

# Informant: Detecting Sybils Using Incentives

N. Boris Margolin and Brian N. Levine

Department of Computer Science, Univ. of Massachusetts, Amherst, MA, USA  
{margolin,brian}@cs.umass.edu

**Abstract.** We propose an economic approach to Sybil attack detection. In our *Informant* protocol, a *detective* offers a reward for Sybils to reveal themselves. The detective accepts from one identity a security deposit and the name of target peer; the deposit and a reward are given to the target. We prove the optimal strategy for the informant is to play the game if and only if she is Sybil with a low opportunity cost, and the target will cooperate if and only if she is identical to the informant. Informant uses a Dutch auction to find the minimum possible reward that will reveal a Sybil attacker. Because our approach is economic, it is not limited to a specific application and does not rely on a physical device or token.

## 1 Introduction

Networked applications often assume or require that identities over network have a one-to-one relationship with individual entities in the external world. A single individual who controls many identities can disrupt, manipulate, or corrupt peer-to-peer applications and other applications that rely on redundancy; this is commonly called the Sybil attack [1].

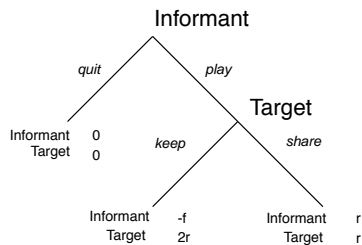
While there has been quite a bit of research on deterring Sybil attacks using such techniques as identity certification, resource testing, and reputation systems [1–6], detection of Sybil attacks has received less attention. Existing detection methods are applicable only in very specific circumstances and applications, such as mobile sensor networks [7–9].

In this work, we take an economic approach in proposing a novel detection protocol called *Informant*. Our protocol provides an incentive to Sybil attackers to reveal two or more controlled identities in exchange for a payment. When the offered incentive exceeds the attackers opportunity cost for this admission, rational attackers will participate. Sybil identities by definition cannot be easily linked to any real-world identity — if they could, preventing the Sybil attack would be trivial. Thus, these revelations do not reveal the attacker herself, which removes an impediment towards participating.

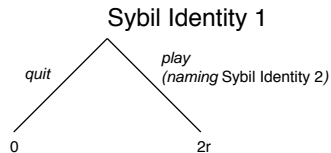
One of the key challenges in designing our incentives scheme is avoiding false claims — a presentation of identities that are *not* controlled by a single entity. To solve this problem, we introduce a *trust game* that makes false claims financially risky for the claimant. Figure 1 outlines the game, which takes place between three identities: a *detective*, an *informant*, and a *target*. The informant and the target may be independent or they may be part of a Sybil controlled by a single

---

This paper was supported in part by National Science Foundation award NSF-0133055.



**Fig. 1.** Utilities in the Trust game when the Identities are independent.



**Fig. 2.** Utilities in the Trust game when the Identities are part of a Sybil attack.

entity; the detective offers a reward as incentive for the informant to reveal the truth. In a real peer-to-peer application, the informant and target are simply participants; the detective can be any entity. For an informant to claim to be part of a Sybil with a target, she must provide a security deposit  $f$  to the detective. The target then receives a payment of  $\$(2r + f)$  from the detective. She is free to do whatever she wants with the reward.

If the identities are controlled by the same entity, that entity has made a profit of  $\$2r$  by revealing its nature. If they are independent, the informant has lost  $\$f$ . She might ask the informant to give her back  $\$(r + f)$  so that they share the reward of  $\$2r$ , but she has no way of imposing this desire, so a rational target will not cooperate. (We defer discussion of repeated games Section 4.) As we show more formally in this paper, when informants are rational, the detective can believe any claim.

This game does not distinguish between Sybil attackers and friends with a very high degree of trust in each other. But entities with high trust in each other, even if not actually malicious, are a matter for concern. They are capable of attacking or distorting the operation of an application, and break the assumption of redundancy required for availability, security, or privacy in many distributed applications [10–12]. While trust in the real world is complex, we are concerned only with a simple, monetary definition of trust and trustworthiness.

### Contributions.

- We define the *Trust Game*, which measures trust with a simple two-player economic protocol. We prove the optimal strategies for each participant: the informant will accept the game if and only if she has a high trust in the target, and the target will cooperate if and only if she has a high trustworthiness towards the informant.
- We define a more sophisticated game, called the *Sybil Game*, that includes the economic benefit to the detective of learning of Sybils and the economic cost to informant and target of revealing that Sybils are present. We prove the optimal strategies for each participant. The detective will offer the game if and only if it will determine her choice about using the application in which these identities participate. The informant will accept the game if and only

if she is Sybil with a low opportunity cost, and the target will cooperate if and only if she is identical to the informant. Both games are presented in Section 3.

- We propose the *Informant* protocol for detecting Sybils, based on our Sybil Game, in Sections 4 and 5. Informant uses a Dutch auction to find the minimum possible reward that will still reveal a Sybil attacker.

We present the background and related work in Section 2.

