Fake VIP Attacks and Their Mitigation via Double-Blind Reputation

Jerzy Konorski
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology
Gdansk, Poland
jekon@eti.pg.gda.pl

***Abstract*—In a generic setting subsuming communication networks, resource sharing systems, and multi-agent communities, a client generates objects of various classes, to which a server assigns class-dependent service quality. We identify a class of *Fake VIP attacks* as false declarations of a high class to acquire undue service quality, with an awareness that a defense via object signature detection is costly and so invoked reluctantly. We show that, unexpectedly, such attacks can be mitigated by a double-blind reputation scheme at the server side. We offer a minimum-information framework for Fake VIP attacks and a stochastic analysis of a two-player Stackelberg game to find optimum attack and defense strategies, as well as to identify regions of operation where both the client and the server find the reputation scheme beneficial.**

Keywords—service provision; Fake VIP attack; signature detection; reputation; Markovian analysis; Stackelberg game

I. INTRODUCTION

In distributed multi-agent systems, interactions between agents can be modeled in the Client-Server paradigm: Client requests service of some quality level, which Server grants. To reflect the workings of today's communication networks or resource sharing systems, this paradigm can be extended to include Sender and Relay. Client generates *objects* (e.g., packets, queries, or transactions) of various *intrinsic classes*; the latter determine the service quality the object is entitled to. For simplicity, we only distinguish two intrinsic classes, *low* (L) and *high* (H). Sender, acting on behalf of Client, appends intrinsic class-dependent information to each object (e.g., suitable metadata) and passes the object to Relay as a request for service. The appended information, called *declared class*, is used by Relay to decide the *assigned class* for the object, i.e., the service quality to be granted; subsequently, Relay passes the object on to an appropriate Server. Thus Relay acts on behalf of Server and embodies, e.g., an intermediate network node with a routing scheme, a load balancer, a front-end processor performing service discovery etc.

When service quality differentiation is supported, a serious threat to system security and performance are *usurpation* attacks, where Sender declares a higher class than an object's intrinsic class to ensure better service quality for Client. This imposes undue effort (service level) upon Server, whose interests Relay should protect. Consider a packet network where a source node (Sender) appends voice/video headers to best-effort packets generated by a local application (Client). The packets then enjoy priority queuing/medium access before onward transmission at the nearest neighbor node (Relay) and at all subsequent nodes (Servers) en route to destination, instead of non-priority queuing/medium access that these packets are entitled to. An example is the *traffic remapping attack* (TRA) [1] in wireless networks employing the IEEE 802.11 MAC protocol in Enhanced Distributed Channel Access (EDCA) mode [2]. In a distributed resource sharing system, a similar attack can be launched by an intelligent terminal (i.e., Sender) issuing a transaction request on behalf of Client, by falsely stating the urgency or nature of the transaction. In the user-centric Social Internet of Things [3], an entity may pretend to be a "trusted friend" and request a higher level of service from another entity within reach.

To defend against usurpation attacks, Relay should infer the intrinsic class of an arriving object, e.g., via a *signature detection* scheme. A *signature* is an abstraction of intrinsic class-dependent features of an object that Sender cannot modify, e.g., packet length and/or data content, client's credentials and/or contextual information in a query, transaction security data etc. A classical defense approach consists in tight access rights control that involves checking the detected signature against a trusted database. However, signature detection may be costly: in packet networks it amounts to Deep Packet Inspection, e.g., using pattern matching [4], which is hard to perform online at high transmission rates; in high-volume transaction processing systems frequent communication with a remote trusted database would foster devastating denial-of-service attacks. Therefore, Relay may be reluctant to invoke signature detection and Sender may hope for an attack to go undetected. This is why we use here the term *Fake VIP attack* to conjure up a cheeky impostor whose legitimacy no one dares to check lest they run into trouble.[1] Besides being costly, signature detection in general is also imperfect: voice/video packets are typically short, but so can be best-effort packets; client's credentials may remain the same even though successive objects have different intrinsic class etc.

[1] One particularly daring Fake VIP attack was portrayed in Nikolai Gogol's *The Government Inspector*; another was actually launched in the famous historical episode known as "the Captain of Koepenick."

Several postulates aggravate the problem of Fake VIP attacks and outline the *minimum-information framework* of our subsequent analysis:

(i) launching a Fake VIP attack is costless for Sender (unless payments are imposed for merely requesting high service quality, which is often impractical),

(ii) Relay has no means of learning an arriving object's intrinsic class: the declared class may result from a Fake VIP attack, and signature detection is imperfect,

(iii) Relay's decision whether to invoke signature detection in general cannot be conditioned upon the demanded class, since the latter may have been incorporated into the signature at Sender, or only be revealed after the decision has been made,

(iv) providing high-quality service is costly for Server and not necessarily accompanied by payment to Relay or Server,

(v) Sender cannot learn the class assigned at Relay—her (and Client's) perception of service quality only forms across a long sequence of class assignments to successive objects,

(vi) Relay is not allowed to cheat on class assignment, i.e., assign a low class when detected signature indicates a high intrinsic class, since it could jeopardize the mission of the system.

In our TRA example, postulate (i) is obvious given that a Fake VIP attack consists in simply substituting a false packet header. Postulate (ii) is due to best-effort packets sometimes having features (such as length, source/destination port or certain bit patterns) characteristic of voice/video packets and *vice versa*. Postulate (iii) arises when Deep Packet Inspection has to be performed in a cut-through fashion, as a packet's bits flow through the node, the demanded class being appended in the packet's trailer. Postulate (iv) is realistic, since priority handling of packets being part of a TRA consumes bandwidth dedicated to other Clients. Postulate (v) is realistic too, since the source node cannot read headers (priority assignments) of packets forwarded by distant en route nodes, and perceived end-to-end performance of an individual packet or session may be misguiding, as it is influenced by encountered congestion, failures, or rerouting. Finally, postulate (vi) reflects the adverse effect of mishandling real-time traffic. In distributed environments other than wireless networks, the above postulates can be justified similarly.

