

$$p = d_3 \ . \tag{31}$$

The overall NE is

p	p_1	q_1	p_2	q_2	p_3	q_3
d_3	1	0	1	0	$\frac{\sigma-\epsilon_1-\epsilon_2}{f_s}$	$1-p_3$

Case 2.2: $\epsilon_2/h_2 > f_s/h_1$.

The inequality $\epsilon_2/h_2 > f_s/h_1$ implies that

$$c_2 < c_3 \text{ and } d_2 < d_3 \ . \tag{32}$$

A simple adaptation of the analysis of the cases 2.1.1 and 2.1.2 gives the following results for cases 2.2.1 and 2.2.2, respectively.

Case 2.2.1: $\epsilon_1 < \sigma < \epsilon_1 + \epsilon_2$

In this case, $p = d_3$ and the overall NE is

p	p_1	q_1	p_2	q_2	p_3	q_3
d_3	1	0	0	1	$\frac{\sigma-\epsilon_1}{f_s}$	$1-p_3$

Case 2.2.2: $\epsilon_1 + \epsilon_2 \leq \sigma$.

In this case, $p = d_2$ and the overall NE is

p	p_1	q_1	p_2	q_2	p_3	q_3
d_2	1	0	$\frac{\sigma-\epsilon_1-f_s}{\epsilon_2}$	$1-p_2$	1	0

Case 2.3: $\epsilon_2/h_2 = f_s/h_1$.

In this case,

$$d_2 = d_3 \text{ and } c_2 = c_3 \ . \tag{33}$$

The case $p_1 < 1$ can easily be excluded, because it would imply $p_2 = p_3 = 0$, making Equation 9 infeasible. Thus, we conclude that $p_1 = 1$. From Equation 9 we obtain that $p_2 + p_3 > 0$. Any pair of values p_2 and p_3 satisfying $\epsilon_2 p_2 + f_s p_3 = \sigma - \epsilon_1$ gives a NE. In this case the SpitGame has the following continuous range of NE

p	p_1	q_1	p_2	q_2	p_3	q_3
d_2	1	0	p_2	$1-p_2$	$\frac{\sigma-\epsilon_1-\epsilon_2 p_2}{f_s}$	$1-p_3$

where $d_2 = d_3$ and the range of values for p_2 is

$$\max\{0, \frac{\sigma - \epsilon_1 - f_s}{\epsilon_2}\} \leq p_2 \leq \frac{\sigma - \epsilon_1 - f_s \max\{0, \frac{\sigma - \epsilon_1 - \epsilon_2}{f_s}\}}{\epsilon_2} \ . \tag{34}$$

From the above case analysis of the SpitGame we conclude that:

Theorem 2. *The SpitGame has a unique NE equilibrium for the assumptions made earlier except for the Case 2.3. The closed forms of the NE for each case are summarized in Table 8.*

4.2 The NE without audio CAPTCHAs

We examine now the NE of the SpitGame if users did not have the option to use audio CAPTCHAs. We can assume that the action CAPTCHA is removed from the game or equivalently that $u_c > 2\ u_l$. If $u_c > 2\ u_l$, then all coefficients B_i would be negative

$$B_i < 0 \text{ for } i = 1, 2, 3. \tag{35}$$

and consequently the probability of submitting an audio CAPTCHA would be $q_i = 0$, for all infomation sets. We will call the model without audio CAPTCHAs SpitGame$'$.

From Equations 25 and 21 we obtain that in the SpitGame$'$ the strategy of Player I satisfies $p \leq \max\{c_1, c_2, c_3\} = c_1$.

Case 1: $\epsilon_1 \geq \sigma$

Let us first consider the case $\epsilon_1 > \sigma$. From Equation 9 we obtain that $p_1 < 1$. Thus in Equation 12 we have $p_1 > 0$ and $q_1 > 0$. Recall, that the values of p_1 and q_1 at any NE have to maximize the utility U_{2L}. Given that $B_1 < 0$, the only way the expression U_{2L} is maximized for p_1 and q_1 both strictly positive is if $A_1 = 0$. This requires that $p = c_1$.

Since $p = c_1$ implies $A_2 < 0$ and $A_3 < 0$, we obtain that $p_2 = p_3 = 0$ (and thus $r_2 = r_3 = 1$). Using this in Equation 9 we obtain that $p_1 = \frac{\sigma}{\epsilon_1}$.

Thus, the SpitGame′ has the following unique NE

p	p_1	r_1	p_2	r_2	p_3	r_3
c_1	$\frac{\sigma}{\epsilon_1}$	$1-p_1$	0	1	0	1

There is an evident analogy with the corresponding NE of the original SpitGame. The strategy of Player I is c_1 instead of d_1, while the strategy of Player II is the same if we swap the values of q_i and r_i. In Section 4.3 we will show that the probability of SPIT calls c_1 is $c_1 > d_1$, for $u_c < 2u_l$. That is, the rate of SPIT calls in the SpitGame′ is increased in comparison with the corresponding case of the SpitGame. We will also compare the corresponding utilities of Player II in both models.

Working in the same way it is straightforward to adapt the rest of the analysis of the original SpitGame to the SpitGame′. The results are presented below.