Our contributions are part of a growing set of results that apply economic games to security and p2p applications, including distributed hash table applications [13], multicast applications [14], file-sharing applications [15], anonymous communication systems [16], the Sybil attack [17], and digital rights management [18]. Because these emerging approaches rely on incentives and other economic mechanisms, they are not as exact as cryptographic mechanisms. However, they are able to address many problems that have resisted traditional cryptographic solutions.

## 2 Background

In this section, we discuss the prior work on incentive systems, define the Sybil attack, and discuss prior work on discouraging and detecting Sybil attacks.

**Incentives.** Many researchers have evaluated the behavior of rational participants in peer-to-peer networks, with a focus on the problems of free-riding. Shneidman and Parkes [19] give evidence that participants in p2p networks behave rationally. Our own work on incentive systems includes SPIES [20, 18], an incentive system for digital rights management, and an analysis of the incentives for malicious Sybil attacks [17].

**Sybils.** Our formal definitions of the Sybil attack are adapted from Douceur [1]. *Entities* are economically rational agents that control one or more *identities*, which are the actors within a network protocol. Identities can send messages to each other through pipes connected to a *communications cloud* that obscures link-level information. Messages received by an identity are communicated to its controlling entity out-of-band. We assume that it is not possible to eavesdrop on these Identity to Entity communications.

**Sybil Deterrence.** The most popular approaches to Sybil deterrence are resource testing and trusted certification, both of which are discussed in Douceur [1]. Other methods include reputation systems, task verification, temporary identities, recurring payments, and per-resource payments.

A full discussion of these methods is beyond the scope of this paper, but none of them are completely effective at eliminating Sybil attacks. Resource testing is ineffective for most systems, and trusted certification is usually too expensive [1, 8, 9]. Most reputation systems can be easily subverted by Sybil attackers, and those that cannot are less informative and little used in practice [21, 4, 22]. Task verification is only applicable for a small subset of applications [23]. Temporary

identities, recurring payments, and per-resource payments require either the use of electronic cash or of significant human effort [24]. Since deterrence is not a completely solved problem, we believe that Sybil Detection is a critical part of a layered defense.

**Sybil Detection.** The Sybil detection research has focused on direct observation, Douceur’s term for link-level or application-level observations dependent on a weakly anonymizing communications cloud. This type of detection is applicable to Sybil attacks on specific types of applications, such as sensor networks. Newsome et al [8] suggest active position verification as a method of direct observation in sensor networks. Of course, two independent sensors may be close together, but it is unlikely that a large number will be consistently close together. In our previous work [25], we suggest a passive approach to detecting sybil attacks in wireless networks.

In the more general case, Kohno, Broido, and Claffy [26] use clock skews to correlate messages originating from a single computer. This method has been used successfully against previous versions of *honeypd* (<http://www.honeypd.org>), a utility for making honeynets, a special class of Sybil identities. However, more recent versions of *honeypd* resist clock skew classification, and it appears that, in general, knowledgeable and determined Sybil attackers can defeat clock skew analysis.

Direct observation focuses on devices rather than common control, so it is ineffective against attackers who, for example, are able to acquire a geographically diverse set of zombie machines. Our Sybil detection protocol does not rely on direct observation, but on the economic unity of Sybil identities, so its areas of application depend on the specific utilities of the participants rather than on the protocol details.

The work most similar to ours concerns attacks on auction systems. Yokoo et al [27] discuss the Sybil attack in combinatorial auctions, but the results are not generalizable to other applications. Rubin et al [28] use techniques similar to ours to find collusion in eBay auction records. Their work differs in that they do not address the general problem of Sybil detection, and they are focused on passive, rather than active, Sybil detection.

### 3 The *Trust* and *Sybil* Games

We introduce our solution to the problem of Sybil detection in three steps. First, in this section, we analyze the optimal strategies of the simple Trust game that we introduced in Section 1. Second, later in this section, we introduce the more realistic Sybil game, which extends the Trust game to take into account the self-interests of the detective and informant. Third, in the next section, we use the Sybil game as the core mechanism of a protocol for Sybil detection.

We define trust in terms of the conflict between self-interest and cooperation in an economic game involving two entities, an informer and a target. These entities may be identical, as in the case of a Sybil attack, or they may be independent. We discuss distinct, collaborating identities to Section 3.3.

### Trust Game

1. The informer chooses whether to *quit* or to *play*.
  - If *quit*, the game ends with a net profit of \$0 to both players.
  - If *play*, she gives a security deposit of \$ $f$  to the detective, identifies the target, and the game proceeds.
2. The target receives a payment of  $\$(2r + f)$ . If the informant and target are part of a Sybil, the Sybil attacker has made  $\$2r$ .  
If they are independent, then the target chooses whether to *keep* or *share* the money.
  - If *keep*, she earns  $\$(2r + f)$  (and the informant loses \$ $f$ )
  - If *share*, she and the informant both earn \$ $r$ .

**Fig. 3.** The Trust game.

### 3.1 The Trust Game

The Trust game, shown in Figures 1 and 2, measures the relationship between two identities with a simple economic protocol. The possible outcomes are *quit*, *(play, cooperate)*, and *(play, defect)*; the outcome that is reached depends on the trust the informer has in the target and the trustworthiness of the target. A formal definition of the game is in Figure 3.

