Hardening Trust Models against Slandering Attacks in Relayed Content Delivery Services

1st Francesco Buccafurri
Dept. DIIES
University of Reggio Calabria
Reggio Calabria, Italy
bucca@unirc.it

2nd Vincenzo De Angelis
Dept. DIIES
University of Reggio Calabria
Reggio Calabria, Italy
vincenzo.deangelis@unirc.it

3rd Maria Francesca Idone
Dept. DIIES
University of Reggio Calabria
Reggio Calabria, Italy
mariafrancesca.idone@unirc.it

4th Cecilia Labrini
Dept. DIIES
University of Reggio Calabria
Reggio Calabria, Italy
cecilia.labrini@unirc.it

***Abstract*—There are a number of application contexts in which services are delivered by a given provider to some clients through other relay clients in a collaborative fashion. This is, for example, the case of sensor networks, vehicular networks, D2D, and so on. In this case, a security problem arises. Indeed, when a service is relayed by a client, it is not sure that it is relayed correctly. Therefore, the final client could be deceived if malicious relay clients exist. The classical way to contrast this problem is to use a trust mechanism, managed by the provider, based on the feedback returned by the clients. Thanks to this mechanism, the trust of malicious relay clients can be decreased, then reducing (even cancelling) the negative effects of these clients in the community. The trust mechanisms of this type often suffer from weakness against slandering attacks. Dishonest final clients can fraudulently decrease the trust of relay clients, by reporting a false feedback. In this paper, we propose a general approach to fortify the trust mechanism against this kind of attacks.**

***Index Terms*—Trust, reputation, network security**

I. INTRODUCTION AND RELATED WORK

In different application contexts, the providers of a given service, such as a data delivery service, cannot reach all the intended clients for various reasons. For example, if the service is delivered in multicast by a provider through a network and the network capabilities of the clients are non-uniform, it can happen that setting the provision rate at the lowest level of the clients can degrade too much the overall performance of the process. In this case, if a collaborative approach can be adopted, the provision rate can be fixed by selecting a subset of efficient clients, asking them to play as relay clients, by forwarding the service to the final clients. This way, the set of clients can be divided into *relay clients*, which receive the service directly, and *indirect clients*, which receive the service through the relay clients. Solution of this type can be found in various cases, like sensor networks, vehicular networks, D2D, etc. [1]–[6].

Although this approach can be effective to maximize the overall provision efficiency, it can open some security problems. Indeed, we cannot assume that all the relay clients behave honestly. Depending on the type or the service, this fact may also have critical consequences for the indirect clients. Consider for example the case of a vehicular network and suppose that the delivered information is an alert regarding a deviation necessary to bypass an anomalous obstacle in the road. In this case, a tampered relayed information might also have catastrophic consequences for the victim [7], [8].

To contrast this problem, trust mechanisms can be adopted. The provider, assumed as a trusted party, can continuously update the trust of any relay client to decrease the relevance of a malicious component in the community and push down it in the list in such a way that it is not chosen for the intended purpose. This way, its malicious intents are prevented. Independently of the specific trust mechanism and trust model we consider, certainly, feedbacks coming from indirect clients are necessary to build the trust of the community. Possibly, feedbacks are supported also by the certification of the interaction [9] to avoid that the indirect client can invent the interaction. This can be done with the purpose of damaging the relay client (*slandering attack*) [10], [11] by reporting a negative feedback to the provider. However, slandering attacks, in this kind of trust model, can occur also when the interaction has really happened. Therefore, the approaches cited earlier, available in the literature, cannot work in this case. In the field of trust models, to contrast the fact that some parties can lie when evaluating a really happened interaction, external information giving the proof of the truthfulness of an evaluation can be also used, as in [12].

However, external information is not always available, so that the only approach we can follow is to have a *witness* in the interaction [13].