From the system design viewpoint, postulates (i) and (ii) create a powerful incentive for Sender's Fake VIP attacks, which can be launched with impunity and at no cost; Relay can only attempt to *mitigate* their effects, i.e., make them not too beneficial for Client and not too damaging for Server, while keeping the signature detection cost acceptable. However, postulate (iii) prevents easy savings at Relay, e.g., by only invoking signature detection when the demanded class is high; it also implies that Relay cannot easily punish a suspected Fake VIP attack (when the detected signature does not match the demanded class). Postulates (iv) and (v) create a *moral hazard* situation for Relay [5], who might feel incentivized to cheat, e.g., by assigning low-quality service regardless of the declared class or detected signature. However, object-by-object cheating is prevented by postulate (vi). Moreover, in connection with postulate (v) it can be assumed that Sender perceives the *statistical* impact of Relay's assignments; therefore Relay should refrain from long-term cheating as well, as it might raise suspicions in Sender. Clearly, postulate (v) rules out the use of online prediction algorithms [6] by a smart Sender who would attempt to learn the rules behind Relay's assignments and so contrive good attack strategies (unless Sender and Relay collude, in which case the former can compare declared and assigned classes of successive objects).

Postulates (ii), (iii), and (v) constitute a demanding minimum-information framework. Given the little information on Relay's behavior, Sender might reasonably resort to a *probabilistic* attack strategy, whereby a Fake VIP attack on a specific object is launched with an intrinsic class-dependent probability; this probability Sender should optimize based on the statistical expectation of perceived service quality. Given the inability to learn an object's intrinsic class, Relay should find a balance between the cost of frequently providing high-quality service and that of frequently executing signature detection, subject to good enough statistical perception at Sender. However, Relay has too little information to learn Sender's attack strategy − inferring it from the statistics of declared classes and detected signatures would require the knowledge of the statistics of intrinsic class generation, which is Sender's private information.

If the signature detection cost is significant, Relay would rather trust declared class than invoke signature detection, i.e., absorb the damage caused by Fake VIP attacks. On the other hand, a small signature detection cost permits Relay to never trust declared class. However, Relay might reason that Sender will then refrain from Fake VIP attacks, which will render signature detection unnecessary. Thus trusting declared class should be subject to a clever policy at Relay, to which Sender should respond with a clever Fake VIP attack strategy. In this paper we provide a novel design and evaluation framework for a *reputation scheme* at Relay. It helps Relay to decide for an arriving object whether to invoke a signature detection (reducing the risk of undue high-quality service provision) or skip it and trust the declared class (reducing the cost of signature detection). A number of *reputation states* are distinguished and updated on an object-by-object basis, and only the highest one (the *trust* state) permits to skip signature detection. In non-trust reputation states, the comparison of the declared class and detected signature governs reputation state transitions. To thwart sophisticated attack strategies, current reputation state is not revealed to Sender. The operation of the proposed reputation scheme is thus double-blind: it is unable to observe the true behavior of the agent under scrutiny, nor is the agent aware of the inferred reputations or decisions they lead to.

While our double-blind approach may raise doubts as to the term *reputation* (since by standard definitions, reputations should be publicly known [7]), we believe any measure de-

rived from an agent's past behavior and used for trust decisions deserves that name. Note that revealing reputations to agents may not always be rational: it may prompt agents to exploit their current reputations [8] or discover the logic of the reputation scheme to later manipulate it [9], [10].

The contributions of this paper can be stated as follows:

- We identify Fake VIP as a class of usurpation attacks aware of signature detection being costly (and imperfect), and offer a game-theoretic analytical framework to evaluate Sender's and Relay's expected utilities under optimized Sender's attack strategy and Relay's defense.
- We consider a demanding minimum-information framework to address real-world environments where Sender's and Relay's actions are not transparent to each other.
- We design a double-blind reputation scheme at Relay and use Markovian analysis to identify regions of operation where Fake VIP attacks are mitigated and moreover, both Client and Server find the scheme beneficial. Unaware of any other existing mitigation approach in the minimum-information framework, we compare our reputation scheme with a baseline reputation-free scenario.

The rest of the paper is organized as follows. After a brief survey of related work (Section II), in Section III we specify the system model and define the agents' *utilities*. Sender's utility (benefit) reflects the effectiveness of Fake VIP attacks as compared with a system without a reputation scheme; Relay's utility (cost) combines the high-quality service provision and signature detection costs. Expected utilities are derived in Section IV using a simple Markovian analysis. Next, in Section V we consider a *Stackelberg game* [5] between Sender and Relay, in which Relay sets the parameters of the reputation scheme so as to minimize her cost, anticipating that Sender will respond with an attack strategy that maximizes her benefit. That is, the players reach a Stackelberg equilibrium (SE). We show that for a range of signature detection costs, the presence of the reputation scheme improves the expected SE utilities of *both* players: Sender launches Fake VIP attacks with restraint, while Relay trusts declared class quite frequently. We also consider Sender's off-equilibrium (possibly malicious) play. In Section VI we briefly discuss the impact of the reputation scheme. Section VII concludes the paper.

II. Related work

Local and multihop wireless networks face a variety of Fake VIP attack known as the traffic remapping attack (TRA) [1]. The reason is that Deep Packet Inspection (signature detection) [4] is rarely practical. The authors of [11] propose that nodes should report to an access point the traffic priority they claim and truthful reporting is incentivized via Vickrey-Clarke-Groves payments. This amounts to Relay charging Sender for declaring class H. However, a payment scheme may be hard to implement and our solution does not require it. In [12], MAC-layer parameters are configured to improve transmission characteristics of low-priority traffic and so disincentivize claiming high priority for such traffic – in our framework, Relay claims from Servers more equitable treatment of objects with assigned class L relative to class H objects. However, such claims may not be supported by the Relay-to-Server communication protocol or service provision mechanisms at Servers. In an opposite approach [13], nodes can take, or threat to take, punitive measures, e.g., selective radio jamming, against high-priority traffic that Deep Packet Inspection detects as intrinsically low-priority. This would require an extension of our framework whereby class H assigned at Relay could be subsequently double-checked at Server. Likewise, the solution in [1] has nodes broadcast a predefined primitive to signal their dissatisfaction with current service quality, presumably due to some other node's TRA. Translated into our framework, Relay signals her discomfort due to frequent assignment of class H and so threatens Sender to invoke a punishment, e.g., assign class L to subsequent objects even if signature detection advises otherwise. In [14], a traffic source node is supposed to occasionally correct traffic priority claimed by client applications. This amounts to Sender performing signature detection on behalf of Relay instead of acting on behalf of Client.