**Optimal Strategies in the Trust Game.** Rational agents in the Trust game will act to maximize their profit. We assume that both agents are rational and that this is common knowledge, and show that the informer will cooperate when her trust in the target is high, while the target will cooperate when her trustworthiness towards the informer is high. The outcomes of the Trust game given  $f$  and  $r$  therefore provide information about the relationship between the two entities involved. We now evaluate the players' strategies formally.

**Theorem 1.** *In the Trust Game, the behavior of rational identities is as follows:*

- (i) *If the target is independent from the informant, she will choose to keep the money. (If they are identical, she has no choice to make.)*
- (ii) *A Sybil identity will play the game and name another Sybil identity.*
- (iii) *An identity will not name an independent identity.*

*Proof.* We use the common game theoretic technique of backwards induction [29] to determine the strategies of the players in the Trust game. The target makes  $\$(2r + f)$  if she keeps the money, and \$ $r$  if she shares it; so a rational target will keep it. This establishes (i). A Sybil identity makes  $\$2r$  if she names another Sybil identity, and \$0 if she does not play. So a rational Sybil attacker will choose to play. This establishes (ii). To show (iii), note that the informant is aware of the target's rationality, and so can predict that the target will keep the money. So if she names an independent target, she expects to lose \$ $f$ , versus a profit of \$0 for not playing, or  $\$2r$  for naming part of a shared Sybil. Therefore, neither Sybil identities nor ordinary identities will name an independent identity.  $\square$

### Sybil Game

1. The detective chooses whether to *offer* the protocol or *quit*.
  - If the detective quits, the game ends. The net profit to each participant is \$0.
  - Otherwise, the game continues.
2. The informant chooses whether to *reject* the game or *play*.
  - If *rejects*, the game ends and the net profit to each participant is \$0.
  - If *play*, she gives a security deposit of \$ $f$  to the detective, identifies the target. The target receives a payment of  $\$(2r + f)$ .
3. If the target is identical to the informant, the game ends with net profits of:  $\$(b - 2r)$  for the detective;  $\$(2r - c)$  for the sybil.  
Otherwise, the target chooses whether to *keep* the money or *share* it.
  - If *keep*, the net profits are:  $\$(b - 2r)$  for the detective;  $\$(-f - c)$  for the informant;  $\$(2r - c)$  for the target.
  - If *share*, the net profits are:  $\$(b - 2r)$  for the detective;  $\$(r - c)$  for the informant;  $\$(r - c)$  for the target.

Fig. 4. The Sybil game.

The actions of the informant thus signal its type: Sybils will play the game, naming another Sybil identity, and ordinary identities will not (since they could only name independent identities. Collaborators may or may not play the game; they are discussed in Section 3.3.

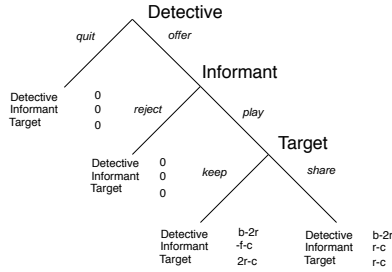
### 3.2 The Sybil game

The Sybil game is an extension of the Trust game that includes the economic benefit to the detective of learning of Sybils and the economic cost to Sybils of revealing their relationship.<sup>1</sup> The Sybil game is again between three players: the detective, who offers the game, an informer, and a target. As before, the informer and target may be identical. The game has four variables:  $f$  the security deposit;  $r$  the reward;  $b$  the detective’s monetary benefit for learning about a Sybil relationship;  $c$  the attacker’s cost for revealing a Sybil attack. The values  $r$  and  $f$  are known to all participants, while the others are private. The steps of the Sybil game are defined in Figure 4.

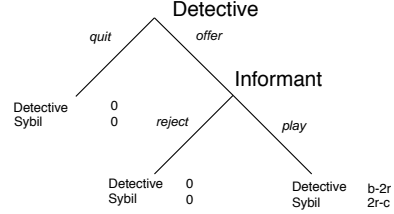
In some cases  $b$  and  $c$  could be related, but this does not change the analysis of a particular instance of the game, since neither the detective or the attacker is able to affect  $b$  or  $c$  by their choices.

The game has the four outcomes *quit*, *(offer, reject)*, *(offer, play, keep)*, and *(offer, play, share)*. We now consider under which conditions each strategy is an equilibrium.

<sup>1</sup> A reasonably careful Sybil attacker can both reveal Sybil identities and continue to participate in the application, so the cost to Sybils is in additional precautions that application users may take, not in being excluded from the application or prosecuted.



**Fig. 5.** Utilities in the Sybil game when the identities are independent.



**Fig. 6.** Utilities in the Sybil game when the identities are part of a Sybil attack.

**Strategies in the Sybil game.** The strategies for the Sybil game are similar to those of the Trust game for the informer and target. The detective is best-off offering the game whenever her benefit exceeds the rewards she pays out.