Case 2: $\epsilon_1 < \sigma$

Case 2.1: $\epsilon_2/h_2 < f_s/h_1$

Case 2.1.1: $\epsilon_1 < \sigma < \epsilon_1 + \epsilon_2$

p	p_1	r_1	p_2	r_2	p_3	r_3
c_2	1	0	$\frac{\sigma-\epsilon_1}{\epsilon_2}$	$1-p_2$	0	1

Case 2.1.2: $\epsilon_1 + \epsilon_2 \leq \sigma$.

p	p_1	r_1	p_2	r_2	p_3	r_3
c_3	1	0	1	0	$\frac{\sigma-\epsilon_1-\epsilon_2}{f_s}$	$1-p_3$

Case 2.2: $\epsilon_2/h_2 > f_s/h_1$.

p	p_1	r_1	p_2	r_2	p_3	r_3
c_3	1	0	0	1	$\frac{\sigma-\epsilon_1}{f_s}$	$1-p_3$

Case 2.2.2: $\epsilon_1 < \sigma < \epsilon_1 + \epsilon_2$.

p	p_1	r_1	p_2	r_2	p_3	r_3
c_2	1	0	$\frac{\sigma-\epsilon_1-f_s}{\epsilon_2}$	$1-p_2$	1	0

Case 2.3: $\epsilon_2/h_2 = f_s/h_1$.

p	p_1	r_1	p_2	r_2	p_3	r_3
c_2	1	0	p_2	$1-p_2$	$\frac{\sigma-\epsilon_1-\epsilon_2 p_2}{f_s}$	$1-p_3$

where $c_2 = c_3$ and the range of values for p_2 is

$$\max\{0, \frac{\sigma-\epsilon_1-f_s}{\epsilon_2}\} \leq p_2 \leq \frac{\sigma-\epsilon_1-f_s \max\{0, \frac{\sigma-\epsilon_1-\epsilon_2}{f_s}\}}{\epsilon_2} \,. \tag{36}$$

The closed forms of the NE for all cases of the SpitGame and the SpitGame$'$ are summarized in Table 8.

4.3 The benefit of supporting audio CAPTCHAs

We can now compare the NE of the SpitGame and the SpitGame$'$ in order to assess the effect of audio CAPTCHAs on the properties of the corresponding NE. We are interested in the rate of SPIT calls at the NE and the corresponding utility of Player II, the VoIP user.

Note that the strategy of Player I is always some value d_i, for $i \in \{1,2,3\}$ in the SpitGame, and c_i for the same index value of i in the corresponding SpitGame$'$. Using Equation 20 it is straightforward to show that $c_i > d_i$, for any $i \in \{1,2,3\}$, which implies a reduced rate of SPIT calls at the NE of the SpitGame. For example the ratio c_1/d_1 is

$$c_1/d_1 = \frac{2u_l(f_l u_c + \epsilon_1 u_s)}{u_c(2f_l u_l + \epsilon_1 u_s)} > 1. \tag{37}$$

Similarly, the ratios c_2/d_2 and c_3/d_3 can also be shown to be larger than 1.

Theorem 3. *At NE, the rate of SPIT calls is strictly less when users have the*

option to submit audio CAPTCHA's.

The utility of Player II at any NE of the SpitGame is larger than in the NE of the corresponding SpitGame. For example, the difference of the utility of Player II in Case 1 of the SpitGame minus the utility of the NE of the corresponding NE in the SpitGame$'$ is

$$U_2 - U_2' = \frac{\epsilon_1 h_2 (2u_l - u_c) u_s (f_l u_l + \epsilon_1 u_s)}{(f_l u_c + \epsilon_1 u_s)(2 f_l u_l + \epsilon_1 u_s)} > 0. \tag{38}$$

Note, that we use the difference for the utilities instead of the ratio, because Player II may have a negative utility in the SpitGame$'$. For Case 2.1.1 the difference is

$$U_2 - U_2' = \frac{\epsilon_2 h_2 (2u_l - u_c) u_s (h_2 u_l + \epsilon_2 u_s)}{(h_2 u_c + \epsilon_2 u_s)(2 h_2 u_l + \epsilon_2 u_s)} > 0. \tag{39}$$

In the same way, the difference of the utilities of Player II at NE in the SpitGame and the SpitGame$'$ can be shown to be positive for the remaining cases of the game.

Theorem 4. *At NE, the utility of Player II is larger in the SpitGame than in the corresponding SpitGame$'$.*

5 Experimental Study

For the experimental analysis we produce the theoretically predicted Nash Equilibria properties independently from the theoretical analysis. We have selected realistic values for the filter's ability to discern legitimate calls from SPIT calls based on the analysis performed in Sections 3 and 4. The experimental analysis was performed for three filter specification cases, shown in Table 9, and for each case we examined the NE of both SpitGame and SpitGame$'$.

The first filter specification case represents the most realistic case: the filter has significant difficulties in identifying SPIT calls resulting in a large percentage of SPIT calls being classified as *Unknown*, but can classify legitimate calls with relatively high accuracy. The second filter specification represents the conditions in a large organisation which receives calls from a large pool of people. As a result, it tends to classify both SPIT and legitimate calls as *Unknown*. The third filter specification represents a smaller organisation with a much smaller pool of frequent callers. Therefore, it tends to identify SPIT and legitimate calls much more accurately than in the previous two cases.

In order to reduce the original problem from a 5D parameter space into an equivalent 3D exploration space we take advantage of the conditions on the parameters shown in Table 2 to set $u_l = 100$ and $s_a = 100$. In order to further reduce the number of problem instances to solve, we take integral values for u_s, u_c and s_r. The restrictions convert the original 5D parameter space into a 3D exploration space shown in Table 10. Additionally, we performed adaptive exploration of the games for values of s_r near the boundary conditions for each case.

We automatically computed the Nash equilibria of these games using the *gambit-lcp* program supplied with Gambit [25] and fitted the resulting data to functions independently from the theoretical analysis.

5.1 Experimental Results & Discussion

The first result is that the Nash equilibria are unique, i.e., for each set of distinct values of u_c, u_s and s_r, the game produces exactly one Nash equilibrium. This has also significantly simplified our results and their analysis. It also means that there are no other equilibriums, with potentially worse outcomes for the user, for the game. As a result, the user's selection of strategies, given the SPIT sender's pay-offs always leads to exactly one equilibrium state. We have also verified empirically the validity of Theorem 1, by finding that all the NE, in all game instances, are mixed.