The contribution of this paper is to design a witness-based approach applicable to any trust model for relayed services working when the indirect client can lie also when the interaction has actually occurred and when no extra information can be drawn from the system to check the truthfulness of the feedback. This is obtained by changing the delivering scheme, from *one-to-one* to *one-to-two* (i.e., one relay client to two indirect clients) in such a way that the witness mechanism is mixed within the delivering mechanism. This is done to reduce dummy interactions and to obtain, simultaneously, the witness that contrasts slandering attacks and the provision of the service to both the involved indirect clients. Our approach is innovative per se, because its aim is not to propose a new trust model in competition with the existing models, but to provide a general method to enhance trust models used for relayed services, not resistant to slandering attacks, to solve this drawback.

The structure of the paper is the following.

The proposed solution is presented in II and its security is analyzed in III. We perform a cost analysis in terms of exchanged messages in Section IV. Finally, in Section V, we draw our conclusions.

II. OUR SOLUTION

Our proposal refers to any existing trust model T for relayed services, which assigns trust values to the proper actors with a global role. This means that the trust includes a component that impacts the role of the specific actor in the whole community, and not only with respect to another specific actor. This is very common in existing trust models [14], [15].

A. The Basic Model

The scenario of relayed services is modeled as follows. We have three actors: (1) the *provider*, (2) the *relay client*, and (3) the *indirect client*.

A provider P delivers a given content S to the client community. S is delivered directly only to the k relay clients $\langle r_1, r_2, \dots r_k\rangle$. There are also t indirect clients $\langle c_1, c_2, \dots c_t\rangle$ that can receive the service S indirectly by the relay clients. Specifically, as typically happens in concrete relayed service schemes, the mapping between relay clients and indirect clients is one-to-one, in the sense that it can happen that the relay client r_i can serve two different indirect clients c_a and c_b, but the two interactions are independent of each other. Therefore, we can always represent an *interaction* as a pair $I = \langle r_i, c_j\rangle$, where r_i is a relay client and c_j is an indirect client. The interaction I consists of the delivery of the service S, which is broadcasted by the provider to the entire client community.

We consider a centralized trust model T, relying on feedbacks that indirect clients can send the provider. Feedbacks are generated by indirect clients after interactions. The model is centralized in the sense that the computation of the trust is done by the provider, owing all the information needed to reach this goal. We assume that the provider can apply any existing strategy to authenticate correctly all the clients [16]. Therefore, we do not consider the case of impersonation or Sybil attacks [17]. Moreover, we assume that the provider can test if the interaction really occurred. The latter assumption, as highlighted in the introduction, does not eliminate the problem of possible slandering attacks.

Formally, T is a tuple $\langle P, R, C, S_I, F, f_t, A, W\rangle$, where P is the provider $R = \langle r_1, r_2, \dots r_k\rangle$ is the set of relay clients, $C = \langle c_1, c_2, \dots c_t\rangle$ is the set of indirect clients, S_I is the (growing) set of interactions (defined as above), F is the (growing) set of feedbacks sent by the indirect clients to P, $f_t : R \times C \rightarrow (0, 1)$ is the trust function, defining, for each pair $\langle r_i, c_j\rangle$, the trust the provider has assigned to r_i with respect to (future interactions with) c_j, A is the function the provider uses to update the function f_t, and W is the function the provider uses to establish the one-to-one mappings for the next interactions. Clearly, W strictly depends on f_t, but also from some other features possibly related to technological aspects (e.g., strength of the signal, distance, etc.).

For every interaction $I = \langle r_i, c_j\rangle$, a feedback $f(I)$ is sent by the indirect client c_j to P. The interaction I is inserted into the set S_I and the feedback $f(I)$ is inserted into the set F. In which phase this feedback can be produced is strictly depending on the specific application context. It might happen that the feedback can be generated only after the content is exploited, or that it can be immediately obtained by examining and verifying the relayed content. However, this aspect does not influence our approach too.

According to this feedback, the trust value of r_i is updated. This is done on the basis of A. We represent this update as follows: $f'_t = A(f_t, f(I))$, meaning that the function f_t is updated on the basis of the received feedback.