**Theorem 2.** *In the Sybil Game, the optimal strategies for rational players are as follows:*

- (i) *The detective will offer the game when  $b \geq 2r$ , and to quit otherwise.*
- (ii) *The informant will accept the game if it is a Sybil with  $c \leq 2r$ , and reject it otherwise.*
- (iii) *If the target is independent from the informant, it will choose to keep the money.*

*Proof.* First note that parts (ii) and (iii) are exactly as in Theorem 1. For (i), note that the detective expects she will receive \$0 if she quits, and somewhere between \$0 and  $b - 2r$  if she offers the game. So she will offer the game as long as  $b \geq 2r$ .  $\square$

### 3.3 Collaborators

Collaborators may act like independent entities or like Sybil identities in the Trust and Sybil games, depending on the level of trust between them. Suppose that the informant and the target are collaborators, and they have agreed to share the money received from the detective. When she is considering whether to name the target, the informant must consider how likely the target is to share the money received. Let  $p$  be her estimation of the probability of sharing, let  $r$  be the reward, and let  $f$  be the security deposit. Then naming the target is more profitable than not when  $pr - (1 - p)f > 0$ , so the informant will play the game and name the target when  $f < \frac{p}{1-p}r$ . Collaborators also may have a much higher opportunity cost than Sybils in revealing their identities, since these identities may be easier to link to the collaborators' real-world information.

Because of this additional complexity, we focus on Sybils and independent entities in this paper. However, the Informant protocol, which we define in the next section, will also detect collaborators if the security deposit is set low enough; it can be tuned to detect only Sybils by setting a high security deposit.

## 4 The Informant Protocol

In this section, we present Informant, a protocol to detect Sybils. The protocol is based on an extension of the Sybil Game. Rather than using a fixed reward  $r$ , it uses a *reverse Dutch auction* to determine the minimum possible reward that will still reveal a Sybil attacker. This benefits the detective financially without reducing the number of Sybils detected.

Informant has several positive characteristics. It reveals a Sybil attacker when her aversion to being detected is less than the detective’s interest in detecting her. It benefits the detective when the possible presence of highly-averse Sybil attackers does not deter her from participating in the underlying peer-to-peer application. These properties are demonstrated in our analysis of the protocol in Section 5.

### 4.1 Assumptions

Informant is built on several assumptions: that entities are rational, that an application has a recurring join cost, and that the questioner is trusted. Each assumption is discussed in greater detail below.

**Rational Entities.** We assume that attackers and other entities participating in the application are rational in the economic sense. In other words, we assume that entities have goals and beliefs, and that they act according to their beliefs in order to achieve these goals. This assumption excludes “attacks” from malfunctioning client software or user error. Entities need not be specific human beings; they may also be corporate or criminal entities that are sufficiently organized to act in an economically rational way.

Following traditional game theory, we do not assume that an attacker’s *goals* are reasonable. An attacker’s goals could include disfiguring websites or feeding false information, based on personal conflict, basic malice, or capriciousness. Making money is a basic goal that we assume all entities participating share. Again following traditional game theory, we assume that some specific monetary value can be assigned to each of an attacker’s goals.

**Recurring per-identity join cost.** We require a specific and recurring per-identity cost to participating in the distributed application. This cost can come in the form of a monetary entry fee, a CAPTCHA [30] solution requirement, proof of receipt of an SMS message, or some other form. Each entity with an identity in the protocol should be able to verify (at least manually) that another identity has borne this entry cost. This can be done, for example, by having each identity participate in generating challenges such as CAPTCHAs in a challenge round, or by using a trusted “payment certification” authority.<sup>2</sup>

We do not require a specific method, and we do not further discuss entry fee verification. However, the authors have available an extensive analysis of using recurring per-identity join costs as a method of limiting Sybil attacks [17]. Unlike

---

<sup>2</sup> Such an authority requires much less trust than an identity certification authority that needs to gather non-electronic information and deal with key-distribution problems.

non-recurring fees, our results show that the difficulty of launching a Sybil attack increases with the number of other identities present in the application. In fact, many researchers rely on recurring per-identity costs (often in the form of limited resources) to limit the effectiveness of Sybil attacks [31, 3, 32, 24, 33–35].

**Trusted Querier.** Since our Sybil detection protocol allows the querier to cheat the respondents, the querier must be trusted to follow the protocol correctly. Since the querier does not need to be anonymous, and thus can be held accountable for her actions, this assumption is reasonable in many cases. Nevertheless, this does limit the scenarios in which Informant can be used.

## 4.2 Protocol Details

The variables and notation we use to describe our detection protocol are summarized below.

- $i$  The ID number of a particular instance of the protocol.
- $\tau$  The time between rounds.
- $r_0$  The initial reward for established Sybil identity claims.
- $r$  The current reward for established Sybil identity claims.
- $n$  A nonce used to ensure freshness.
- $f$  The security deposit required of the informant.