Another interesting result is the percentage of legitimate calls that the SPIT sender decides on (or conversely, the percentage of SPIT calls, as they are complementary) in the NE as a function of u_c and u_s. In the filter specification cases 1 and 2 there are two s_r value groups ($1 \leq s_r \leq 11.\overline{1}$ and $11.\overline{1} < s_r \leq 99$), while filter specification case 3 has three s_r value groups ($1 \leq s_r \leq 5.263$, $5.263 < s_r < 42.86$ and $42.86 < s_r \leq 99$). These results are shown in Fig. 2. These s_r value groupings are correspond to the two base cases (1 and 2) illustrated in Table 8 for both SpitGame and SpitGame$'$. As an example, for the first filter specification and for case 1 in Table 8:

$$\begin{aligned} \epsilon_1 &\geq \sigma \Leftrightarrow \\ 0.1 &\geq s_r/(s_a + s_r) \Leftrightarrow \\ 0.1 &\geq s_r/(100 + s_r) \Leftrightarrow \\ 11.\overline{1} &\geq s_r \end{aligned}$$

Case 2 is the complement of case 1, so $11.\overline{1} < s_r$.

In these Nash Equilibria and within each of the value groups for s_r the percentage of legitimate calls as a function of u_c, u_s is identical. Realistically, the actual s_r value for SPIT senders will be relatively low and probably $\leq 11.\overline{1}$ (i.e., the cost of attempting a SPIT call is $\leq 11.\overline{1}\%$ of the value gained if

the SPIT call goes through) given the resources needed to make SPIT calls. The difference between the value groups pertains to the rate with which the percentage of the legitimate calls decreases in relation to u_s and u_c. The figure illustrates that as the cost of deploying the CAPTCHA mechanism increases, the number of SPIT calls also increases (legitimate percentage decreases). At the same time, as the disutility of accepting SPIT calls (u_s) increases for the user, the number of legitimate calls increases, purportedly due to the increase in CAPTCHA use providing a strong disincentive to the SPIT sender.

The percentage $[0.0 - 1.0]$ of legitimate calls as a function of u_c and u_s has been fitted to the functions L_1, L_2, L_3 as shown in Table 11, which are identical to the ones produced in the theoretical analysis. For each of the fitted function cases, we have parenthesised its corresponding case in the NE Table 8.

5.2 Comparison of SpitGame and SpitGame′

In order to examine whether the use of the CAPTCHA challenge provides benefits to the users, we created a game model where all the *CAPTCHA* challenge actions have been removed (SpitGame′) and only *Accept* and *Reject* actions are present. Using the same value ranges for u_l, u_s, s_r, s_a and disregarding u_c (since there are no CAPTCHA challenges present) we performed the same experiments at the same granularity as before. Our findings from comparing the model without CAPTCHA (SpitGame′) to the model with CAPTCHA (SpitGame) are summarized in Table 12.

The use of CAPTCHA leads in to notable improvements to the percentage of legitimate calls since in no case does the percentage of legitimate calls drop. The improvement in percentage of legitimate calls is shown in Fig. 3. It is notable that for the filter specifications 1 and 2 when $s_r \leq 11.\overline{1}$ and for the filter specification 3 when $s_r \leq 5.263$, the CAPTCHA-less model performs so

badly that the measure of improvement is almost identical to the performance of the model with CAPTCHA.

Further discoveries include the fact that for the first filter specification, for all values of s_r, when the filter identifies a call as *SPIT*, only the *CAPTCHA* action is used (never *Reject* or *Accept*). Also, even when the call is identified as either *Legitimate* or *Unknown*, the *Reject* action is never used. Furthermore, when the filter identifies a call as *Legitimate* and $s_r > 11.\overline{1}$ the user always selects *Accept*. Finally, when the filter identifies a call as *Unknown* and $s_r \leq 11.\overline{1}$ the user never selects *Accept*. These discoveries, summarized in Table 13, mean that for the more realistic values of s_r ($\leq 11.\overline{1}$) the *Accept* action can be removed without impact when the filter identifies a call as *Unknown* or *SPIT*.

6 Conclusions and further research

Spam over Internet Telephony is a significant threat for VoIP communications, which may become a serious problem just like ordinary spam is for email. In this paper, we focused on the strategic interaction between SPIT senders and legitimate VoIP users. We assumed the existence of incoming call filters and effective audio CAPTCHAs and armed the VoIP users with the option to accept an incoming call, to reject it or to request an audio CAPTCHA based on a filter's verdict.

The main contribution of our work is the derivation of game-theoretic model that captures the interaction of independent, selfish SPIT senders and VoIP users. Through theoretical arguments and a comprehensive experimental analysis we studied the properties of the proposed game and identified its Nash equilibria.

The outcomes of our approach show that the use of the above mentioned defensive mechanisms lead to desirable Nash equilibria, where audio CAPTCHAs

contribute to the utility of the legitimate users. Moreover, if the user and SPIT sender pay-offs are known, then the game always leads to exactly one equilibrium state with predictable characteristics.

In is noteworthy that in our model we allow for the attacker (SPIT user) to already know the performance characteristics of our filter. As a result, we are not vulnerable to attacks which would uncover the filter's characteristics. In addition, at NE, all players, hence SPIT senders too, have full knowledge of the strategies of their opponents, but still cannot achieve a better outcome. This means that in our approach we are not attempting to secure through obscurity.

The game-theoretic model of this work can be extended in several aspects to capture more properties of the real problem. An interesting topic for further research could be to refine the audio CAPTCHAs, for example, with additional parameters to model the solvability of the audio CAPTCHA. We have assumed here that the audio CAPTCHA are always solvable by a legitimate user and never solvable by a SPIT sender (automated SPIT application). New research works [5][42] have proven that about 10% of the humans are unable to solve them and that the success rate of the bots is about 5%. This new parameter would cover these edge cases.

