# Trust Evaluation Model for Attack Detection in Social Internet of Things

Wafa Abdelghani, Corinne Amel Zayani, Ikram Amous, Florence Sèdes

# Trust Evaluation Model for Attack Detection in Social Internet of Things

Wafa Abdelghani[1,2(✉)], Corinne Amel Zayani[2], Ikram Amous[2], and Florence Sèdes[1]

[1] IRIT Laboratory, Paul Sabatier University, Toulouse, France
{abdelghani.wafa,sedes}@irit.fr
[2] Miracl Laboratory, Sfax University, Sfax, Tunisia
{corinne.zayani,ikram.amous}@isecs.rnu.tn

**Abstract.** Social Internet of Things (SIoT) is a paradigm in which the Internet of Things (IoT) concept is fused with Social Networks for allowing both people and objects to interact in order to offer a variety of attractive services and applications. However, with this emerging paradigm, people feel wary and cautious. They worry about revealing their data and violating their privacy. Without trustworthy mechanisms to guarantee the reliability of user's communications and interactions, the SIoT will not reach enough popularity to be considered as a cutting-edge technology. Accordingly, trust management becomes a major challenge to provide qualified services and improved security.

Several works in the literature have dealed with this problem and have proposed different trust-models. Nevertheless, proposed models aim to rank the best nodes in the SIoT network. This does not allow to detect different types of attack or malicious nodes.

Hence, we overcome these issues through proposing a new trust-evaluation model, able to detect malicious nodes, block and isolate them, in order to obtain a reliable and resilient system. For this, we propose new features to describe and quantify the different behaviors that operate in such system. We formalized and implemented a new function learned and built based on supervised learning, to analyze different features and distinguish malicious behavior from benign ones. Experimentation made on a real data set prove the resilience and the performance of our trust model.

**Keywords:** Social Internet of Things · Social networks · Trust management · Trust attacks · Resilience

This work was financially supported by the PHC Utique program of the French Ministry of Foreign Affairs and Ministry of higher education and research and the Tunisian Ministry of higher education and scientific research in the CMCU project number 18G1431.

## 1 Introduction

The Internet of Things is expected to be dominated by a huge number of interactions between billions of persons and heterogeneous communications among hosts, smart objects and among smart objects themselves. It provides a variety of data that can be aggregated, fused, processed, analyzed and mined in order to extract useful information [6]. Unquestionably, the main strength of the IoT vision is the high impact on several aspects of every-day life and behavior of potential users. However, many challenging issues prevent the vision of IoT to become a reality, such as interoperability, navigability, trust, privacy and security management and resource discovery in such heterogeneous and decentralized network. To resolve some of the cited issues, a new paradigm called Social Internet of Things (SIoT) is born.

Integrating social networking concepts into the Internet of Things has led to the Social Internet of Things paradigm, enabling people and connected devices to interact with offering a variety of attractive applications. SIoT appeared as a result of an evolutionary process that affected modern communication by the advent of IoT in telecommunication scenarios [3]. The first step of this process consists in making objects smart. The next step consists in the evolution of objects with a degree of smartness to pseudo-social objects [3] which can interact with the surrounding environment and perform a pseudo-social behavior with other objects. The last step consists of the appearance of social objects includes being able to autonomously establish relationships with other objects, to join communities and build their own social networks whose can differ from their owner's ones [3]. Adopting such vision is, therefore, a promising new trend, with numerous advantages. First, navigability and resources discovery are improved by narrowing down their scopes to a manageable social network of everything [2]. Second, the scalability is guaranteed like in the human social networks [4]. Third, the heterogeneity of devices, networks and communication protocols is resolved by the use of social networks [2]. And a larger data source becomes available as it comes from a set of users. The continuous feeding of data from communities gives us big data [10]. Quantity and variety of contextual data have increased allowing improved services intelligence and adaptability to users' situational needs [2].

However, the SIoT paradigm does not allow to fix trust, security and privacy issues. Furthermore, the numerous advantages of SIoT such as improving navigability, and increasing quantity and variety of contextual data, make privacy and security of IoT users more compromised.

In the literature, trust mechanisms have been widely studied in various fields. Several works in the literature have dealed with this problem. They have proposed different trust-models, based on different features and measures, aiming to rank the best nodes in the SIoT network. Regarding the existing related works, our contribution in this paper are summarized as follow:

1. Unlike most existing reputation and trust management schemes in the literature, our goal is to detect malicious nodes. This allows us to isolate (or block)

the malicious nodes, limit the interactions made with them, and obtain a trustworthy system (network). Classifying trustworthy nodes would not prevent malicious ones from performing their malicious behaviors that could break the basic functionality of the given system.
2. To achieve the goals of ensuring a reliable and trustworthy system, we first present an informal description of each kind of trust-related attack. Then, we propose a new trust model based on new features derived from the description of each type of trust-related attack. Works in the literature use more global features such as the centrality of the node or the number of friends in common between two nodes. These features have no relation (from a semantic point of view) with the mentioned trust-related attacks.
3. To combine the proposed features, the majority of related works use the weighted mean. However, the performed behaviors for each type of trust attack are different. A weighted mean cannot detect all types of attacks since the features considered and the weights assigned to each feature may differ from one type of attack to another. We propose new features in our work and a new way to combine them using machine learning techniques, in order, to classify nodes into benevolent nodes and malicious nodes.