According to the global component of the trust model, A has a *global* behavior. This means the following. For any indirect client c_s with $1 \leq s \leq t$, an ordered sequence of trust values can be extracted from the function f_t. Denote by $\langle t_{1,s}, t_{2,s}, \dots t_{k,s}\rangle$ this list, where $\langle t_{1,s}$ is the maximum value. $t_{i,s}$ is the trust of c_s about the relay node r_i. Formally, $t_{i,s} = f_t(r_i, c_s)$. The global nature of A means that the feedback $f(I)$, related to the interaction $I = (r_i, c_j)$, affects the value $t_{i,x}$, for any $1 \leq x \leq t$. Depending on the specific trust model we consider, the feedback can be modeled in various ways, but, in general, we can assume that $f(I)$ can either *positive* or *negative*. Therefore, from now on we assume that feedbacks are binary information (i.e., *positive*$= 1$, *negative*$= 0$). Coherently, a positive feedback tends to move up $t_{i,x}$ in the corresponding list $\langle t_{1,x}, t_{2,x}, \dots t_{k,x}\rangle$, for any x. Conversely, a negative feedback moves down in the list $t_{i,x}$.

B. The Enhanced Model

It is easy to realize that in the model T, there is the possibility for a malicious indirect client to perform a *slandering attack* on a relay client just by producing a negative feedback even though the relayed content is correct. The slandering attack would have an impact on the trust of the relay client with respect to the entire community. Our approach is aimed to contrast the above threat. The basis of our strategy is to mix the concept of *witness* [13] with the relay mechanism itself. In this section, we show how to modify T into T^* to embed this strategy.

We observe in advance that the required changes can be applied to *any* trust model as a *composition* transformation. For this reason, our approach can be considered orthogonal with respect to the original trust model.

The first change regards the notion of interaction. The one-to-one mapping between relay clients and indirect clients is transformed into a *one-to-two* mapping in the following way. An interaction I is now defined as a triple $I = \langle r_i, c_a, c_b\rangle$, where r_i is a relay client and c_a and c_b are indirect clients. c_a plays the role of *witness* for c_b and c_b plays the role of *witness* for c_a. However, both c_a and c_b are clients requiring the services. As in the standard case, the association between relay client and indirect client is done on the basis of the function

W. Therefore, from the function f_t, a list $\langle t_{1,s}, t_{2,s}, \ldots t_{k,s} \rangle$ of trust values can be extracted for any indirect client c_s. Differently from the standard case, the fact that an interaction is a triple $I = \langle r_i, c_a, c_b \rangle$, means that the relayed content S is sent by r_i to both c_a and c_b. Moreover, c_a and c_b should be chosen in such a way that a direct interaction between c_a and c_b is possible (for example, they can interact in D2D fashion, in the case of D2D applications). Then, when c_a and c_b are able to verify the quality of the received content, they exchange a message. Specifically, if the result of the check performed by c_a (c_b, resp.) is positive, then the content S is forwarded to c_b (c_a, resp). Otherwise, a failure message is sent. Then, the proper feedback is sent to the provider, depending on the result of the check. Therefore, the provider receives two feedbacks $f_a(I)$ and $f_b(I)$ possibly different from each other.

The second change regards the function A, that becomes A^*. Indeed, the function f_t is updated as follows. $A^*(f_t, f_a(I), f_b(I)) = A(f_t, f_a(I) \vee f_b(I))$, where $\vee$ represents the Boolean operator OR. The rationale of the above construction will be clear in Section III. Indeed, it is the basis of the robustness of our approach against slandering attacks.

III. SECURITY ANALYSIS

In this section, we analyze how our approach protects the modified trust model against slandering attacks. Clearly, we have also to show that T^* preserves the security properties of T under the assumption that at most one client per interaction can behave maliciously. In other words, in the *threat model* we consider in the first part of our analysis, the only possible type of adversary is a relay client, acting by maliciously performing the relay operation.