Building upon the theoretical arguments and the experimental results presented here, we plan to work on performing a complete theoretical analysis of the SpitGame [14]. As part of this analysis, we plan to investigate how different filter parameters influence the Nash equilibria and lead the VoIP users and the SPIT sender to adjust their behavior. This will aid in further informing the decisions on trade-offs when implementing real CAPTCHA-based anti-SPIT systems.

7 Acknowledgments

This work was performed in the framework of the SPHINX (09SYN-72-419) Project, which is partly funded by the Hellenic General Secretariat for Research and Technology (http://sphinx.vtrip.net).

8 References

1. I. Androutsopoulos, E. Magirou and D. Vassilakis, A game theoretic model of spam e-mailing, in: *Proc. of the 2nd Conference on Email and Anti-Spam*, Stanford University, USA, 2005.

2. V. Balasubramaniyan, M. Ahamad and H. Park, Callrank: Combating spit using call duration, social networks and global reputation, in: *Proc. of the 4th Conference on Email and Anti-Spam*, USA, August 2007.

3. K. Basu, The traveler's dilemma: Paradoxes of rationality in game theory, *American Economic Review*, 84(2):391–95, May 1994.

4. D. Braess, Über ein Paradoxon aus der Verkehrsplanung. Unternehmensforschung 12, pp. 258-268, 1968.

5. E. Bursztein, S. Bethard, C. Fabry, J. Mitchell and D. Jurafsky, How good are humans at solving CAPTCHA? A large scale evaluation, in: *Proc. of the 2010 IEEE Symposium on Security and Privacy*, pp. 399-413, USA, 2010.

6. H. Cavusoglu and S. Raghunathan, Configuration of detection software: A comparison of decision and game theory approaches, *Decision Analysis*, Vol. 1, No. 3, pp. 131-148, 2004.

7. C. Daskalakis, P.W. Goldberg and C.H. Papadimitriou, The complexity of computing a Nash equilibrium, *Commun. ACM*, 52(2):89–97, February 2009.

8. S. Dritsas, B. Tsoumas, V. Dritsou , P. Konstantopoulos and D. Gritzalis, OntoSPIT: SPIT Management through Ontologies, *Computer Communications*, vol. 32, no. 2, pp. 203-212, 2009.

9. Federal Communications Commission, FCC Strengthens Consumer Protections Against Telemarketing Robocalls, In the Matter of Rules and Regulations Implementing the Telephone Consumer Protection Act of 1991, CG Docket No. 02-278, February 15, 2012.

10. Federal Trade Commission, Do-Not-Call Implementation Act of 2003, Public Law No. 108-10, June 2003.

11. Federal Trade Commission, National Do Not Call Registry Data Book for Fiscal Year 2012, October 2012.

12. Federal Trade Commission,FTC Settles "Rachel" Robocall Enforcement Case (http://www.ftc.gov/opa/2013/07/aplus.shtm , retrieved 23 October 2013).

13. D. Graham-Rowe, A sentinel to screen phone calls technology, *MIT Review*, 2006.

14. D. Gritzalis, P. Katsaros, S. Basagiannis and Y. Soupionis, Formal analysis for robust anti-SPIT protection using model-checking, *International Journal of Information Security*, vol. 11, no. 2, pp. 121-135, 2012.

15. D. Gritzalis, V. Katos, P. Katsaros, Y. Soupionis, J. Psaroudakis and A. Mentis, The Sphinx enigma in critical VoIP infrastructures: Human or

botnet?, in: *Proc. of the 4th International Conference on Information, Intelligence, Systems and Applications*, IEEE Press, 2013.

16. D. Gritzalis, G. Marias, Y. Rebahi, Y. Soupionis and S. Ehlert, SPIDER: A platform for managing SIP-based spam over Internet Telephony, *Journal of Computer Security*, Vol. 19, No. 5, pp. 835-867, 2011.

17. B. Johnson, J. Grossklags, N. Christin and J.Chuang, Are Security Experts Useful? Bayesian Nash Equilibria for Network Security Games with Limited Information, in: *Proc. of the 15th European Symposium on Research in Computer Security*, pp. 588-606, Greece, September 2010.

18. A. Johnston, *SIP: Understanding the Session Initiation Protocol*, 2nd edition, Artech House, 2004

19. C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. Voelker, V. Paxson and S. Savage, Spamalytics: An empirical analysis of spam marketing conversion, in: *Proc. of the 15th ACM Conference on Computer and Communications Security*, pp. 3-14, USA, October 2008.

20. A. Keromytis, Voice-over-IP Security: Research and Practice, *IEEE Security and Privacy*, vol. 8, no. 2, pp. 76-78, 2010.

21. A. Keromytis, A Comprehensive Survey of Voice over IP Security Research, *IEEE Communications Surveys & Tutorials*, vol. 14, no. 2, pp. 514-537, 2012.

22. Y. Kim, Y. Park and J. Lee, Using stated-preference data to measure the inconvenience cost of spam among Korean e-mail users, *Applied Economics Letters*, Vol. 13, No. 12, pp. 795-800, 2006.

23. D. Lowd and C. Meek, Good word attacks on statistical spam filters, in: *Proc. of the 2nd Conference on Email and Anti-Spam*, pp. 21-22, USA,

2005.

24. M.H. Manshaei, Q. Zhu, T. Alpcan, T. Bacşar and J.P. Hubaux, Game theory meets network security and privacy, *ACM Comput. Surv.*, 45(3):25:1–25:39, July 2013.

25. J. McKelvey, D. Richard, A. McLennan and T. Turocy, *Gambit: Software tools for game theory*, ver. 0.2010.09.01. http://www.gambit-project.org.