The rest of the paper is organized as follows. In Sect. 2, we present background about main concepts. In Sect. 3, we analyze and compare related works. In Sect. 4, We give a formal presentation of the proposed trust evaluation model. In Sect. 5, we detail the design of the proposed features. In Sect. 6, we detail the proposed classification function which allows to aggregate proposed feature in order to distinguish malicious behavior from benign ones. In Sect. 7, we present evaluations that enabled us to validate the resilience our trust evaluation model. Finally, we conclude in Sect. 8 with providing future insights.

## 2 Background

The Social Internet of Things paradigm allows people and objects to interact within a social framework to support a new kind of social navigation. The structure of the SIoT network can be shaped as required to facilitate the navigability, perform objects and services discovery, and guarantee the scalability like in human social networks. However, trust must be ensured for leveraging the degree of interaction among things.

Trust is a complicated concept used in various contexts and influenced by many measurable and non-measurable properties such as confidence, belief, and expectation on the reliability, integrity, security, dependability, ability, and other characters of an entity [20]. There is no definitive consensus about the trust concept in the scientific literature. Indeed, although its importance is widely recognized, the multiple approaches towards trust definition do not lend themselves to the establishment of metrics and evaluation methodologies.

Trust can be defined as a belief of a trustor in a trustee that the trustee will provide or accomplish a trust goal as trustor's expectation. In SIoT environment, trustors and trustees can be humans, devices, systems, applications, and services.

Measurement of trust can be absolute (e.g., probability) or relative (e.g., level of trust). The trust goal is in a broad understanding. It could be an action that the trustee is going to perform; it could also be an information that the trustee provides.

Trust management mechanisms and trust evaluation models are proposed to ensure trust in different types of systems. Their roles consist of providing (computing) a trust score, which will help nodes to take decision about invoking or not, services provided by other nodes. There are several varieties of attacks that are designed to specifically break this functionality. We present in this section the main trust-related attacks cited in the literature [1,5,7]. We also explain the differences between trust management mechanisms and trust evaluation models.

### 2.1 Trust Attacks in SIOT Networks

An attack is a malicious behavior established by a malicious node launched to break the basic functionality of a given system and to achieve a variety of malicious ends. A malicious node, in general, can perform communication protocol attacks to disrupt network operations. We assume such attack is handled by intrusion detection techniques [9,16] and is not addressed in our work.

In the context of SIoT, we are concerned about trust-related attacks that can disrupt the trust system. In this kind of attacks, a malicious node could boost its own reputation to gain access to higher functions or generally be disruptive in a manner that brings down the overall efficiency of the system. Thus, a malicious IoT device (because its owner is malicious) can perform the following trust-related attacks. We assume that there are some others attacks which can be autonomously launched by devices. We will consider them in future works. In this paper, we focus on attacks performed by IoT devices under the control of them malicious owners.

- **Self Promoting attacks (SPA):** is an attack where malicious nodes, provide bad-quality service, try to boost their reputation (by giving good rates for themselves) in order to be selected as service providers.
- **Bad Mouthing Attacks (BMA):** is an attack where malicious nodes try to destroy the reputation of well-behaved nodes (by giving them bad rates) in order to decrease their chance to be selected as service providers.
- **Ballot Stuffing Attacks (BSA):** is an attack where malicious nodes try to promote the reputation of other malicious nodes in order to increase their chance to be selected as service providers.
- **Discriminatory Attacks (DA):** is an attack where malicious nodes attack discriminatory other nodes, without a strong social relationship with them, because of human propensity towards strangers.

In Table 1, we propose an informal specification of the malicious behavior for each type of trust-related attack.

**Table 1.** An informal description of malicious behavior for each type of trust attacks.

| | Invoker ($u_i$) | Provider ($u_j$) | Interactions I($u_i, u_j$) |
|---|---|---|---|
| BMA | Malicious node:<br>- Provides poor quality services<br>- Provides bad votes that do not reflect his actual opinion to destroy the reputation of $u_j$ | Benign node:<br>- Has a good reputation<br>- Provides good quality services | - A lot of interaction<br>- The majority of votes provided by $u_i$ to $u_j$ are negative |
| BSA | Malicious node:<br>- It has a good reputation<br>- It gives high scores that do not reflect his actual opinion in $u_j$ in order to promote his reputation | Malicious node:<br>- Provides good quality services<br>- Has a bad reputation in the network | - A lot of interaction<br>- The majority of votes provided by $u_i$ to $u_j$ are positive |
| SPA | Malicious node:<br>- Provides services of poor quality<br>- Has a bad reputation in the network<br>- Provides high ratings that do not reflect his opinion to $u_j$ in order to promote reputation | Malicious node:<br>- Provides poor quality services | - A lot of interaction<br>- The majority of votes provided by $u_i$ to $u_j$ are positive<br>- $u_i$ and $u_j$ are often nearby<br>- They have same interests<br>- They provide same services |
| DA | Malicious node:<br>- Provides bad votes to the majority of other users | Malicious/benign node | The majority of votes provided by $u_i$ to $u_j$ are negative |