In this threat model, we have to show that, if the original model T is able to detect that the malicious behavior of a relay client r_i compromises the correct reception of the service for a given indirect client c_a (thus updating the trust of r_i accordingly), this occurs also in T^*. This is modeled by the following theorem.

Theorem 3.1: Let $I_a = \langle r_i, c_a \rangle$ be an interaction occurring in T in which r_i behaves maliciously and then the correct reception of the service S for c_a is compromised. Let $I = \langle r_i, c_a, c_b \rangle$ be the corresponding interaction in T^* involving the same indirect client c_a (and another indirect client c_b selected according to W). Then, if r_i behaves maliciously and the correct reception of the service S for c_a is compromised, then T^* is able to detect the malicious behavior of r_i as T and the trust of r_i is updated in T^* in the same way as in T.

Proof 3.1: According to T, if in $I_a = \langle r_i, c_a \rangle$ r_i behaves maliciously, then c_a reports a negative feedback $f(I_a)$. Consider now the interaction $I = \langle r_i, c_a, c_b \rangle$ in which r_i behaves maliciously. Two cases may hold: either (1) r_i behaves maliciously with either c_a or c_b, or (2) r_i behaves maliciously with both c_a and c_b. In case (1), if c_b is the only indirect client receiving a bad content, then, obviously, we are not in the case of the hypothesis, because c_a receives the good content. Suppose now that c_a is the only indirect client receiving a bad content. According to the message exchanging enabled in T^* between c_a and c_b, c_a receives the good content by c_b anyway. Therefore, we are not in the case of the hypothesis, because the correct reception of the service S for c_a is not compromised. In case (2), both $f_a(I)$ and $f_b(I)$ are negative (i.e., their value is 0). Since $f(I_a)$ in T is the same as $f_a(I)$ in T^*, $f(I_b)$ in T is the same as $f_b(I)$ in T^*, and $A^*(f_t, f_a(I), f_b(I)) = A(f_t, f_a(I) \vee f_b(I))$, we have that $A^*(f_t, f_a(I), f_b(I)) = A(f_t, 0)$. Therefore, the malicious behavior of r_i is detected in T^*, and the trust of r_i is updated in T^* in the same way as in T. This concludes the proof.

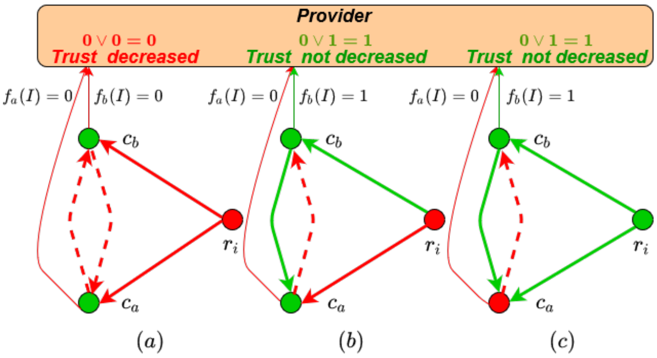

Fig. 1. (a) Consistent malicious relay client; (b) Byzantine malicious relay client; (c) Slandering attack.

The proof of the above theorem allows us to better formalize the possible malicious behavior of a relay client when the enhanced trust model T^* is adopted. Specifically, the malicious behavior of a given relay client r_i during an interaction $I = \langle r_i, c_a, c_b \rangle$ can be:

- *Consistent*, if it is applied to both c_a and c_b (meaning that the right content is relayed neither to c_a nor to c_b).
- *Byzantine*, if the bad content is sent by r_i to only one between c_a and c_b. The right content is sent to the other indirect client.