26. R.B. Myerson, *Game Theory: Analysis of Conflict*, Harvard University Press, Cambridge, MA, 1991.

27. S. Niccolini, S. Tartarelli, M. Stiemerling and S. Srivastava, *SIP Extensions for SPIT Identification*, Internet Draft, Network Working Group, 2007, draftniccolini-sipping-feedback-spit-03.

28. N. Nisan, T. Roughgarden, E. Tardos and V.V. Vazirani, *Algorithmic Game Theory*, Cambridge University Press, New York, NY, USA, 2007.

29. M. Osborne and A. Rubinstein, *A Course in Game Theory*, The MIT Press, 1994.

30. M. Osborne, *An Introduction to Game Theory*, Oxford University Press, 2003.

31. G. Owen, *Game Theory*, Academic Press, 1982.

32. C.H. Papadimitriou, Algorithms, Games, and the Internet, in: *Proceedings of the 33rd ACM STOC*, pages 749–753, New York, NY, USA, 2001.

33. M. Parameswaran, H. Rui and S. Sayin, A game theoretic model and empirical analysis of spammer strategies, in: *Proc. of the Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, 2010.

34. P. Patankar, G. Nam, G. Kesidisand and C. Das, Exploring anti-spam models in large scale voip systems, in: *Proc. of the 28th International Conference on Distributed Computing Systems*, China, June 2008.

35. J. Quittek, S. Niccolini, S. Tartarelli, M. Stiemerling, M. Brunner and T. Ewald, Detecting SPIT calls by checking human communication patterns, in: *Proc. of the IEEE International Conference on Communications*, pp. 1979-84, UK, 2007.

36. J. Rosenberg and C. Jennings, The Session Initiation Protocol (SIP) and Spam, *Network Working Group*, RFC 5039, January 2008.

37. S. Sawda and O. Urien, SIP security attacks and solutions: A state-of-the-art review, in: *Proc. of the IEEE International Conference on Information and Communication Technologies*, pp. 3187-3191, April 2006.

38. A.B. Shahroudi, R.H. Khosravi, H.R. Mashhadi and M. Ghorbanian, Full Survey on SPIT and prediction of how VoIP providers compete in presence of SPITTERS using Game-Theory, in: *Proc. of the 2011 IEEE International Conference on Computer Applications and Industrial Electronics (ICCAIE)*, pp. 402 - 406, Abu Dhabi, 2011

39. D. Shin, J. Ahn and C. Shim, Progressive multi gray-leveling: a voice spam protection algorithm, *IEEE Network*, Vol. 20, No. 5, pp. 18-25, 2006.

40. H. Sinnreich and B. A. Johnston, *Internet Communications Using SIP: Delivering VoIP and Multimedia Services with Session*, Second Edition, Wiley Publishing Inc., 2006

41. Y. Soupionis and D. Gritzalis, ASPF: An adaptive anti-SPIT policy-based framework, in: *Proc. of the 6th International Conference on Availability,*

Reliability and Security, pp. 153-160, Austria, August 2011.

42. Y. Soupionis and D. Gritzalis, Audio CAPTCHA: Existing solutions assessment and a new implementation for VoIP telephony, *Computers & Security*, Vol. 29, No. 5, pp. 603-618, 2010.

43. Y. Soupionis, S. Dritsas and D. Gritzalis, An adaptive policy-based approach to SPIT management, in: *Proc. of the 13th European Symposium on Research in Computer Security*, pp. 446-460, Springer, 2008.

44. M. Tambe, M. Jain, J.A. Pita and A.X. Jiang, Game theory for security: Key algorithmic principles, deployed systems, lessons learned, in: *50th Annual Allerton Conference on Communication, Control, and Computing*, pages 1822–1829, 2012.

45. D. Vassilakis, I. Androutsopoulos and E. Mageirou, A game-theoretic investigation of the effect of human interactive proofs on spam e-mail, in: *Proc. of the 4th Conference on Email and Anti-Spam*, USA, 2007.

46. T. Walsh and D. Kuhn, Challenges in securing voice over IP, *IEEE Security and Privacy*, Vol. 3, No. 3, pp. 44-49, 2005.

47. T. Wilson, *Competition may be driving surge in botnets and spam.* www.darkreading.com/security/security-management/208803799

Tables

Table 1: Game-theoretic model utilities

	User/Player II			**SPIT sender/Player I**		
Message	***Accept***	***Reject***	***CAPTCHA***	***Accept***	***Reject***	***CAPTCHA***
Legitimate	u_l	$-u_l$	$u_l - u_c$	0	0	0
SPIT	$-u_s$	0	0	s_a	$-s_r$	$-s_r$

Table 2: Player preferences parameters

Player	**Parameter**	**Description**	**Conditions(absolute values)**
User/Player II	u_l	Measure of user utility of accepting legitimate call	$u_l > u_s > u_c > 0$
	u_s	Measure of user disutility of accepting SPIT call	
	u_c	Measure of user disutility of sending CAPTCHA	
SPIT sender/Player I	s_a	Measure of SPIT sender utility of getting a SPIT call accepted	$s_a > s_r > 0$
	s_r	Measure of SPIT sender disutility of getting a SPIT call rejected	

Table 3: The filter verdicts.