If possible, some discrete round of the distributed application where each identity receives some service should complete before Informant starts. This approach minimizes the opportunity cost to attackers in revealing some of their Sybil identities because any Sybil attack during the service phase will have already succeeded or failed. In the next round of the distributed application, the entity operating the Sybil attack can use all new identities, so knowledge of attacker identities in the current application incarnation will not prevent the attacker from introducing identities in the next incarnation. If the distributed application does not support discrete phases, Informant can still be used, but the opportunity costs to attackers will be higher.

In the protocol description below,  $\$x$  represents an electronic payment of  $\$x$  dollars usable by the message recipient<sup>3</sup>.  $A \rightarrow * : m$  represents the broadcast of message  $m$  to all distributed application participants. This broadcast is assumed to be reliable.

Informant requires each participating identity  $A$  to have a public/private key pair, designated  $K_A^+$  and  $K_A^-$ , respectively. In contrast to the key pairs used in many protocols, these are not part of any PKI and cannot be used to establish an entity’s true identity; in particular, nothing prevents a single entity from computing or acquiring an arbitrarily large number of new pairs for each round of the application. We assume the application that Informant supports has a method of obtaining or exchanging each identity’s public key.

Informant is broken into a number of steps. Figure 7 shows the control flow of the protocol.

<sup>3</sup> We do not require any specific form of electronic payment; it does not need to be anonymous electronic cash, although this is an acceptable form.

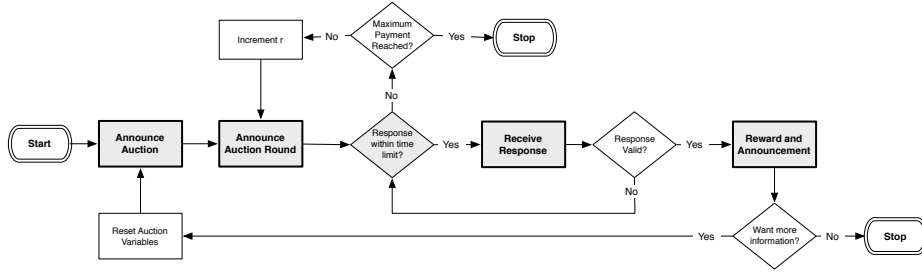


Fig. 7. Steps in the Sybil Discovery protocol.

1.  $Q \rightarrow * : [i, \tau, f]_{K_Q^-}$   
*Announce Auction.* The detective  $Q$  announces a new auction and its parameters as a signed message.
2.  $Q \rightarrow * : [i, f, r, n]_{K_Q^-}$   
*Announce Auction Round.* The detective sets the current price  $r = r_0$ . She announces to all application participants that she is willing to pay  $\$r$  for knowledge of a single Sybil relationship, and includes nonce  $n$ .
3. If no responses are received in  $\tau$  secs,  $Q$  increments  $r$  by any desired amount, chooses a new nonce  $n$ , and reruns Step 2. If  $r = b/2$ , the protocol ends — nothing further can be learned.
4.  $A \rightarrow Q : [n, \$f, A, B]_{K_A^-}$   
*Receive Response.* Otherwise, a claim of common control between  $A$  and  $B$  and a payment of  $\$f$  has been received. Only the first valid message received is valid. The use of the nonce  $n$  prevents identities from sending messages before the round is announced.
5.  $Q \rightarrow B : \$(f + r)$   
*Reward Payment.* The detective sets up an authenticated channel with the candidate Sybil identity  $B$ , and pays her  $\$(f + r)$ .
6.  $Q \rightarrow * : [i, A, B]_{K_Q^-}$   
*Announcement.* The detective broadcasts to all application participants the claim of common control between  $B$  and  $C$ . This announcement prevents the claim from being sold to multiple detectives, as well as providing valuable information to the application participants. This broadcast includes the auction ID  $i$ .

At the end of Step 6,  $Q$  has good reason to believe in a common control (Sybil) relationship between the identities  $A$  and  $B$ . Optionally, the protocol could be amended to include in Step 4 a hashed nonce supplied by  $A$ ; before the reward is paid,  $Q$  would require  $B$  to provide a signed copy of the nonce and  $n$ . This addition would prevent false claims if that was a concern.

If  $Q$  still wants to learn more information about Sybils present, she resets  $i$ ,  $\tau$ , and  $r_0$ , and restarts the protocol. Since common control is transitive, over multiple runs the initiator can learn about larger Sybil groups of identities.

Note that a Sybil attacker does not know whether other Sybils are present, nor what the questioner’s maximum possible payment  $b/2$  is. It is therefore in her interest to respond soon after  $r$  exceeds her opportunity cost  $t$ .

## 5 Sybil Detection Protocol Analysis

In this section, we analyze Informant in economic terms, giving the best strategies for the various participants and validating the claims of usefulness of Section 4. We also discuss the problem of opportunistic Sybils and how they can be avoided.