In this paper we find that reputation can come to an unexpected rescue under Fake VIP attacks. A large body of work on reputation and trust in multi-agent systems exists, cf. [7], [15]. The mission of the proposed reputation scheme is for Relay to learn Sender's behavior under imperfect and occasionally skipped signature detection. In doing so, Relay should not let Sender learn the workings of the reputation scheme; accordingly (and contrary to the common understanding of reputation [7]), she does not reveal Sender's current reputation level. This is in line with studies indicating that a knowledgeable Sender might manipulate the reputation scheme, e.g., through oscillation [9] or whitewashing [16], which sometimes justifies "security by obscurity" [17].

Sender might attempt to learn her present reputation level to see whether she is presently trusted so that a Fake VIP attack will go unpunished. E.g., she might recreate Relay's comparisons of the declared class and signature detection output for recent objects, or she might compare their declared and assigned classes. Both approaches are in the spirit of online prediction [6], a field of study at the junction of machine learning and pattern recognition. A pertinent game-theoretic setting [18] has the learner match the adversary's action (in our context, Sender seeks a coincidence of declared class L and assigned class H). The trouble with any learning approach applied in our framework is that neither the signature detection output nor assigned class is observable to Sender, hence she can only judge her Fake VIP attack strategy by her long-term utility.

In a game-theoretic Intrusion Detection System model (e.g., [19]), a strategic intruder (in our context, Sender) pretends to be the honest type and so attacks with restraint to avoid being detected as the attacker type, whereas a defender (in our context, Relay) does or does not invoke a defensive mechanism (in our context, signature detection) based on Bayesian updating of perceived Sender type. However, Fake VIP attacks cannot be punished due to postulate (vi), hence the risk of revealing her-

self as the attacker does not stop Sender and Fake VIP attack is her dominant strategy.

III. System Model and Players' Utilities

A. System Model

Sender and Relay communicate via a noiseless channel (Fig. 1). At Sender, abstract objects sequentially generated by Client are passed to Relay as requests for service provided by Server. An object can be of intrinsic class *low* (L) or *high* (H).

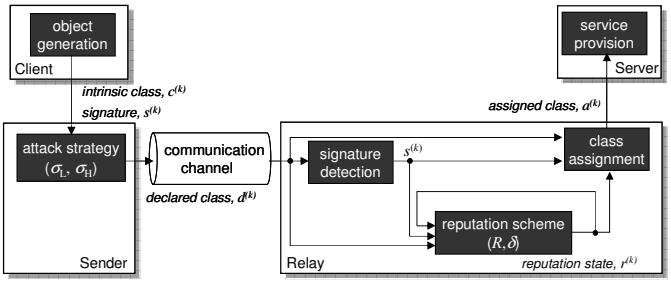

Fig. 1. System operation.

Object generation follows a stationary memoryless random process. Let $c^{(k)}$ be the intrinsic class of the k^{th} generated object (k = 1,2,…) and ρ be the proportion of class H objects, i.e., probability that a generated object is of intrinsic class H. Denote by $rand(\rho)$ an event occurring with probability ρ. Then

$$c^{(k)} = \begin{cases} \text{H}, & rand(\rho), \\ \text{L}, & rand(1-\rho). \end{cases} \tag{1}$$

A generated object bears a *signature*, i.e., a set of features that Relay defines as relevant to the class to be assigned (the intrinsic class not being observable). Like intrinsic classes, signatures are exogenous to Sender and cannot be modified; yet Sender is aware of them provided that she knows Relay's signature detection scheme. It is realistic to assume that signatures are intrinsic class-dependent, but not in a deterministic way: an intrinsic class L object may "accidentally" bear an H signature and vice versa. E.g., in the context of packet classification into Access Category under EDCA, a packet's signature can be defined as byte length, with short packets recognized as voice/video and assigned high-quality service, and long ones recognized as best-effort and assigned low-quality service. However, best-effort packets are of variable length, hence some can bear an H signature. Let $s^{(k)}$ represent the k^{th} object's signature, and $\varepsilon_c = \Pr[s^{(k)} \neq c^{(k)} \mid c^{(k)} = c]$, c = L or H, be the (stationary) signature error rates, i.e., probabilities that a generated intrinsic class c object has a "wrong" signature. Then

$$s^{(k)} \begin{cases} \neq c^{(k)}, & rand(\varepsilon_{c^{(k)}}), \\ = c^{(k)}, & rand(1-\varepsilon_{c^{(k)}}). \end{cases} \tag{2}$$

When passing the k^{th} object to Relay, Sender declares its class as $d^{(k)}$ based on $c^{(k)}$, $s^{(k)}$ and a stationary Fake VIP attack strategy (σ_L, σ_H), where $\sigma_c = \Pr[d^{(k)} = \text{H} \mid s^{(k)} = \text{L} \wedge c^{(k)} = c]$; clearly, if $s^{(k)}$ = H then $d^{(k)}$ = H is only plausible. That is,

$$d^{(k)} = \begin{cases} \text{H}, & s^{(k)} = \text{H} \vee rand(\sigma_{c^{(k)}}), \\ \text{L}, & \text{otherwise}. \end{cases} \tag{3}$$

At Relay, assignments of class (service quality) to successively arriving objects are based on Sender's declared class, signature detection and Sender's current *reputation state*. Let $r^{(k)} \in \{1,\ldots,R\}$ be the reputation state just before the arrival of the k^{th} object and let $a^{(k)}$ denote its assigned class, where $R \geq 2$ and k = 1,2,…. In the *trust state* R, Relay trusts the object and passes it to Server as is, implying $a^{(k)} = d^{(k)}$; in the spirit of our minimum-information framework, in such a case we assume that Relay does not observe $d^{(k)}$. In the non-trust states, Relay disregards declared class and invokes signature detection to decide the assigned class $a^{(k)}$:

$$a^{(k)} = \begin{cases} d^{(k)}, & r^{(k)} = R, \\ s^{(k)}, & r^{(k)} < R. \end{cases} \tag{4}$$

(note that in accordance with postulate (vi), under no circumstances can $(s^{(k)},a^{(k)})$ = (H,L) occur).