A consistent malicious behavior results in a compromised service but is detected by the trust model T^*. Instead, a byzantine malicious behavior is not able to produce damage to the victim, so it is not detected by the model as an event causing reduction of trust. The above cases are summarized in Figures 1.(a) and 1.(b). Therein, a red circle is malicious, a green circle is behaving correctly, a red arrow with continuous line is a corrupted content, a green arrow with continuous line is a good content, while dashed arrows represent failure messages. Slight arrows represent feedbacks, red if negative, green if positive.

So far, we have only checked that the modified trust model T^* preserves the security property of the original model T. Now, we consider slandering attacks, to show that T^* enhances T. To do this, we have to change our threat model (keeping the assumption of a single attacker). Specifically, in this case, the adversary is an indirect client, which wants to damage an honest relay client by reporting a negative feedback although the content has been delivered correctly. It holds that T^* is resistant to this type of attack, while T is not, as stated in the theorem below.

Theorem 3.2: Let $I_a = \langle r_i, c_a \rangle$ be an interaction occurring in T in which c_a behaves maliciously by reporting a negative feedback although the content S has been delivered correctly. Let $I = \langle r_i, c_a, c_b \rangle$ be the corresponding interaction in T^* involving the same indirect client c_a (and another indirect client c_b selected according to W). Then, T does not prevent the effect of attack performed by c_a whereas T^* does this.

Proof 3.2: The fact that T does not prevent the effect performed by c_a is straightforward. Indeed, the provider cannot be able to distinguish this case from the case of a negative feedback sent by c_a when r_i behaved maliciously. Therefore, in T, the trust of r_i is badly updated according to $A(f_t, 0)$. Concerning T^*, as r_i is behaving correctly, it relays the content S correctly to both c_a and c_b. As a consequence, whilst $f_a(I) = 0$, due to the malicious behavior of c_a, $f_b(I) = 1$. It turns that $A^*(f_t, f_a(I), f_b(I)) = A(f_t, f_a(I) \vee f_b(I)) = A(f_t, 1)$. Therefore, the trust of r_i is not decreased. The attack is then nullified.

The case of slandering attack is summarized in Figure 1.(c).

IV. COST ANALYSIS

To comparative analyze the cost of T and T^* in terms of exchanged messages, we detail here how the communication protocol enforced by T^* can be efficiently defined.

1) r_i sends the content S to both c_a and c_b. This costs 2 *content* messages;
2) c_a and c_b check the content and send the proper feedback $f_a(I)$ and $f_b(I)$. This costs 2 *feedback* messages;
3) c_a (c_b, resp.) sends only a *flag* to c_b (c_a, resp.), positive if it received the right content, negative otherwise. This has a negligible cost, since only a bit is sent.
4) c_a (c_b, resp.) sends the received content S to c_b (c_a, resp.) only if it received a good content from r_i and a negative flag from c_b (c_a, resp.). Therefore, if both the clients received a good content, then they do not exchange further content. Therefore, this step may cost either 0 content messages or 1 content message.

Observe that, for an interaction, the total cost in terms of number of messages is:

- *Worst case*: 3 content messages and 2 feedback messages
- *Best case*: 2 content messages and 2 feedback messages.

As far as the *average case*, we notice that the worst case occurs only in the case of a malicious attempt performed by either r_i or one of the two indirect clients (this is also clear in Figure 1). Therefore, if malicious interactions are only a small percentage of the overall interactions, then we can argue that the average case leads to the same cost as the best case. In the original model T, instead, we have always 2 content messages and 2 feedback messages (equal to the best case of T^*), because the interaction is one-to-one. This analysis shows that T^* has not a relevant price in terms of exchanged messages.

V. CONCLUSIONS

In this paper, we studied how to protect relay clients in relayed content delivery services against slandering attacks. The proposal is done by also taking into account the number of exchanged messages, to avoid that the solution has a relevant price in terms of traffic overhead. We obtain a quite general result applicable to existing trust models, being orthogonal with respect to the method used to build the trust. As a short paper, the aim is just to present the idea by including a conceptual and formal validation. As a future work, we will validate our proposal in a specific application context and a specific trust model also through simulation and experiments.