### 2.2 Trust Evaluation and Trust Management

Some researchers have focused on developing trust management mechanisms dealing with trust establishment, composition, aggregation, propagation, storage and update processes [11]. However, we focus in this work on the main step which is the trust establishment step. We will focus on the other steps in future works. The trust establishment step consists of developing a trust evaluation model and represents the main component of trust management mechanisms. Indeed, the performance of the trust management system essentially depends on the model introduced to evaluate the degree of trust that can be granted to the various entities involved in the system. We consider that a trust evaluation model is mainly composed of two steps, namely (i) the composition step and (ii) the aggregation step. The other steps such as propagation, updating, and storage will provide other properties such as system response time and scalability.

**(i) The Composition Step** consists of choosing features to be considered in calculating trust values. Several features have been considered in the literature such as honesty, cooperativeness, profile's similarity of profiles, reputation,... These features can be categorized into various dimensions: (i) global or local; (ii) implicit or explicit; or (iii) related to users, devices or provided services. To

measure these different features, the authors use information related to nodes, such as their position, their interaction history, their centrality in the network.

**(ii) The Aggregation Step** consists of choosing a method to aggregate values of different features in order to obtain the final trust value. For this purpose, authors in the literature use static weighted mean, dynamic weighted mean, fuzzy logic, probabilistic models, and so on.

## 3 Related Works

Various trust-models are proposed in the literature in order to ensure trustworthy services and interactions in SIoT environments. In this section, we try to analyze and compare these different models based on two criteria: (i) the proposed trust evaluation model; and (ii) the resilience face trust-attacks.

Trust evaluation models are composed of two steps, namely (i) **The composition step** and (ii) **The trust aggregation step**. For the trust composition step, authors propose different features such as recommendation, reliability, experience, and cooperativeness. Those features represent abstract concepts aiming to quantify the nodes trust level and are computed by different measures depending on authors goal and background. For example, in [14], the recommendation feature is measured as the number of nodes directly connected to a given node $u_i$. However, in [17], the recommendation feature is measured as the total mean of rates given to a node $u_i$. This same measure (mean of rates) is called reputation in some other works. The cooperativeness feature is considered as an indicator to measure a node's knowledge in [17] and is computed as the level of the social interactions between two nodes. However, in [7] the cooperativeness feature is computed as the number of common friends between two nodes.

Given that there is no consensus about trust concept definition, and given the divergence of the proposed features, as well as the divergence of the measure of each feature, this can give birth to thousands of trust evaluation models with different combinations between the features calculated with different measures. We believe that a trust evaluation model must above all fulfill the role of guaranteeing the reliability of the system in which it is involved. This reliability is compromised by the different types of trust-related attacks.

We have chosen in this work to start from the definition of each type of attack. We present an informal description of each attack that we formalized using mathematical measures and equations. We believe that some features and measures proposed in the literature, such as the number of friends in common or the number of relationships in the network, have no relation to the cited trust attacks. Moreover, as it is common in the classic social networks, a malicious node could try to increase the number of its relations in general or the number of its common relations with a given node, before proceeding to attacks. Some other measures, such as the mean rates, could give an idea about the history of a node's interactions and could, therefore, permit to detect some types of attacks. The features proposed in the literature remain insufficient to detect the different types of attacks. Indeed, none of the proposed features can detect, for example, the SPA attack in which a node is hidden under a false identity.

To conclude, the performance of a trust evaluation model mainly depends on the features and measures chosen in the composition phase. Nevertheless, it also depends on the method chosen in the aggregation phase. The weighted mean is most used aggregation method. However, the performed behaviors for each type of trust-related attack are not similar. A weighted mean cannot detect all types of attacks since the features considered and the weights assigned to each feature may differ from one type of attack to another. The problem of detecting malicious nodes being considered as a complex problem and requiring an in-depth analysis of nodes behaviors, and thus, we propose the use of machine learning techniques.

The second criterion of comparison concerns the resilience to trust-related attacks. Some of the cited works focus on trust-attack detection. However, they do not prove the ability of the proposed model to detect trust-attack through evaluations or experimentation. The majority of related works propose model permitting to assign a trust degree to each node in the network. Their goal is to rank nodes according to their trust-values. However, this kind of model does not allow to detect malicious attacks and malicious nodes. This gives the malicious nodes free access to establish different types of attacks in the network. The purpose of our work is to detect malicious nodes in order to block them and obtain a trustworthy system (Table 2).

**Table 2.** Comparison of related-works

| | | [18] | [13] | [17] | [7] | [15] | [8] |
|---|---|---|---|---|---|---|---|
| Trust evaluation model˜ | Trust composition | Knowledge reputation experience | Consistency intention ability | Recommend reputation experience | Honesty coopertiveness community-interest | Reliability reputation | Reputation, social relationship, energy-level |
| | Trust agregation | Fuzzy logic | Weighted mean | Fuzzy logic | Combinatorial logic | Weighted mean | Weighted mean |
| Goal | Node ranking | ✓ | ✓ | ✓ | | ✓ | |
| | Attack-detection | | | | ✓ | | ✓ |

# 4 System Modeling

## 4.1 Notations

Let $G_u^t = (U, S)$ be a directed graph representing users social network at time t. U is the node set $U = \{u_i | 0 < i \leq n\}$ where n is the number of users. S is the edges set and reflects the friendship relation between users. Each user $u_i \in U$, can be modeled by the 5-tuple $\langle id, age, city, country, devices \rangle$ where *devices* represents the list of devices that belong to the user $u_i$.