If $r^{(k)} < R$ and $(s^{(k)},d^{(k)})$ = (L,H), i.e., a Fake VIP attack is suspected (of which Relay cannot be certain, since $s^{(k)}$ = L does not imply $c^{(k)}$ = L), then Sender's reputation state is lowered. On the other hand, a perceived honest demand of class L (i.e., $(s^{(k)},d^{(k)})$ = (L,L)) raises the reputation state. In the trust state R, $s^{(k)}$ is not detected and only declared class is observed; a cautionary policy is then to lower the reputation state if $d^{(k)}$ = H. In all other cases the reputation state remains unchanged. The reputation scheme also defines a parameter $\delta \in [0, 1]$ that measures the tendency to lower a reputation state (including state R). Both R and δ are Relay's private information. Formally, reputation state transitions are as follows, where $\Phi = (r^{(k)} = R \wedge d^{(k)} = \text{H}) \vee (1 < r^{(k)} < R \wedge (s^{(k)},d^{(k)}) = (\text{L,H}))$ and $\Theta = (r^{(k)} < R \wedge (s^{(k)},d^{(k)}) = (\text{L,L}))$:

$$r^{(k+1)} = \begin{cases} r^{(k)} - 1, & \Phi \wedge rand(\delta), \\ r^{(k)} + 1, & \Theta \wedge rand(1-\delta), \\ r^{(k)}, & \text{otherwise}. \end{cases} \tag{5}$$

B. Utilities

Under the reputation scheme, Sender's utility associated with the k^{th} generated object is a unit benefit if a Fake VIP attack has been successful, a unit loss if Relay has wrongly assigned class L, and neutral otherwise:

$$u_{\text{Sender}}^{\text{reput}}(c^{(k)},a^{(k)}) = \begin{cases} +1, & (c^{(k)},a^{(k)}) = (\text{L,H}), \\ -1, & (c^{(k)},a^{(k)}) = (\text{H,L}), \\ 0, & \text{otherwise}. \end{cases} \tag{6}$$

As a baseline utility, Sender may take the *never-trust scenario*, where Relay does not employ a reputation scheme and so never trusts declared class, i.e., $a^{(k)} \equiv s^{(k)}$:

$$u_{\text{Sender}}^{\text{never-trust}}(c^{(k)}, s^{(k)}) = \begin{cases} +1, & (c^{(k)}, s^{(k)}) = (\text{L},\text{H}), \\ -1, & (c^{(k)}, s^{(k)}) = (\text{H},\text{L}), \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Another baseline, ideal from Sender's perspective, is the *always-trust scenario*, where Relay does not employ signature detection and always trusts declared class, thus allowing unpunished Fake VIP attacks with $(\sigma_{\text{L}}, \sigma_{\text{H}}) = (1,1)$:

$$u_{\text{Sender}}^{\text{always-trust}}(c^{(k)}) = \begin{cases} +1, & c^{(k)} = \text{L}, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

Relay's utility can be defined as a linear combination of signature detection and high-quality service provision costs. The former is suffered whenever an object arrives in a non-trust state, and the latter when class H is assigned (recall that Relay seeks to protect Server from undue effort). Thus

$$u_{\text{Relay}} = \beta \cdot 1_{r^{(k)} < R} + 1_{a^{(k)} = \text{H}}, \quad (9)$$

where $\beta > 0$ is the relative weight of the signature detection cost, and $1_Z = 1$ if Z = true and 0 otherwise. Sender optimizes $(\sigma_{\text{L}}, \sigma_{\text{H}})$ and Relay optimizes (R, δ) with a view of the statistical expectation of utility across a long sequence of objects.

IV. Players' Expected Utilities

A. Markov Chain Analysis

At Relay, arriving objects determine successive reputation levels so that $(r^{(k)})_{k=1,2,\ldots}$ is a homogeneous Markov chain over state space $\{1,\ldots,R\}$. Its one-step transition matrix is $\mathbf{T} = (t_{ij})_{i,j=1,\ldots,R}$, where $t_{ij} = \Pr[r^{(k+1)} = j \mid r^{(k)} = i]$. Based on (5), Fig. 2 depicts the state transitions (self-loops are not drawn). The relevant probabilities occurring in t_{ij} can be expressed as follows (whenever no confusion arises, the superscripts k are omitted): $\Pr[d = \text{L}] = \Pr[(s,d) = (\text{L},\text{L})] = \omega(1-x)$, $\Pr[d = \text{H}] = 1 - \omega(1-x)$, and $\Pr[(s,d) = (\text{L},\text{H})] = \omega x$, where

$$\omega = (1-\rho)(1-\varepsilon_{\text{L}}) + \rho\varepsilon_{\text{H}}, \quad (10)$$

$$x = \frac{(1-\rho)(1-\varepsilon_{\text{L}})\sigma_{\text{L}} + \rho\varepsilon_{\text{H}}\sigma_{\text{H}}}{\omega} = \Pr[d = \text{H} \mid s = \text{L}]. \quad (11)$$

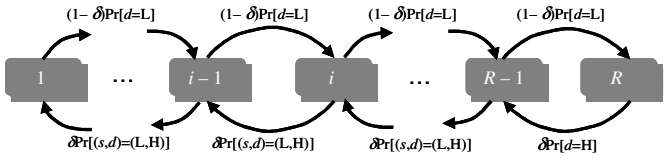

Fig. 2. Reputation state transitions.