	Filter verdict		
Type of call	*Legitimate*	*Unknown*	*SPIT*
SPIT call	ϵ_1	ϵ_2	f_s
Legitimate call	f_l	h_2	h_1

Table 4: The strategy of Player I at a NE

Action of Player I	Probability
SPIT call	p
Legitimate call	1-p

Table 5: The strategy of Player II at a NE

		Action of Player II		
	Information Set (Filter verdict)	*Accept*	*CAPTCHA*	*Reject*
1	*Legitimate* call	p_1	q_1	$r_1 = 1 - p_1 - q_1$
2	*Unknown*	p_2	q_2	$r_2 = 1 - p_2 - q_2$
3	*SPIT* call	p_3	q_3	$r_3 = 1 - p_3 - q_3$

Table 6: The coefficients for Equation 15

i	A_i	B_i	C_i	D_i
1	$2f_l u_l(1-p) - \epsilon_1 u_s p$	$f_l(2u_l - u_c)(1-p)$	$-f_l u_l(1-p)$	$f_l(1-p) + \epsilon_1 p$
2	$2h_2 u_l(1-p) - \epsilon_2 u_s p$	$h_2(2u_l - u_c)(1-p)$	$-h_2 u_l(1-p)$	$h_2(1-p) + \epsilon_2 p$
3	$2h_1 u_l(1-p) - f_s u_s p$	$h_1(2u_l - u_c)(1-p)$	$-h_1 u_l(1-p)$	$h_1(1-p) + f_s p$

Table 7: Boundary values of p

Equation	Condition	Equation	Condition
$A_1 = 0,$	if $p = \frac{2f_l u_l}{2f_l u_l + \epsilon_1 u_s} = c_1$	$A_1 = B_1,$	if $p = \frac{f_l u_c}{f_l u_c + \epsilon_1 u_s} = d_1$
$A_2 = 0,$	if $p = \frac{2h_2 u_l}{2h_2 u_l + \epsilon_2 u_s} = c_2$	$A_2 = B_2,$	if $p = \frac{h_2 u_c}{h_2 u_c + \epsilon_2 u_s} = d_2$
$A_3 = 0,$	if $p = \frac{2h_1 u_l}{2h_1 u_l + f_s u_s} = c_3$	$A_3 = B_3,$	if $p = \frac{h_1 u_c}{h_1 u_c + f_s u_s} = d_3$

Table 8: The NE of SpitGame and SpitGame′ (without CAPTCHAs).
The ranges of values for p_2 in case 2.3 of SpitGame and 2.3 of SpitGame′ are given in Equations 34 and 36, respectively.

			Player II (Information Sets)								
		Player I	*Legitimate*			*Unknown*			*SPIT*		
SpitGame Case		$p \quad 1-p$	p_1	q_1	r_1	p_2	q_2	r_2	p_3	q_3	r_3
1	$\epsilon_1 \geq \sigma$	$d_1(1-d_1)$	$\frac{\sigma}{\epsilon_1}$	$1-\frac{\sigma}{\epsilon_1}$	0	0	1	0	0	1	0
2	$\epsilon_1 < \sigma$										
2.1	$\epsilon_2/h_2 < f_s/h_1$										
2.1.1	$\epsilon_1+\epsilon_2 > \sigma$	$d_2(1-d_2)$	1	0	0	$\frac{\sigma-\epsilon_1}{\epsilon_2}$	$1-p_2$	0	0	1	0
2.1.2	$\epsilon_1+\epsilon_2 \leq \sigma$	$d_3(1-d_3)$	1	0	0	1	0	0	$\frac{\sigma-\epsilon_1-\epsilon_2}{f_s}$	$1-p_3$	0
2.2	$\epsilon_2/h_2 > f_s/h_1$										
2.2.1	$\epsilon_1+\epsilon_2 > \sigma$	$d_3(1-d_3)$	1	0	0	0	1	0	$\frac{\sigma-\epsilon_1}{f_s}$	$1-p_3$	0
2.2.2	$\epsilon_1+\epsilon_2 \leq \sigma$	$d_2(1-d_2)$	1	0	0	$\frac{\sigma-\epsilon_1-f_s}{\epsilon_2}$	$1-p_2$	0	1	0	0
2.3	$\epsilon_2/h_2 = f_s/h_1$	$d_2(1-d_2)$	1	0	0	p_2	$1-p_2$	0	$\frac{\sigma-\epsilon_1-\epsilon_2 p_2}{f_s}$	$1-p_3$	0
SpitGame′ Case		$p \quad 1-p$	p_1	q_1	r_1	p_2	q_2	r_2	p_3	q_3	r_3
1	$\epsilon_1 \geq \sigma$	$c_1(1-c_1)$	$\frac{\sigma}{\epsilon_1}$	0	$1-\frac{\sigma}{\epsilon_1}$	0	0	1	0	0	1
2	$\epsilon_1 < \sigma$										
2.1	$\epsilon_2/h_2 < f_s/h_1$										
2.1.1	$\epsilon_1+\epsilon_2 > \sigma$	$c_2(1-c_2)$	1	0	0	$\frac{\sigma-\epsilon_1}{\epsilon_2}$	0	$1-p_2$	0	0	1
2.1.2	$\epsilon_1+\epsilon_2 \leq \sigma$	$c_3(1-c_3)$	1	0	0	1	0	0	$\frac{\sigma-\epsilon_1-\epsilon_2}{f_s}$	0	$1-p_3$
2.2	$\epsilon_2/h_2 > f_s/h_1$										
2.2.1	$\epsilon_1+\epsilon_2 > \sigma$	$c_3(1-c_3)$	1	0	0	0	1	0	$\frac{\sigma-\epsilon_1}{f_s}$	0	$1-p_3$
2.2.2	$\epsilon_1+\epsilon_2 \leq \sigma$	$c_2(1-c_2)$	1	0	0	$\frac{\sigma-\epsilon_1-f_s}{\epsilon_2}$	0	$1-p_2$	1	0	0
2.3	$\epsilon_2/h_2 = f_s/h_1$	$c_2(1-c_2)$	1	0	0	p_2	0	$1-p_2$	$\frac{\sigma-\epsilon_1-\epsilon_2 p_2}{f_s}$	0	$1-p_3$

Table 9: The experimental filter verdicts.