Let $G_d^t = (D, R)$ be a directed graph representing devices network at time t. D is the node set $D = \{d_j | 0 < j \leq m\}$ where m is the number of devices. R is the edges set where $R \in \{or; sr, wr, lr, pr\}$ represents the different kinds of

relations which can occur between devices. *or* represents the owner relationship which occurs between two devices having the same owner. *sr* represents the social relationship which occurs between two devices when their owners have a social relationship. *lr* represents the relation between two devices which are in proximity. *wr* represents the relation between two devices which interact to perform a common task. *pr* represents the relation between two devices which belong to the same category.

Each device $d_j \in D$, can be modeled by the 5-tuple $<id, category, longitude, latitude, services>$ where *services* represent the list of services provided by the device $d_j$. Each service $s_k$ is modeled by the 4-tuple $<id, endpoint, domain, qos>$ where qos (Quality of Service) represents the service's non-functional characteristics such as its availability, response time, latency, and reliability.

### 4.2 Problem Definition and Formalization

With the presented notations and definitions, our main problem is to detect malicious users. In a formal way, given a training set of N instances of users associated with a class label $l_i \in$ {malicious, bening}, the problem is turned first to design an advanced set of M features extracted from the training set. Then, the features designed are used to learn or build a binary classification model $y$ using a given training set such that it takes features X as an input and predicts the class label of a user as an output, defined as $y : u_i \rightarrow$ {malicious, bening}.

## 5 Features Design

In this section, we present the composition step of our trust evaluation model. We propose new features permitting to describe and quantify the different behaviors operating in SIoT systems. Our features are derived from the informal description of each type of trust-related attack and allow to distinguish malicious behavior from benign ones.

### 5.1 Reputation

This feature represents the global reputation of a user $u_i$ in the overall network and is denoted as $Rep(u_i)$. It is computed as the quotient between the number of positive interactions and the total number of interactions (Eq. 1). Positive interactions are interactions with a high rate value. Nodes with a high reputation value are more likely to be attacked by other nodes. Nodes with a low reputation value are more likely to perform trust attacks. The reputation feature, combined with other features, can help in revealing BMA, BSA, SPA and DA attacks.

$$Rep(u_i) = \frac{\sum_{s_k \in S(u_i), (rt(u_j, s_k) >= 3)} m}{|I(u_i, u_j)|} \tag{1}$$

where $rt(u_j, s_k)$ is the rate given by the user $u_i$ to the service $s_k$ and $I(u_i, u_j)$ the set of interactions occurred between $u_i$ and $u_j$.

### 5.2 Honesty

Honesty represents whether a user is honest and is denoted as $Hon(u_i)$. A user is considered honest if his rates reflect his real opinion, which means that he doesn't try to give wrong rating values to enhance or decrease other users reputation. Indeed, in BMA, BSA and SPA attacks, the malicious node presents a dishonest behavior. In the BMA attack, the malicious node gives bad votes to a node that provides good quality services, in order to ruin its reputation. In the BSA attack, the malicious node gives good votes to another malicious node that provides poor quality services, with the aim of helping it to promote its reputation. In the SPA attack, the malicious node tries to promote its own reputation by giving itself good votes while its services have poor quality.

The Honesty feature is, therefore, a key feature, which associated with other features, may reveal different types of attacks. To measure and quantify this feature, we compare the user rating vector $Rvec(u_i)$ with the rating matrix using Cosine Similarity Measure (Eq. 2).

$$Hon(u_i) = \sum_{x_j \in Rvec(u_i), x'_j \in Mvec} \sqrt{(x_j - x'_j)^2} \tag{2}$$

Where $x_j$ is the rate of $i^{th}$ user on $j^{th}$ item, $x'_j$ the average of rates given by all network nodes on item j, $Rvec(u_i)$ is the rating vector of the $i^{th}$ node and Mvec is the mean rating vector representing the average of the rating matrix

### 5.3 Quality of Provider

Quality of provider represents whether services provided by the user $u_i$ present a good or bad QoS. It is denoted as $QoP(u_i)$. Indeed, malicious node aims at propagating services with bad quality. Services with good quality naturally reach a good reputation in the network. The malicious node must resort to malicious behavior to propagate bad services and will, therefore, perform BMA, BSA, SPA and DA attacks to achieve this goal. QoP feature is therefore essential to distinguish the nodes that likely perform malicious behaviors from other nodes that provide good quality services and do not need to carry out attacks to propagate them.

$$QoP(u_i) = \frac{\sum_{d_j \in D(u_i), s_k \in S(d_j))} QoD(d_j) * QoS(s_k)}{\sum_{d_j \in D(u_i)} QoD(d_j)} \tag{3}$$

where $S_{u_i}$ is the set of services provided by the $i^{th}$ node, qos($s_k$) is the QoS value of the service $s_k$, and $\alpha$ is a threshold.