To facilitate analysis, we make several simplifying assumptions. First, we assume that there is either one entity attacking, with probability  $\gamma$ , or none, with probability  $1 - \gamma$ . We exclude situations with multiple Sybil attackers. Second, we assume that a Sybil attacker belongs to one of just two classes. Attackers in a *low-cost* class reveal themselves if offered a reward of  $\$c < 2r$ , where  $2r$  is the maximum reward the detective is willing to pay. Attackers in the *high-cost* class reveal themselves if offered a reward of  $\$C > 2r$ . Attackers belong to the low-cost class with probability  $1 - \delta$  and to the high-cost class with probability  $\delta$ . This simplifies the analysis, but does not significantly alter the results. Third, we assume that the informant is always able to guess the maximum reward that will be offered, so that the target always receives the maximum payment of  $\$(2r + f)$ . This is a conservative assumption. Finally, we assume that the detective’s interest in running the Informant protocol is to decide whether to stay in the underlying application or to leave it. If the detective instead makes some other choice based on the outcome of the tests, our analysis applies so long as the choice has the same incentive structure.

We represent the presence or absence of a Sybil in the underlying application, and the Sybil’s type if present, as a choice of the class of the informant, made randomly by a player representing the role of nature<sup>4</sup>; non-sybil with probability  $(1 - \gamma)$ , low-profit Sybil with probability  $\gamma(1 - \delta)$ , and high-profit Sybil with probability  $\gamma\delta$ . This type of game is called a *signaling game* [29]; the informer, by accepting or rejecting the game, in effect signals her (otherwise unobservable) type to the detective.

The game tree, player utilities, and equilibrium strategies are shown in Fig. 8.

**Theorem 3.** *For rational Detectives, Informants, and Targets,*

- (i) *If the detective runs Informant, she will stay if there are no Sybils reported, and leave if there is a Sybil reported.*
- (ii) *Low-cost Sybils will play the game and announce themselves. Non-Sybil identities and high-cost Sybils will not play the game.*

<sup>4</sup> The somewhat counterintuitive use of a “nature” player is a standard game-theoretic technique.

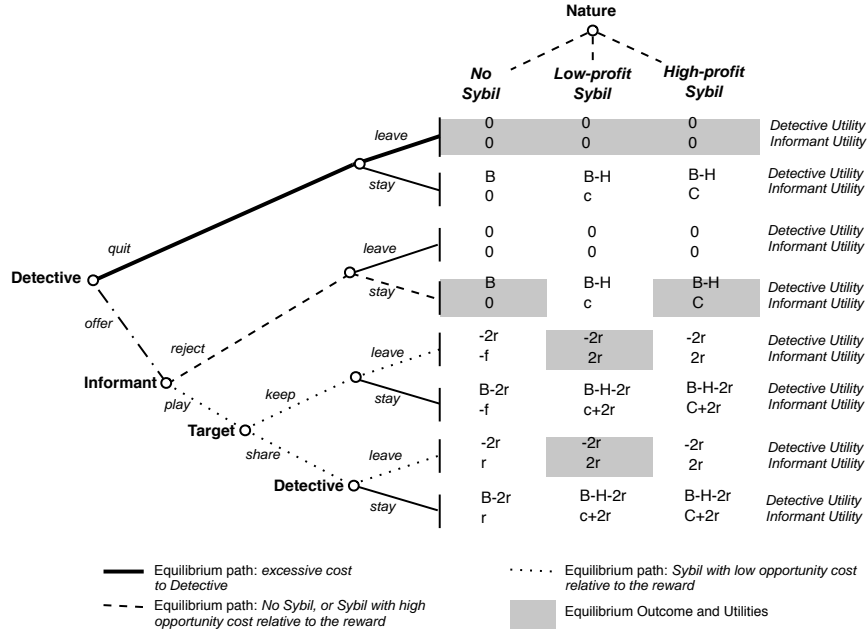


Fig. 8. Equilibria in Informant

(iii) The detective will run Informant unless  $(1 - \delta + \delta\gamma)B < \delta\gamma H + 2\delta(1 - \gamma)r$  where  $\delta$  is the probability of a Sybil;  $\gamma$  the probability a Sybil is high-cost (and therefore highly averse to detection);  $B$  is the benefit of the protocol; and,  $H$  is the harm from the Sybil. Otherwise she will leave the application.

In other words, Informant is useful so long as the expected harm from high-profit Sybils ( $\delta\gamma H$ ) does not greatly exceed the benefit  $B$  provided by the application.

*Proof.* Parts (i) and (ii) are directly implied by the utilities in Figure 8: any change of strategy from the shaded Nash equilibria results in a lower utility for either the detective or the informant.

Given (i) and (ii), the detective's expected utility from running Informant is  $B$  if there are no Sybils,  $-2r$  if there is a low-cost Sybil, and  $B - H$  if there is a high-cost Sybil that has a high aversion to detection. The probabilities of these outcomes are  $(1 - \delta)$ ,  $\delta(1 - \gamma)$ , and  $\delta\gamma$ , respectively, so the detective's expected utility when running Informant is  $(1 - \delta)B + \delta(1 - \gamma)(-2r) + \delta\gamma B$ , which is positive when  $(1 - \delta + \delta\gamma)B > \delta\gamma H + 2\delta(1 - \gamma)r$ . This establishes (iii).  $\square$

## 5.1 Opportunistic Sybils

We have evaluated Informant under the assumption that the Sybil attacker has a malicious interest in the underlying application. However, careless use of our Sybil detection protocol could induce *opportunistic Sybils*, which are attackers

that have no interest in the application itself; instead, they participate and form a Sybil in the hopes of being paid to reveal their presence. To avoid this problem, detectives must run the protocol with a frequency and unpredictability such that the cost of maintaining a Sybil exceeds an opportunistic Sybil’s profit.