Filter Specification	Type of call	Filter verdict *Legitimate*	*Unknown*	*SPIT*
1	*SPIT*	0.1	0.6	0.3
	Legitimate	0.7	0.25	0.05
2	*SPIT*	0.1	0.6	0.3
	Legitimate	0.3	0.6	0.1
3	*SPIT*	0.05	0.25	0.7
	Legitimate	0.7	0.25	0.05

Table 10: Solution exploration space

$\mathbf{u_s}$		$\mathbf{u_c}$		$\mathbf{s_r}$		**# Instances**
$2 \dots 99$	$\times$	$1 \dots u_s - 1$	$\times$	$1 \dots 99$	$=$	~ 500000

Filter Spec.	Fitted functions for % of legitimate calls	Abs. Fitting Error
1	$L_1(u_c, u_s) = \frac{u_s}{u_s + \alpha u_c}, \alpha = \begin{cases} 7, & 1 \le s_r \le 11.\overline{1} \quad (1) \\ 0.41\overline{6}, & 11.\overline{1} < s_r \le 99 \quad (2.1.1) \end{cases}$	$\le 5.2 \times 10^{-11}$
2	$L_2(u_c, u_s) = \frac{u_s}{u_s + \alpha u_c}, \alpha = \begin{cases} 3, & 1 \le s_r \le 11.\overline{1} \quad (1) \\ 1, & 11.\overline{1} < s_r \le 99 \quad (2.1.1) \end{cases}$	$\le 5.01 \times 10^{-11}$
3	$L_3(u_c, u_s) = \frac{u_s}{u_s + \alpha u_c}, \alpha = \begin{cases} 14, & 1 \le s_r \le 5.263 \quad (1) \\ 1, & 5.263 < s_r < 42.86 \quad (2.1.1) \\ 0.0714, & 42.86 \le s_r \le 99 \quad (2.1.2) \end{cases}$	$\le 5.14 \times 10^{-11}$

Table 11: Fitted functions for % of legitimate calls $((1-p) * 100)$ (function of u_c and u_s for the s_r value groups)

Table 12: Major findings from comparison of models with (SpitGame) and without CAPTCHA (SpitGame′) in NE

Filter Spec.	Property	Model	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**
			$\mathbf{1 \leq s_r \leq 11.\overline{1}}$		$\mathbf{11.\overline{1} < s_r \leq 99}$			
1	Legit.	SpitGame	12.60%	93.45%	70.79%	99.58%		
	Calls	SpitGame′	0.14%	6.60%	2.34%	54.30%		
	User	SpitGame	0.13	92.52	46.85	99.24		
	Utility	SpitGame′	-6.60	-0.14	0.74	17.19		
			$\mathbf{1 \leq s_r \leq 11.\overline{1}}$		$\mathbf{11.\overline{1} < s_r \leq 99}$			
2	Legit.	SpitGame	25.19%	97.06%	50.25%	99.00%		
	Calls	SpitGame′	0.33%	14.16%	0.99%	33.11%		
	User	SpitGame	0.50	0.96	10.86	98.2		
	Utility	SpitGame′	-14.16	-0.33	-19.87	-0.59		
			$\mathbf{1 \leq s_r \leq 5.263}$		$\mathbf{5.263 < s_r < 42.86}$		$\mathbf{42.86 \leq s_r \leq 99}$	
3	Legit.	SpitGame	6.73%	87.61%	50.25%	99.00%	93.40%	99.93%
	Calls	SpitGame′	0.07%	3.42%	0.99%	33.11%	12.28%	87.39%
	User	SpitGame	0.13	86.73	33.02	98.65	86.86	99.86
	Utility	SpitGame′	-3.42	-0.07	0.30	9.93	10.53	74.91

Table 13: Summary of actions used based on filter call identification and the value s_r in the first filter specification case

Call identified as	**Actions Used**	
	$\mathbf{1 \leq s_r \leq 11.\overline{1}}$	$\mathbf{11.\overline{1} < s_r \leq 99}$
Legitimate	*Accept*, *CAPTCHA*	*Accept*
Unknown	*CAPTCHA*	*Accept*, *CAPTCHA*
SPIT	*CAPTCHA*	*CAPTCHA*

Figure Captions

Fig. 1. The game-theoretic model

Fig. 2. % of legitimate calls ($(1-p)*100$) (function of u_c and u_s for the s_r value groups)

Fig. 3. Improvement (absolute difference) of % of legitimate calls with CAPTCHA (SpitGame) vs. without CAPTCHA (SpitGame$'$)