### 5.4 Similarity

Similarity refers to the similarity between user $u_i$ and user $u_j$ and it is denoted as $sU(u_i, u_j)$. This feature is computed based on different features such as profiles,

interests, provided services, used devices and the frequency of proximity between a couple of users. It aims to detect affinity between users but can also reveal Self-Promoting Attack (SPA) in which the same user tries to promote his own reputation under a false identity.

### 5.5 Rating-Frequency

Rating-Frequency refers to the frequency of rating attributed by a user $u_i$ to a user $u_j$, denoted as $RateF(u_i, u_j)$. It is computed as the number of rates given by a user $u_i$ to a user $u_j$ divided by the total number of rates given by the user $u_i$. Indeed, if a user $u_i$ performs an attack against a user $u_j$, we will probably find a high number of rates given by user $u_i$ to user $u_j$. According to whether these rates are positives or negatives and according to some other features such as the reputation and the QoP of the target user $u_j$, we can detect a Ballot-Stuffing Attack or a Bad-Mouthing Attack.

### 5.6 Direct-Experience

Direct-Experience refers to the opinion of a node i about its past interactions with a node j, denoted as $dExp(u_i, u_j)$. It is computed as the quotient of successful interactions between node $u_i$ and node $u_j$, divided by the total number of interactions between them. The direct experience feature can not therefore directly reveal an attack. But, combined with other features, it helps to distinguish what kind of attack it is. Indeed, taking the example of two nodes $u_i$ and $u_j$ where $u_i$ is a node that provides bad services and therefore has a low QoP value. The Rating frequency value $RateF(u_i, u_j)$ shows that the node $u_i$ is striving to give rates to the $u_j$ node. Indeed, $u_i$ gives a total of 10 votes, of which 6 are attributed to $u_j$. In this case, it is probably an attack. Other features, such as $u_j$'s reputation and QoP, as well as $u_i$'s honesty, may confirm this hypothesis. The direct experience feature can finally decide whether it is a BMA or BSA attack. Indeed, in the BMA attack, node $u_i$ aims to ruin the reputation of $u_j$ and will, therefore, provide negative rates which result a low value of $dExp(u_i, u_j)$. Whereas, in the BSA attack, the node $u_i$ aims to promote the reputation of $u_j$, which result a high value of $dExp(u_i, u_j)$.

### 5.7 Rating Trend

The rating trend feature is measured by the number of positive votes divided by the total number of votes provided by a user. It aims to reveal if a user is rather optimistic or pessimistic. It permit to detect the discriminatory attack (DA) in which the user provides negative votes randomly.

## 6 Classification Function Design

Once we have chosen the features that describe the behavior of different nodes in the network, the next step consists in choosing a method to aggregate the values

of the different features, in order to obtain the final trust value. In the literature, the most common method is the weighted mean. However, we estimate that the performance of the system depends in this case mainly on weights assigned to each feature. Furthermore, the performed behaviors for each type of trust-related attack are different. A weighted mean cannot detect all types of attacks since the features considered and the weights assigned to each feature may differ from one type of attack to another.

The problem of the detection of malicious nodes being considered as a complex problem and requiring an in-depth analysis of nodes behaviors, we propose to use machine learning techniques. To our knowledge, this technique has never been used to measure trust. We consider our system as a classification problem. Indeed, our objective is to detect if a user is malicious or benign. A user is considered malicious if he tries to perform BMA, BSA, SPA or DA attack. If the user didn't perform any of the cited attacks, he is considered as benign. So, for each users $u_i$, we have two possible classes, namely (i) malicious user class, (ii) benign user class.

Machine learning techniques allowed us to avoid the problem of fixing weights and thresholds. Indeed, the machine learning algorithm will take as input the proposed features, will automatically assign the weights based on the learning data-base and will return as output one of the mentioned classes. The model as proposed in this work does not allow to determine the type of performed attack, but only to detect whether there was an attack or not. We plan in our future work to improve the proposed model, since some attacks may be more dangerous than others depending on the context and the domain. It would be interesting, in this case, to be able to know the type of attack.

# 7 Results and Evaluations

## 7.1 Experimental Setup

**Data-Set Description.** Due to the unavailability of real data, the majority of related works offer experiments based on simulations. In our work, we evaluated the performance of our model based on experiments applied to an enriched real dataset. Sigcomm[1] data-set contains users, their profiles, their list of interests. It contains also social relations between users, interactions occurred between them and frequency of proximity of each couple of users. We generate for each user one or more devices and we divide interactions of a user by his devices. Figure 1 shows statistics and description of the resulting data-set.

**Performance Metrics.** To assess the effectiveness and robustness of our proposed features using machine learning algorithms, we adopt the accuracy and the standard existing information retrieval metrics of precision, recall, and f-measure.