Assume $r^{(1)} = R$. Then it is easy to distinguish the following cases: $\delta = 0$ implies that the Markov chain stays indefinitely at the trust state R, $\delta = 1$ implies that the trust state is never revisited, and if $0 < \delta < 1$ and $x = 1$ then with high probability state R will be visited only finitely many times. In the remaining cases, i.e., $0 < \delta < 1$ and $x < 1$, a stationary probability π_i of visiting each state i can be obtained (for $x > 0$ the Markov chain is ergodic and for $x = 0$ only states R and $R - 1$ are recurrent). Of interest is π_R, the probability of visiting the trust state. Since $\mathbf{T}$ is clearly tridiagonal, the Markov chain is reversible and local balance applies: $\pi_i t_{i,i+1} = \pi_{i+1} t_{i+1,i}$ for $i = 1,\ldots,R-1$. This yields for $i < R$:

$$\pi_i = \pi_R \prod_{j=i}^{R-1} \frac{t_{j+1,j}}{t_{j,j+1}} = \pi_R \left(\frac{\delta}{1-\delta} \frac{x}{1-x} \right)^{R-1-i} \left(\frac{1}{\omega(1-x)} - 1 \right) \frac{\delta}{1-\delta}. \quad (12)$$

Using the normalization constraint $\pi_1 + \ldots + \pi_R = 1$ and combining with the above described nonergodic cases we have:

$$\pi_R = \begin{cases} 1, & \delta = 0, \\ 0, & \delta = 1 \vee (0 < \delta < 1 \wedge x = 1), \\ 1/f(x,\delta), & 0 < \delta < 1 \wedge x < 1, \end{cases} \quad (13)$$

where $f(x,\delta) = 1 + \left(\frac{1}{\omega(1-x)} - 1 \right) \cdot \frac{\delta}{1-\delta} \cdot \sum_{i=0}^{R-2} \left(\frac{\delta}{1-\delta} \cdot \frac{x}{1-x} \right)^i$.

If $f(1,0) = 1$ then the last case in (13) subsumes the former two.

B. Expected Utilities

Under the reputation scheme, Sender's expected utility (net benefit) derived from (6) is

$$\begin{aligned} \mathsf{E}u_{\text{Sender}}^{\text{reput}} &= +1 \cdot \Pr[(c,a) = (\text{L},\text{H})] - 1 \cdot \Pr[(c,a) = (\text{H},\text{L})] \\ &= \omega x \pi_R + (1-\rho)\varepsilon_{\text{L}} - \rho\varepsilon_{\text{H}}. \end{aligned} \quad (14)$$

For the never-trust scenario ($a \equiv s$), we have from (7):

$$\begin{aligned} \mathsf{E}u_{\text{Sender}}^{\text{never-trust}} &= +1 \cdot \Pr[(c,s) = (\text{L},\text{H})] - 1 \cdot \Pr[(c,s) = (\text{H},\text{L})] \\ &= (1-\rho)\varepsilon_{\text{L}} - \rho\varepsilon_{\text{H}}. \end{aligned} \quad (15)$$

If Sender's expected utility under the reputation scheme falls below that under the never-trust scenario, she may be suspicious of Relay cheating on class assignments. Sender is therefore interested in her *excess benefit* defined as:

$$\mathsf{E}u_{\text{Sender}}^{\text{reput}} - \mathsf{E}u_{\text{Sender}}^{\text{never-trust}} = \omega x \pi_R \quad (16)$$

This excess Sender owes to Relay occasionally entering the trust state R. Observe that (16) is nonnegative, and is strictly positive in the generic case $0 < \sigma_{\text{L}}, \sigma_{\text{H}}, \delta < 1$. Hence, the reputation scheme does not put Relay under Sender's suspicion of cheating to exploit the moral hazard situation; at the worst, Sender may surmise that Relay never trusts declared class. For

convenience, (16) is further normalized to the always-trust scenario, where $a \equiv d$ and $(\sigma_L,\sigma_H) = (1,1)$. From (8):

$$\mathsf{E}u_{\text{Sender}}^{\text{always-trust}} = +1\cdot\Pr[c = \text{L}] = 1 - \rho. \quad (17)$$

Finally, Sender's expected utility (benefit) is:

$$\mathsf{E}u_{\text{Sender}} = \frac{\mathsf{E}u_{\text{Sender}}^{\text{reput}} - \mathsf{E}u_{\text{Sender}}^{\text{never-trust}}}{\mathsf{E}u_{\text{Sender}}^{\text{always-trust}} - \mathsf{E}u_{\text{Sender}}^{\text{never-trust}}} = x\pi_R = \frac{x}{f(x,\delta)} \in [0,1] \quad (18)$$

(the extreme values signify never trusted and always trusted declared class). Note that Sender can derive (15) and (17) from the objects' intrinsic classes and signatures.

The statistical expectation of (9) is:

$$\begin{aligned}\mathsf{E}u_{\text{Relay}} &= \beta(1 - \pi_R) + \Pr[a = \text{H}] \\ &= \beta(1 - \pi_R) + \Pr[s = \text{H}] + \Pr[(s,d) = (\text{L},\text{H})]\pi_R. \end{aligned} \quad (19)$$

Combining (18) and (19) we have Relay's utility (cost):

$$\mathsf{E}u_{\text{Relay}} = \beta(1 - \pi_R) + 1 - \omega(1 - \mathsf{E}u_{\text{Sender}}). \quad (20)$$

This cost becomes $\beta + 1 - \omega$ when Relay never trusts declared class, and 1 when Relay always trusts declared class (which is then always set to H).

For illustration assume $\varepsilon_L = \varepsilon_H$, then $\omega \in [0, 1 - \rho]$; a realistic range is, e.g., $\omega \in [0, 0.9]$. As β we take the ratio of throughput degradation due to signature detection, on order of 10%..90% as reported for proprietary DPI solutions [20], and that due to TRAs, on order of 50%..90% [1]. Therefore a realistic range is $\beta \in [0, 2]$. Fig. 3 plots both players' utilities vs. x for $R = 5$, $\omega = 0.8$, $\beta = 0.3$, and various δ.

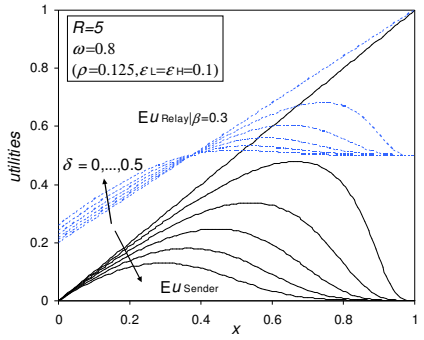

Fig. 3. Expected utilities under Fake VIP attacks.