We denote the cost to enter a single identity in the application  $e$ . We assume that a non-malicious entity’s total utility after entering a single identity — the behavior of an honest user — is greater than zero. Denote this initial utility  $u$ . Since she is not malicious, she gains nothing from additional identities other than the chance of a reward.

First, we evaluate an application in which the Informant protocol is run every round. We assume conservatively that the attacker is certain that she will be the winning respondent in each auction round. For each identity the attacker adds, she pays  $e$  and expects to receive a reward of  $r$  after participating in the protocol. So if she enters two identities into the application she expects a total utility of  $u - e + r$  versus  $u$  if she only enters one identity. Thus it is not profitable for non-malicious entities to form Sybils when  $u - e + r \leq u$ , that is when  $r < e$ , the offered reward is less than the per-identity participation cost.

The detective can avoid additional Sybils by keeping her rewards this low, but she will not be able to detect some Sybils in this case. She can increase her ability to detect the overall prevalence of Sybil attacks by participating only occasionally, but offering higher rewards when she does so. If the maximum reward the questioner is willing to offer is  $r_{\max}$ , she can avoid introducing non-malicious Sybils by randomly running the protocol at most every  $r_{\max}/e$  rounds, so that the expected per-round reward remains less than  $e$ . Even though the protocol is run less frequently, it is still worthwhile for legitimate Sybils to answer when it is run.

If a number of independent questioners are running Sybil detection protocols, a questioner who does not want to increase Sybil attacks should make sure that the *total* expected per round reward does not exceed  $e$ .

Note that in each case, running the Sybil detection protocol may increase the prevalence of Sybil attacks made by weakly malicious entities. If an attacker gains utility greater than zero, but less than  $e$ , from an additional identity introduced into the application, then a reward less than  $e$  may be enough to make the total per-round return positive. This tendency can be avoided by keeping the per-round reward as low as possible.

## 5.2 Discussion

**Legal Concerns.** If Sybil attackers are engaged in illegal activity, they may be reluctant to participate in Informant for fear that Sybil identities could be linked to their real-world identity. If this proves to be a concern, identities can be registered using an anonymity system such as Tor [10] and paid with anonymous electronic cash.

**Electronic Cash.** Informant relies on some form of anonymous payment. The lack of practical anonymous electronic cash is the most significant obstacle to implementation.

**Entry Fees versus Rewards.** In general, the designer of a network application has two competing interests: to maximize participation in the protocol, which requires keeping entry fees low, and to maximize Sybil detection, which requires setting a high reward and thus a high entry fee as well (to avoid opportunistic Sybil attacks.) Where entry fees are low and rewards must be high, the detective can choose to offer Informant unpredictably and only very occasionally. Informant must only be announced after all identities have registered for a particular round of the application, so that opportunistic Sybil attackers must participate and pay the entry fees each round in the hopes of receiving the reward eventually.

## 6 Conclusion

We have designed and analyzed a novel, economic approach to Sybil attack detection protocol called *Informant*. We have proven the optimal strategies for each participant. The informant will accept the game if and only if she is Sybil with a low opportunity cost, and the target will cooperate if and only if she is identical to the informant. Our use of a Dutch auction ensures the minimum possible reward that will still reveal a Sybil attacker. While previous approaches have focused on physical tokens, such as radios [8, 25] or clock skew [26], our approach is more general and not limited to a specific application. Given that the Sybil attack is not preventable without centralized verification of unique identity — all but impossible on a large scale — detection is crucial for protecting p2p applications.

## References

1. Douceur, J.: The Sybil Attack. In: Proc. Wkshp on P2P Systems (IPTPS). (2002)
2. Mathur, G., Padmanabhan, V.N., Simon, D.R.: Securing routing in open networks using secure traceroute. Tech Rep MSR-TR-2004-66, Microsoft Research (2004)
3. Castro, M., Druschel, P., Ganesh, A.J., Rowstron, A.I.T., Wallach, D.S.: Secure routing for structured peer-to-peer overlay networks. In: OSDI. (2002)
4. Kamvar, S.D., Schlosser, M.T., Garcia-Molina, H.: The Eigentrust algorithm for reputation management in P2P networks. In: Proc. WWW Conf. (2003) 640–651
5. Jelasity, M., Montresor, A., Babaoglu, O.: Towards Secure Epidemics: Detection and Removal of Malicious Peers in Epidemic-Style Protocols. Technical Report UBLCS-2003-14, University of Bologna (2003)
6. Levien, R.L.: Attack Resistant Trust Metrics. PhD thesis, UC Berkely (2004)
7. Perrig, A., Stankovic, J., Wagner, D.: Security in wireless sensor networks. Commun. ACM **47**(6) (2004) 53–57
8. Newsome, J., Shi, E., Song, D., Perrig, A.: The Sybil attack in sensor networks: analysis & defenses. In: Proc. IPSN Intl Symp. (2004) 259–268
9. Karlof, C., Wagner, D.: Secure routing in wireless sensor networks: Attacks and countermeasures. Ad hoc Networks Journal (Elsevier) **1**(2–3) (2003) 293–315
10. Dingledine, R., Mathewson, N., Syverson, P.: Tor: The second-generation onion router. In: Proc. USENIX Security Symposium. (2004)
11. Cox, L., Noble, B.: Pastiche: Making backup cheap and easy. In: Proc. USENIX Symposium on Operating Systems Design and Implementation. (2002)