[1] http://crawdad.org/thlab/sigcomm2009/20120715/.

| | | |
|---|---|---|
| Users | | 75 |
| Users profiles | Institute | 75 |
| | City | |
| | Country | |
| Users Interests | | 711 |
| Social relationship between users | | 531 |
| Interaction between users | | 32 000 |
| Devices | | 300 |
| Services | | 364 |
| Proximities | | 285 788 |

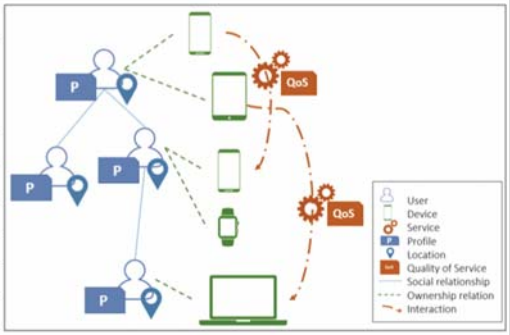


**Fig. 1.** Data-set description

**Learning Methods.** We used the different learning algorithms implemented in WEKA [12] tool, to build the binary classification function $y$. We report here the results of Naive Bayes, Multi-Layer Perceptron and Random Tree learning algorithms (see Fig. 2). We finally opt for the Multi-Layer Perceptron because it has shown the best results in terms of the evaluation metrics. We used 10-fold cross-validation algorithm to evaluate the performance of our features.

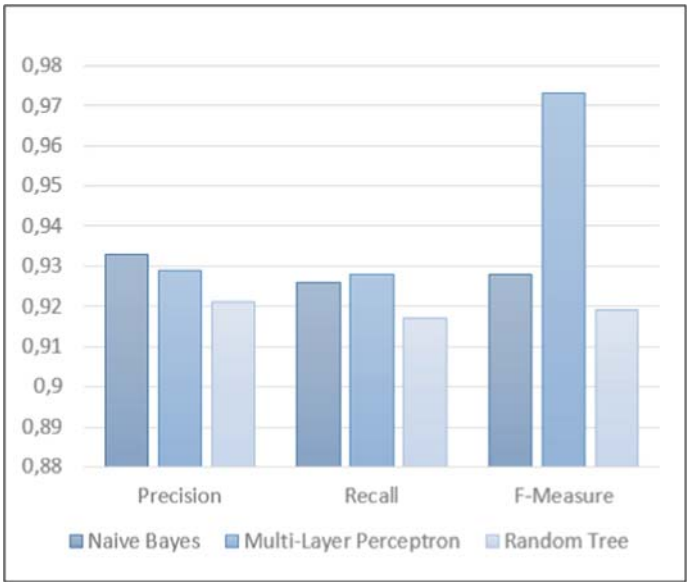


**Fig. 2.** Comparison of machine learning algorithms

**Experiments Procedure.** In this work, we propose a trust evaluation model. For this, we proposed, first, new features and measures to describe the behavior of different users. Secondly, we propose to use a new method of aggregation based on machine learning, able to differentiate malicious users from legitimate users.

To prove the performance of each proposed feature, we first measure the information gain for each feature separately. Then, we compare the features

that we propose with the most used features in the literature based on common evaluation measures such as recall, precision, and F-measurement. For this, we use the most used aggregation method in the literature which is the weighted mean.

To prove the performance of the proposed aggregation method (Machine Learning), we compare (i) the results obtained by the other works with the weighted mean, (ii) our characteristics with the weighted mean and (iii) our characteristics with Machine Learning. Finally, to prove the resilience of the proposed trust evaluation model, we measure the proportion of malicious nodes detected on different networks with different percentages of malicious nodes ranging from 10 to 50%.

### 7.2 Experimental Results

**Single Features Performance.** The Fig. 3 shows the information gain when using one single feature in the learning process. The *similarity* feature has the largest value of information gain. This can be explained by the fact that this is the only feature able to detect Self Promoting Attacks (SPA). *Rating frequency*, *quality of provider*, *rating trend*, *honesty* and *reputation* features present almost equal information gain values. Indeed, they are equally discriminative for the detection of BMA, BSA and DA attacks type. The direct experience attribute has the lowest information gain value. This attribute does not actually detect attacks. But, it allows, as explained previously, to make the difference between a BMA and a BSA attack.

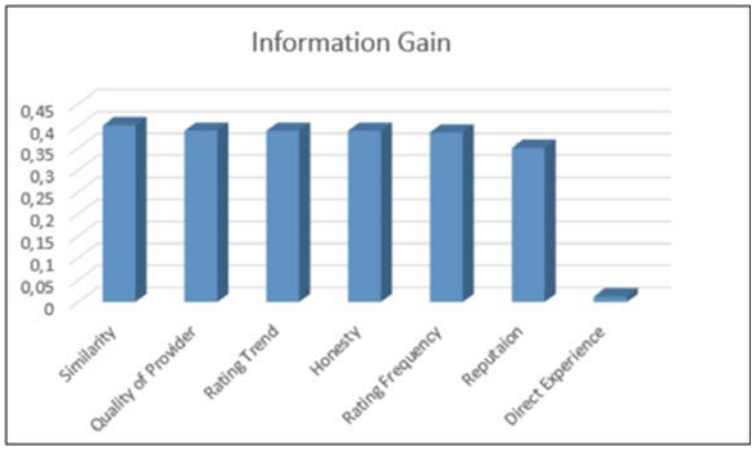


**Fig. 3.** Evaluation of single features performance

**Group Features Performance.** We compare our 7 features with 10 features existing in the literature. We implemented the 10 features with experimenting them on our data-set to have fare comparison. Since related-works that propose these features use the weighted mean for the aggregation step, we had to try

different weights for each feature used in each related-work. We selected the weights that gave the best results for each work (see Table 3).