REFERENCES

[1] M. Wang and Z. Yan, “A survey on security in d2d communications,” *Mobile Networks and Applications*, vol. 22, no. 2, pp. 195–208, 2017.

[2] C. Suraci, S. Pizzi, D. Garompolo, G. Araniti, A. Molinaro, and A. Iera, “Trusted and secured d2d-aided communications in 5g networks,” *Ad Hoc Networks*, vol. 114, p. 102403, 2021.

[3] L. Feng, P. Zhao, F. Zhou, M. Yin, P. Yu, W. Li, and X. Qiu, “Resource allocation for 5g d2d multicast content sharing in social-aware cellular networks,” *IEEE Communications Magazine*, vol. 56, no. 3, pp. 112–118, 2018.

[4] C. Rezende, A. Mammeri, A. Boukerche, and A. A. Loureiro, “A receiver-based video dissemination solution for vehicular networks with content transmissions decoupled from relay node selection,” *Ad Hoc Networks*, vol. 17, pp. 1–17, 2014.

[5] W. Fang, C. Zhang, Z. Shi, Q. Zhao, and L. Shan, “Btres: Beta-based trust and reputation evaluation system for wireless sensor networks,” *Journal of Network and Computer Applications*, vol. 59, pp. 88–94, 2016.

[6] C. V. Anamuro, N. Varsier, J. Schwoerer, and X. Lagrange, “Distance-aware relay selection in an energy-efficient discovery protocol for 5g d2d communication,” *IEEE Transactions on Wireless Communications*, 2021.

[7] S. Su, Z. Tian, S. Liang, S. Li, S. Du, and N. Guizani, “A reputation management scheme for efficient malicious vehicle identification over 5g networks,” *IEEE Wireless Communications*, vol. 27, no. 3, pp. 46–52, 2020.

[8] C. Li, S. Gong, X. Wang, L. Wang, Q. Jiang, and K. Okamura, “Secure and efficient content distribution in crowdsourced vehicular content-centric networking,” *IEEE Access*, vol. 6, pp. 5727–5739, 2018.

[9] F. Buccafurri, A. Comi, G. Lax, and D. Rosaci, “Experimenting with certified reputation in a competitive multi-agent scenario,” *IEEE Intelligent Systems*, vol. 31, no. 1, pp. 48–55, 2015.

[10] D. Wen, W. Huai-Min, J. Yan, Z. Peng *et al.*, “A recommendation-based peer-to-peer trust model,” 2004.

[11] S. Chen, Y. Zhang, Q. Liu, and J. Feng, “Dealing with dishonest recommendation: The trials in reputation management court,” *Ad Hoc Networks*, vol. 10, no. 8, pp. 1603–1618, 2012.

[12] F. Buccafurri, L. Coppolino, S. D’Antonio, A. Garofalo, G. Lax, A. Nocera, and L. Romano, “Trust-based intrusion tolerant routing in wireless sensor networks,” in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2014, pp. 214–229.

[13] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt, “An integrated trust and reputation model for open multi-agent systems,” *Autonomous Agents and Multi-Agent Systems*, vol. 13, no. 2, pp. 119–154, 2006.

[14] A. Jøsang and J. Golbeck, “Challenges for robust trust and reputation systems,” in *Proceedings of the 5th International Workshop on Security and Trust Management (SMT 2009), Saint Malo, France*, vol. 5, no. 9. Citeseer, 2009.

[15] A. Jøsang, S. Marsh, and S. Pope, “Exploring different types of trust propagation,” in *International Conference on Trust Management*. Springer, 2006, pp. 179–192.

[16] C. D. Pham and T. K. Dang, “A lightweight authentication protocol for d2d-enabled iot systems with privacy,” *Pervasive and Mobile Computing*, vol. 74, p. 101399, 2021.

[17] J. R. Douceur, “The sybil attack,” in *International workshop on peer-to-peer systems*. Springer, 2002, pp. 251–260.