Figures

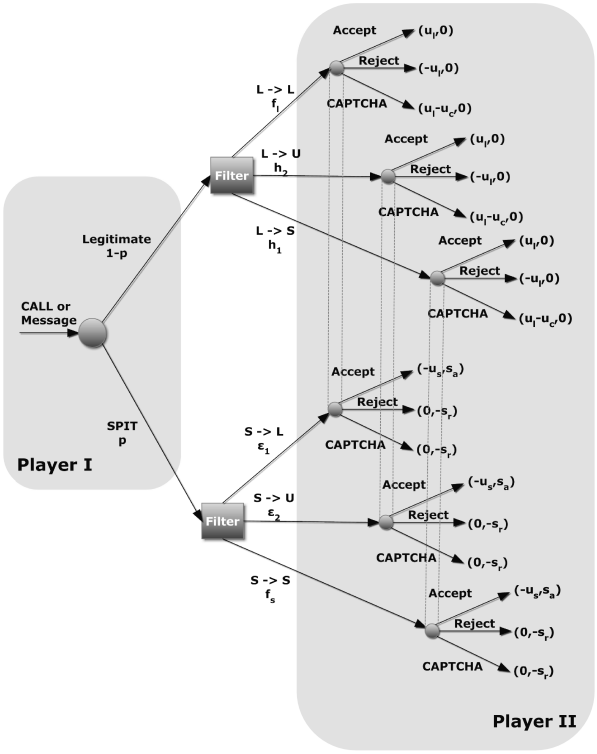

Fig. 1: The game-theoretic model

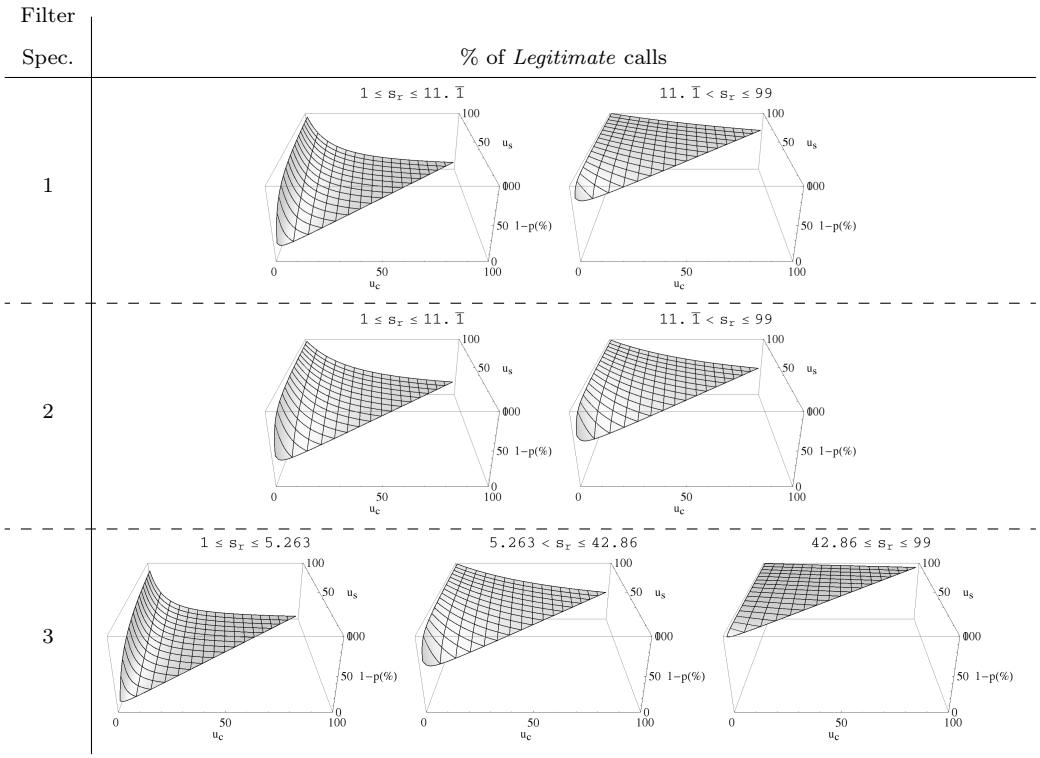

Fig. 2: % of legitimate calls $((1-p)*100)$ (function of u_c and u_s for the s_r value groups)

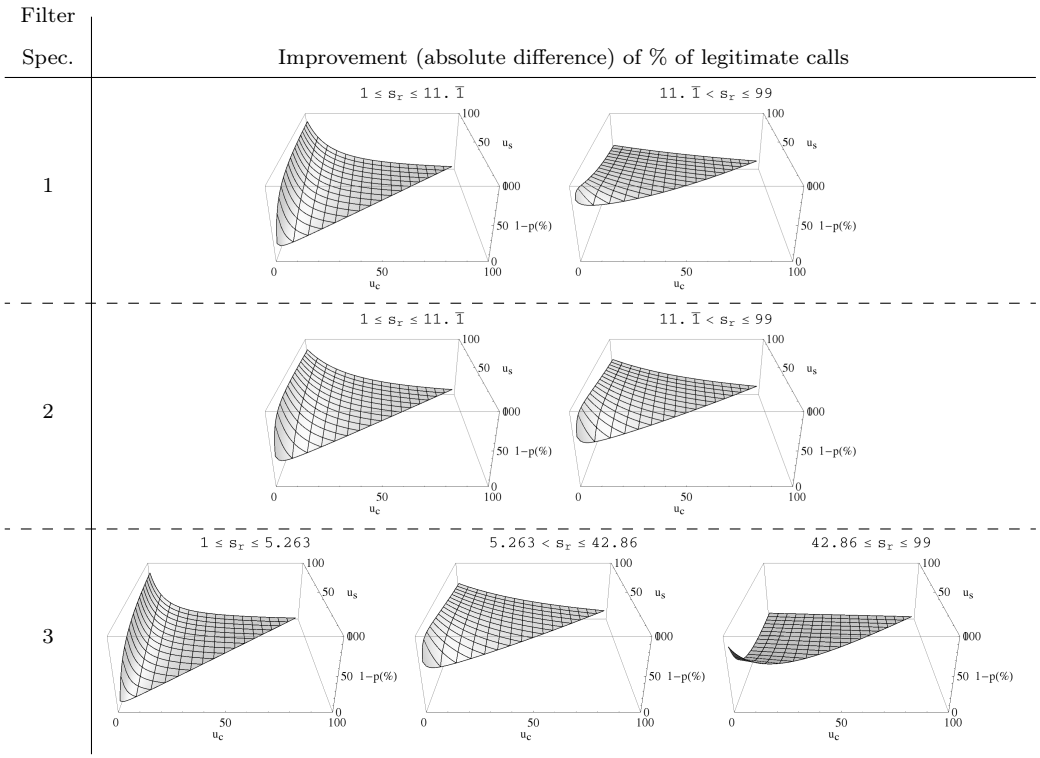

Fig. 3: Improvement (absolute difference) of % of legitimate calls with CAPTCHA (SpitGame) vs. without CAPTCHA (SpitGame′)