V. Equilibrium and Off-Equilibrium Play

We formalize the two-person game arising between Sender and Relay as $\langle P, S' \times S'', (u', u'')\rangle$, where $P = \{\text{Sender},\text{Relay}\}$ is the set of players, $S' = [0, 1]^2$ is the set of Sender's strategies (represented by (σ_L,σ_H)), $S'' = \{2,3,\ldots\} \times [0, 1]$ is the set of Relay's strategies (represented by (R,δ)), and u', u'': $S' \times S'' \to \mathbf{R}$ are the players' utilities (represented by $\mathsf{E}u_{\text{Sender}}$ and $\mathsf{E}u_{\text{Relay}}$). In a Stackelberg game [5], Relay (the "leader") sets her strategy (R,δ) anticipating a best-response strategy (σ_L,σ_H) of a selfish Sender (the "follower"); the players then remain at a *Stackelberg equilibrium* (SE). (Various paths to SE can be envisaged in a dynamic game scenario, given that the players are able to observe their utilities, e.g., via trial-and-error or more sophisticated long-term learning. This aspect will be left out and we will only characterize the static SE.)

Sender selects a best-response (σ_L,σ_H) assuming a fixed Relay's strategy (R,δ). Note that σ_L and σ_H only enter Sender's and Relay's utilities through $x \in [0, 1]$, as given by (11); $x = 0$ corresponds to $(\sigma_L,\sigma_H) = (0,0)$ and $x = 1$ to $(\sigma_L,\sigma_H) = (1,1)$. Having decided on a particular x, Sender still has a freedom of choice of (σ_L,σ_H) as prescribed by (11); in particular, will not lose by setting $\sigma_L = 0$ or $\sigma_H = 0$, i.e., never launching a Fake VIP attack on a class L or class H object, respectively. For a given $\omega \in [0, 1]$ assume that $R \geq 2$ is fixed (we will consider various R separately), and indicate explicitly the dependence of $\mathsf{E}u_{\text{Sender}}$ on x and δ. Then Sender seeks

$$x^*(\delta) = \arg\max_{x\in[0,1]} \mathsf{E}u_{\text{Sender}}(x,\delta), \quad (21)$$

i.e., in the ergodic case, seeks a maximum of $x/f(x,\delta)$, where f, defined in (13), is convex and increases monotonously in x. Clearly, $x^*(0) = 1$ and it is easy to prove that for $\delta > 0$, $x^*(\delta)$, is unique in [0, 1) (cf. Fig. 3). Tedious but straightforward calculation of $f_x(x,\delta)$ moreover reveals that $x^*(\delta)$ decreases in δ regardless of R and ω (cf. Fig. 4). Relay selects a best-response (R,δ) knowing Sender's best response $x^*(\cdot)$ (again, ω is given and R is treated as fixed). For clarity, let us indicate the dependence of $\mathsf{E}u_{\text{Relay}}$ on x, δ and β. By (20), Relay reaches an SE at

$$\delta^*(\beta) = \arg\min_{\delta\in[0,1]} \mathsf{E}u_{\text{Relay}}(x^*(\delta),\delta,\beta) = \arg\min_{\delta\in[0,1]} \frac{\omega x^*(\delta) - \beta}{f(x^*(\delta),\delta)}. \quad (22)$$

Numerical calculation shows that $f(x^*(\delta),\delta)$ increases in δ. Hence for small enough β, $\delta^*(\beta) = 1$, i.e., Relay never trusts declared class and only relies on signature detection, thus it does not need a reputation scheme. For large enough β, $\delta^*(\beta) = 0$, i.e., Relay always trusts declared class, thus it does not need a signature detection mechanism either. Fig. 5 shows sample plots of both players' SE strategies: $\delta^*(\beta)$ and $x^*(\delta^*(\beta))$, as well as $\pi_R = 1/f(x^*(\delta^*(\beta)),\delta^*(\beta))$ against β (some of the plots look ragged due to the numerical inaccuracies of locating the maxima of $\mathsf{E}u_{\text{Relay}}(x,\delta,\beta)$ vs. δ that become quite flat for some β). As the plots illustrate, under a growing cost of signature detection Relay is inclined to trust declared class, to which Sender responds with constant Fake VIP attack. Thus Sender's benefit grows with β in return for Relay's growing cost, cf. Fig. 6. The plot of $\delta^*(\beta)$ reveals a range of β for

which the reputation scheme is *meaningful*, i.e., $0 < \delta^*(\beta) < 1$. For small enough β, Sender launches Fake VIP attacks with restraint ($x^*(\delta^*(\beta)) < 1$), while Relay trusts declared class quite frequently ($\delta^*(\beta)$ is distinctly above zero); this is when Relay may also find the reputation scheme *practical*.

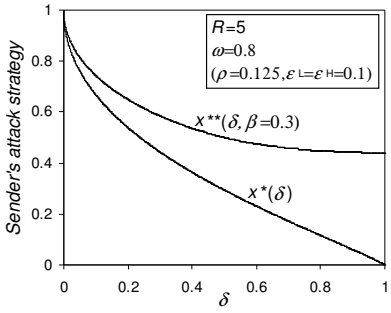

Fig. 4. Sender's SE and OE best responses to (R,δ).

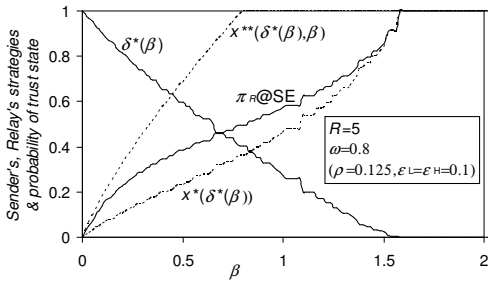

Fig. 5. SE play vs. relative cost of signature detection.

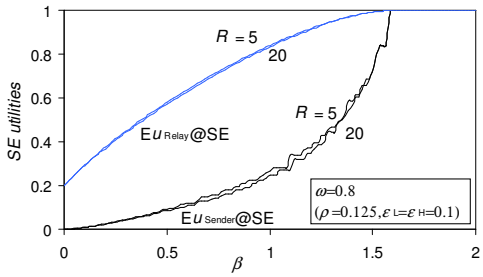

Fig. 6. Expected SE utilities vs. relative cost of signature detection.