12. Adar, E., Huberman, B.A.: Free riding on gnutella. *First Monday* **5**(10) (2000)
13. Ntarmos, N., Triantafyllou, P.: SeAl: Managing Accesses and Data in Peer-to-Peer Sharing Networks. In: *Proc. P2P Computing*, August (2004) 116–123
14. Ngan, T.W.J., Wallach, D.S., Druschel, P.: Incentives-compatible peer-to-peer multicast. In: *Proc. P2PEcon Workshop*. (2004)
15. Anagnostakis, K., Greenwald, M.: Exchange-Based Incentive Mechanisms for Peer-to-Peer File Sharing. In: *Proc. ICDCS*. (2004)
16. Acquisti, A., Dingedine, R., Syverson, P.: On the Economics of Anonymity. In: *Proc. Financial Cryptography (FC)*, Springer-Verlag, LNCS 2742 (2003)
17. Margolin, N.B., Levine, B.N.: Quantifying and discouraging sybil attacks. *Tech Rep 2005-67*, University of Massachusetts Amherst (2005)
18. Margolin, N.B., Wright, M., Levine, B.N.: Analysis of an incentives-based protection system. In: *Proc. ACM Digital Rights Management Workshop*. (2004)
19. Shneidman, J., Parkes, D.C.: Rationality and self-interest in peer to peer networks. In: *Proc. Intl Wkshp on Peer-to-Peer Systems (IPTPS)*. (2003)
20. Margolin, N.B., Wright, M., Levine, B.N.: SPIES: Secret Protection Incentive-based Escrow System. In: *Proc. P2PEcon Workshop*. (2004)
21. Cheng, A., Friedman, E.: Sybilproof reputation mechanisms. In: *Proc. P2PEcon Workshop*. (2005) 128–132
22. Čapkun, S., Hubaux, J.P.: BISS: building secure routing out of an incomplete set of secure associations. In: *Proc. ACM Wireless Security Conf.* (2003) 21–29
23. Srivatsa, M., Liu, L.: Vulnerabilities and security threats in structured overlay networks: A quantitative analysis. In: *Proc. ACSAC*. (2004) 252–261
24. Awerbuch, B., Scheideler, C.: Group Spreading: A Protocol for Provably Secure Distributed Name Service. In: *Proc. Automata, Languages and Programming*. (2004) 183–195
25. Piro, C., Shields, C., Levine, B.N.: Detecting the Sybil Attack in Ad hoc Networks. In: *Proc. IEEE/ACM SecureComm*. (2006)
26. Kohno, T., Broido, A., Claffy, K.C.: Remote physical device fingerprinting. *IEEE Trans. Dependable Sec. Comput.* **2**(2) (2005) 93–108
27. Yokoo, M., Sakurai, Y., Matsubara, S.: The effect of false-name bids in combinatorial auctions. *Games and Economic Behavior* **46**(1) (2004) 174–188
28. Rubin, S., Christodorescu, M., Ganapathy, V., Giffin, J.T., Kruger, L., Wang, H., Kidd, N.: An auctioning reputation system based on anomaly. In: *Proc. ACM conference on Computer and Communications Security*. (2005) 270–279
29. Osborne, M.J., Rubinstein, A.: *A Course In Game Theory*. MIT Press (1994)
30. von Ahn, L., Blum, M., Hopper, N., Langford, J.: CAPTCHA: Using hard AI problems for security. In: *Proc. of Eurocrypt*. (2003) 294–311
31. Nielson, S.J., Crosby, S.A., Wallach, D.S.: A taxonomy of rational attacks. In: *Proc. of IPTPS*. (2005)
32. Cornelli, F., Damiani, E., Samarati, S.: Implementing a reputation-aware gnutella servant. In: *Proc. of Intl Workshop on Peer to Peer Computing*. (2002)
33. Marti, S., Garcia-Molina, H.: Limited reputation sharing in p2p systems. In: *Proc. of the 5th ACM conference on Electronic commerce*. (2004)
34. Maniatis, P., et al.: Preserving peer replicas by rate-limited sampled voting. In: *Proc. ACM SOSP*. (2003) 44–59
35. Vishnumurthy, V., Chandrakumar, S., Siler, E.G.: KARMA: A secure economic framework for p2p resource sharing. In: *Proc. P2PEcon Workshop*. (2003)