In addition, many of the proposed trust evaluation models have objectives of classifying nodes according to their trust degrees, without detecting malicious nodes. So, we had to set the thresholds below which a node is considered as malicious. We have similarly tried different threshold values for each of the related-works and we have chosen the thresholds that give the best results for each model. Table 3 shows the weight and threshold values we have finally selected for each related-work.

**Table 3.** Parameters for group feature performance evaluation

| Related works | Features | Weights | Thresholds |
|---|---|---|---|
| Jayasinghe, U., et al. 2016 [14] | Recommendation<br>Reputation | 0,62<br>0,38 | 0,30 |
| Truong, N.B., et al. 2018 [19] | Reputation,<br>Experience | 0,84<br>0,16 | 0,22 |
| Chen, R., et al. 2016 [7] | Honesty<br>Coopertiveness<br>Community-Interest | 0,74<br>0,12<br>0,14 | 0,22 |
| Militano, L., et al. 2016 [15] | Reliability<br>Reputation | 0,37<br>0,63 | 0,35 |
| Our features with weighted mean | Honesty<br>Reputation<br>Similarity<br>Direct Experience<br>Rating frequency<br>Quality of Provider<br>Rating trend | 0,18<br>0,19<br>0,1<br>0,06<br>0,19<br>0,18<br>0,1 | 0,58 |

We then used the weighted mean for the features we propose in this work. This allowed us to compare and validate the relevance of the features we propose compared to those of the state of the art. Finally, we applied the machine learning on the features that we propose in this work. This allowed us to prove the relevance of the aggregation method that we propose (the machine learning) compared to the most used method in the literature (the weighted mean). Figure 4 shows the results obtained. The features we propose give better results in terms of recall, precision, and f-measurement compared to other works even in the case of aggregation with a weighted mean. The results are even better when we applied the machine learning technique for the aggregation step.

**System Resilience.** Figure 5 presents the proportions of malicious nodes being detected obtained for an increasing number of malicious nodes performing randomly all kinds of trust-related attack. The proportion remains high (89%) even

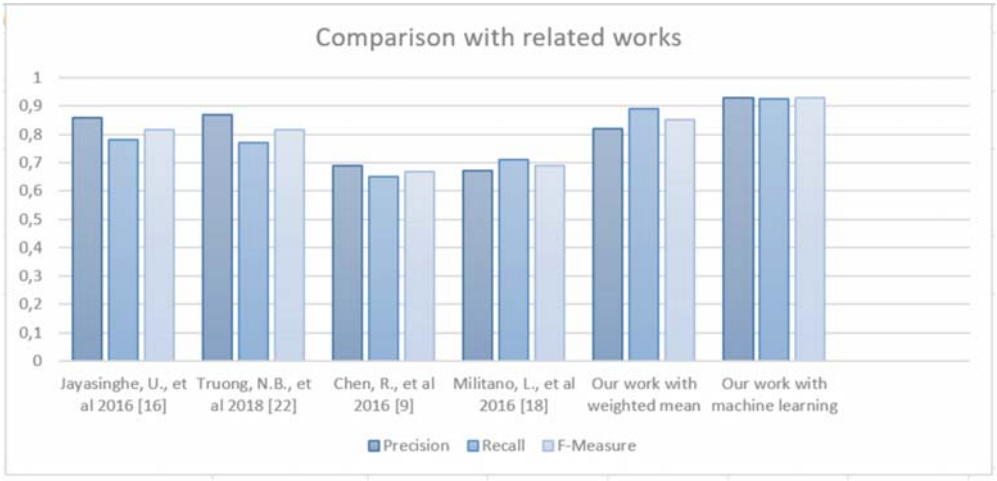


**Fig. 4.** Comparison with related works

for a system when 50% of the nodes are malicious. Those evaluations prove that our system can ensure resiliency toward each kind of trust-related attack, even facing a high percentage of malicious nodes. We have not continued the assessments for higher percentages, for the simple reason that, if a system has more than 50% malicious nodes, then it is a defective system.