Would a reputation scheme of a larger complexity (i.e., R) yield a better tradeoff between both players' SE utilities? For $\omega = 0.8$, Fig. 6 compares $R = 5$ and $R = 20$ and shows that, indeed, a larger R improves Relay's cost as well as worsens Sender's benefit at equilibrium, although the difference is barely perceptible. This qualitative conclusion is largely independent of ω. Fortunately, then, R is not a sensitive parameter and the reputation scheme need not be complex.

To what extent is Relay protected from Sender's off-equilibrium (OE) play resulting from inaccurate observation of $\mathsf{E}u_{\text{Sender}}$ or insufficient intelligence to learn the SE strategy, or perhaps from malice? In the latter (worst) case, Sender seeks to maximize Relay's cost so that her strategy becomes:

$$x^{**}(\delta,\beta) = \arg\max_{x\in[0,1]} \mathsf{E}u_{\text{Relay}}(x,\delta,\beta)\,. \qquad (23)$$

Fig. 4 and Fig. 5 illustrate that $x^{**}(\delta,\beta)$ decreases in δ and that $x^{**}(\delta^*(\beta),\beta)$ is typically larger than $x^*(\delta^*(\beta))$, i.e., a malicious Sender attacks an SE-playing Relay more frequently than does a selfish one. The utilities at OE are depicted in Fig. 7. Relay is prepared for Sender's selfish play, but is exposed to malicious play instead, therefore her utility worsens considerably for intermediate β and approaches $\beta + 1 - \omega$, like in the never-trust scenario. However, Sender's utility suffers even more; hence, Sender's malice is either ineffective or self-damaging. Thus off-equilibrium protection considerations do not reduce the range of β where Relay finds the reputation scheme practical. For large β, $\delta^*(\beta)$ tends to 0 and $x^{**}(\delta^*(\beta))$ tends to 1. Hence at OE, $\mathsf{E}u_{\text{Sender}}$ and $\mathsf{E}u_{\text{Relay}}$ tend to 1, utilities achieved in the always-trust scenario under $(\sigma_L,\sigma_H) = (1,1)$.

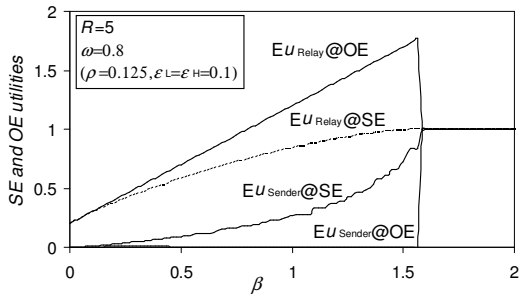

Fig. 7. Comparison of SE and OE expected utilities.

VI. Impact of the Reputation Scheme

To quantify the impact of the reputation scheme upon both players' utilities, introduce the *reputation-free* (RF) scenario to subsume both never-trust and always-trust scenarios (the difference between them being the signature detection scheme). Recall that in the never-trust scenario, Sender's and Relay's expected utilities are 0 and $\beta + 1 - \omega$, respectively, whereas in the always-trust scenario they are equal to 1. Thus in the RF scenario, Relay invokes signature detection when $\beta < \omega$, to which Sender responds with any attack strategy, and skips it when $\beta \geq \omega$, to which Sender responds with $(\sigma_L,\sigma_H) = (1,1)$. Fig. 8 provides a conceptual illustration, featuring also generic plots of $\mathsf{E}u_{\text{Sender}}$ and $\mathsf{E}u_{\text{Relay}}$ at SE against β. The shaded R+ and R++ areas quantify Relay's gains due to the reputation scheme. The S+ and S− areas quantify, respectively, the corresponding Sender's gains at $\beta < \omega$ and Sender's losses at $\beta \geq \omega$.

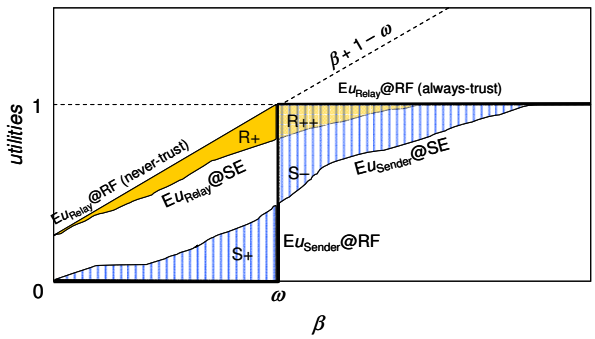

Fig. 8. Conceptual illustration of the impact of reputation scheme.

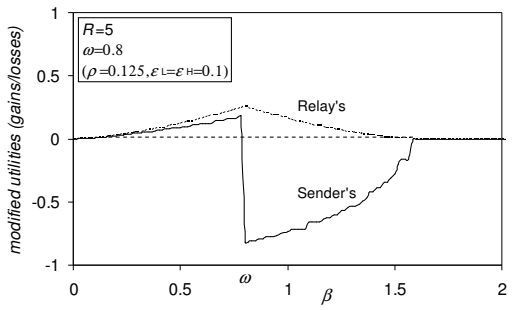

Fig. 9. Expected utility gains and losses due to reputation scheme.

To reflect the gains and losses of SE play with respect to RF play we modify (18) and (20) as follows, using shorthand @SE for $(x^*(\delta^*(\beta)), \delta^*(\beta))$:

$$\mathrm{E}u_{\mathrm{Sender,mod}} = \mathrm{E}u_{\mathrm{Sender}} \,@\, \mathrm{SE} - \mathrm{E}u_{\mathrm{Sender}} \,@\, \mathrm{RF}\,, \qquad (24)$$

$$\mathrm{E}u_{\mathrm{Relay,mod}} = \mathrm{E}u_{\mathrm{Relay}} \,@\, \mathrm{RF} - \mathrm{E}u_{\mathrm{Relay}} \,@\, \mathrm{SE}\,. \qquad (25)$$

Fig. 9 plots (24) and (25) based on Fig. 4. The range of β just beyond $\beta = \omega$ is where the reputation scheme at SE brings the most significant gains to Relay and the most significant losses to Sender, as compared to RF play. Interestingly, when $\beta < \omega$, both players gain: Relay reduces cost by occasionally trusting declared class, while Sender's Fake VIP attacks occasionally succeed (indeed, owing to the term $\beta(1 - \pi_R)$ in (20), (18) and (20) are not necessarily conflicting, i.e., the game need not be antagonistic). Overall, Relay can only gain by employing the reputation scheme, whereas even a clever Sender gains little when $\beta < \omega$ and loses much when $\beta > \omega$.

VII. Conclusion

In the outlined minimum-information framework, no systematic analyses of, or defenses against, Fake VIP attacks are known as yet. We propose that Relay resort to a reputation scheme that dictates whether to trust the declared class or invoke signature detection. Unconventionally, the scheme is double-blind: Relay cannot learn an object's intrinsic class, whereas Sender cannot learn the assigned class or her current reputation. For a naturally arising Stackelberg game setting we offer, among others, the following findings:

- compared to RF play, SE play brings Relay utility gains regardless of the relative signature detection cost β (with maximum gains at $\beta = \omega$), i.e., the reputation scheme is uniformly beneficial against a selfish Sender; interestingly, for small enough β, Sender gains too, whereas for larger she suffers utility loss,
- against Relay's SE play, Sender's OE play, e.g., resulting from malice, is either ineffective or self-damaging; thus off-equilibrium considerations do not reflect upon the practicality of the reputation scheme, and
- the complexity of the reputation scheme (in terms of R) has little bearing on both players' SE utilities.

Sender might probably increase her SE utility (and diminish Relay's) by exploiting some rudimentary idea of the workings of Relay's reputation scheme; this observation stimulates further research into even more intelligent strategies of, and optimum defense against, Fake VIP attacks. How close the presented heuristic design is to that goal is an open question.

References

[1] J. Konorski and S. Szott, "Discouraging traffic remapping attacks in local ad hoc networks," IEEE Transactions on Wireless Communications, vol. 13, 7, pp. 3752–3767, July 2014.

[2] S. Mangold, S. Choi, G. R. Hiertz, O. Klein, and B. Walke, "Analysis of IEEE 802.11e for QoS support in Wireless LANs," IEEE Wireless Communications, vol. 10, 6, pp. 40–50, Dec. 2003.

[3] I.-R. Chen, F. Bao, and J. Guo, "Trust-based service management for Social Internet of Things systems, " IEEE Transactions on Dependable and Secure Computing, vol. 13, 6, pp. 684-696, Nov./Dec. 2016.

[4] Po-Ching Lin, Ying-Dar Lin, Yuan-Cheng Lai, and Tsern-Huei Lee, "Using String Matching for Deep Packet Inspection," Computer, vol. 41, 4, pp. 23–28, April 2008.

[5] Rasmussen E.: *Games and Information: An Introduction to Game Theory*, 3rd ed. Blackwell Publishers. 2001.

[6] P. L. Bartlett. *Online Prediction*. http://stat.berkeley.edu/~bartlett/papers/b-ol-16.pdf. 2015.

[7] A. Josang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," Decision Support Systems, vol. 43, 2, pp. 618–644, March 2007.

[8] Q. Liu, "Information acquisition and reputation dynamics," Review of Economic Studies, vol. 78, no. 4, pp. 1400-1425, 2011.

[9] K. Hoffman, D. Zage, and C. Nita-Rotaru, "A survey of attack and defense techniques for reputation systems," ACM Computing Surveys, vol. 42, 1, Dec. 2009.

[10] Y. Chae, L. C. DiPippo, and Y. L. Sun, "Trust management for defending on-off attacks," IEEE Transactions on Parallel and Distributed Systems, vol. 26, 4, pp. 1178–1191, April 2015.

[11] M. H. Cheung, A. H. Mohsenian-Rad, V. W. Wong, and R. Schober, "Random access protocols for WLANs based on mechanism design," Proc. IEEE Int. Conf. on Comm. ICC'09, Dresden, Germany, June 2009.

[12] S. H. Nguyen, L. L. Andrew, and H. L. Vu, "Service differentiation without prioritization in IEEE 802.11 WLANs," Proc. 36th IEEE Conf. on Local Computer Networks (LCN), Bonn, Germany, Oct 2011.

[13] L. Galluccio, "A game-theoretic approach to prioritized transmission in wireless CSMA/CA networks," Proc. 69th IEEE Vehicular Technology Conf., Barcelona, Spain, April 2009.

[14] M. Li and B. Prabhakaran, "MAC layer admission control and priority re-allocation for handling QoS guarantees in non-cooperative wireless LANs," Springer Mobile Networks and Applications, vol. 10, 6, pp. 947–959, Dec. 2005.

[15] F. Hendrikx, K. Bubendorfer, and R. Chard, "Reputation systems: A survey and taxonomy," J. Parallel and Distrib. Comput., vol. 75, pp. 184-197, Jan. 2015.

[16] Y. L. Sun and Y. Liu, "Security of online reputation systems: The evolution of attacks and defenses," IEEE Signal Proc. Mag., vol. 29, 2, pp. 87–97, March 2012.

[17] R. Kerr and R. Cohen, "Smart cheaters do prosper: Defeating trust and reputation systems," Proc. AAMS'09, Budapest, Hungary, May 2009.

[18] Y. Freund, M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, and R. E. Schapire, "Efficient Algorithms for Learning to Play Repeated Games Against Computationally Bounded Adversaries," Proc. 36th Annual Symp. Foundations of Computer Science, Milwaukee WI, Nov. 1995.

[19] A. Patcha and Jung-Min Park, "A Game Theoretic Formulation for Intrusion Detection in Mobile Ad Hoc Networks," Int. J. of Network Security, vol. 2, 2, pp.131–137, March 2006.

[20] Intel Corporation, "Delivering 160Gbps DPI performance on the Intel® Xeon® processor E5-2600 series using HyperScan," solution white paper, 2013.