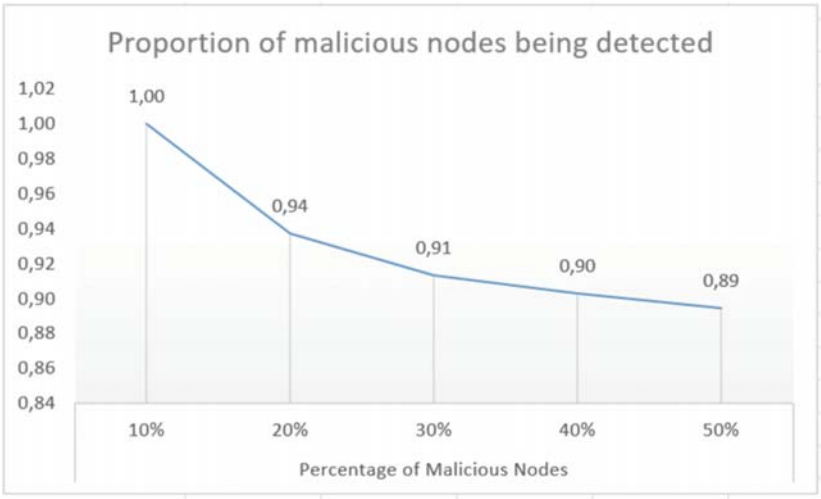


**Fig. 5.** Proportion of malicious nodes being detected

## 8 Conclusion and Future Directions

Social Internet of Things (SIoT) is a paradigm where the Internet of Things (IoT) is fused with Social Networks, offering a variety of attractive services and applications. However, facing this emerging paradigm, people feel wary and cautious. They worry divulgence of their data and violation of their privacy.

In this work, we propose a new trust-evaluation model, able to detect malicious nodes, in order to obtain a reliable and resilient system. Our future prospects are to develop a trust-management mechanism based on the proposed trust-evaluation model. This mechanism must ensure not only trust establishment but also the propagation, storage, and updating of trust. This will raise new issues related to the specific characteristics of SIoT environments, such as the scalability, dynamism and constrained capabilities of IoT devices.

## References


1. Abdelghani, W., Zayani, C.A., Amous, I., Sèdes, F.: Trust management in social internet of things: a survey. In: Social Media: The Good, the Bad, and the Ugly, pp. 430–441. Swanwea (2016)
2. Ali, D.H.: A social Internet of Things application architecture: applying semantic web technologies for achieving interoperability and automation between the cyber, physical and social worlds. Ph.D. thesis, Institut National des Télécommunications (2015)
3. Atzori, L., Iera, A., Morabito, G.: From "smart objects" to "social objects": the next evolutionary step of the internet of things. IEEE Commun. Mag. **52**(1), 97–105 (2014)
4. Atzori, L., Iera, A., Morabito, G., Nitti, M.: The social internet of things (SIoT)-when social networks meet the internet of things: concept, architecture and network characterization. Comput. Netw. **56**(16), 3594–3608 (2012)
5. Bao, F., Chen, I., Guo, J.: Scalable, adaptive and survivable trust management for community of interest based internet of things systems. In: 11th International Symposium on Autonomous Decentralized Systems, Mexico City, pp. 1–7 (2013)
6. Calvary, G., Delot, T., Sedes, F., Tigli, J.Y.: Computer Science and Ambient Intelligence. Wiley, Hoboken (2013)
7. Chen, R., Bao, F., Guo, J.: Trust-based service management for social internet of things systems. IEEE Trans. Dependable Secur. Comput. **13**(6), 684–696 (2016)
8. Chen, Z., Ling, R., Huang, C., Zhu, X.: A scheme of access service recommendation for the social internet of things. Int. J. Commun. Syst. **29**(4), 694–706 (2016)
9. Cho, J.H., Chen, R., Feng, P.G.: Effect of intrusion detection on reliability of mission-oriented mobile group systems in mobile ad hoc networks. IEEE Trans. Reliab. **59**(1), 231–241 (2010)
10. Geetha, S.: Social internet of things. World Sci. News **41**, 76 (2016)
11. Guo, J., Chen, R., Tsai, J.J.: A survey of trust computation models for service management in internet of things systems. Comput. Commun. **97**, 1–14 (2017)
12. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
13. Huang, J., Seck, M.D., Gheorghe, A.: Towards trustworthy smart cyber-physical-social systems in the era of internet of things. In: 2016 11th System of Systems Engineering Conference (SoSE), pp. 1–6. IEEE (2016)
14. Jayasinghe, U., Truong, N.B., Lee, G.M., Um, T.W.: RpR: a trust computation model for social internet of things. In: Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, pp. 930–937. IEEE (2016)

15. Militano, L., Orsino, A., Araniti, G., Nitti, M., Atzori, L., Iera, A.: Trusted D2D-based data uploading in in-band narrowband-IoT with social awareness. In: 2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1–6. IEEE (2016)
16. Mitchell, R., Chen, I.R.: A survey of intrusion detection techniques for cyber-physical systems. ACM Comput. Surv. (CSUR) **46**(4), 55 (2014)
17. Truong, N.B., Um, T.W., Lee, G.M.: A reputation and knowledge based trust service platform for trustworthy social internet of things. Innovations in Clouds, Internet and Networks (ICIN), Paris, France (2016)
18. Truong, N.B., Um, T.W., Zhou, B., Lee, G.M.: From personal experience to global reputation for trust evaluation in the social internet of things. In: GLOBECOM 2017–2017 IEEE Global Communications Conference, pp. 1–7. IEEE (2017)
19. Truong, N.B., Um, T.W., Zhou, B., Lee, G.M.: Strengthening the blockchain-based internet of value with trust. In: International Conference on Communications (2018)
20. Yan, Z., Zhang, P., Vasilakos, A.V.: A survey on trust management for internet of things. J. Netw. Comput. Appl. **42**, 120–134 (2014)