Article

Not peer-reviewed version

Defense Mechanism Against Attacks Promoting Spread Of Wrong Information

Saba Mahmood , Farah Akif , Humaira Ashraf , NZ Jhanjhi *

Posted Date: 8 January 2024

doi: 10.20944/preprints202401.0633.v1

Keywords: Content Credibility; Reputation Attacks; Sybil; Slander; Whitewash

Article

Defense Mechanism Against Attacks Promoting Spread Of Wrong Information

Saba Mahmood [1], Farah Akif [2], Humaira Ashraf [3] and NZ Jhanjhi *

[1] Department of Computer Science, Bahria University Islamabad; saba.mahmood@gmail.com
[2] Department of Electrical Engineering IIU Islamabad; farah.mahmood@gmail.com
[3] Department of Computer Science and Software Eng IIU Islamabad; humaira.ashraf@iiu.edu.pk
[4] School of Computer Science, SCS, Taylors University, Malaysia
* Correspondence: noorzaman.jhanjhi@taylors.edu.my

Abstract: The rise in dependency upon information present on web and social networks has increased the importance of content credibility systems. These sys- tems can help the users to make a right decision related to buying a product, utiliz- ing a service and etc. Application areas including social network blogs of different subjects such as health, food, education, politics and product review/ratings some- times suffer incredible and wrong information flow. The content credibility sys- tems can help users to identify the credible information on these various forums. Recently, researchers have proposed certain mechanisms for these systems, out of which reputation based systems have gained most attention. However, reputa- tion systems are vulnerable to reputation based attacks, like Sybil, Slandering, and Whitewash promoting spread of wrong information. The authors have proposed a defense mechanism against these attacks that is based upon Bayesian Model. The proposed defense mechanism provides protection against these attacks so that fake or wrong information could be prevented. The authors evaluated the proposed mechanism in three different scenarios and presented the results in terms of pre- cision, recall and rate of change in rank. The results reveal almost 88%prevention against these attacks in comparison to the baseline systems.

Keywords: Content Credibility; Reputation Attacks; Sybil; Slander; Whitewash

1. Introduction

The content present on web is not reliable and credible [1][2]. The content can be a piece of text, product reviews, user ratings, recommendations, QA blogs to name a few. Reputation systems are utilized in different domain, it is built upon the success or failure of interactions. Recently Reputation systems [3] are employed to ascertain the credibility of reviews, ratings, rankings and recommendations. The importance of content credibility systems [4][5] have emerged due to a rise in reliance on the information present on web by the people from every age, background and knowledge level. Anyone with or without expert knowledge can contribute on the web in the form of social networks blogs, emails, web content.

An approach towards evaluating content credibility is to evaluate the ranking and rating associated with the contributor of that information. Reputation Systems[6]are widely employed to calculate these ratings or ranks. However, these rating and rankings can be manipulated through reputation attacks [7], that can increase the reputation rank of fake or wrong content. These attacks can be launched on the information dissemination, calculation structure of the reputation systems. These attacks on reputation based systems are also evaluated in cloud environment [8], IoT Systems [9] and social networks [10]. Reputation attacks investigated in a recent research [11] has studied the impact of the ratings and reviews of the users. One approach to mitigate these attacks is to identify the Sybil identities or dishonest nodes. This involves content based analysis[12] of the nodes, that is hard to capture and is sometimes not revealed. Another approach involves graph based analysis to find attack edges. This article however, has proposed an approach that strengthens the calculation structure of the reputation model that can filter the malicious activity while preserving privacy of the nodes. In this article the authors have focused on the attacks that are specifically launched to explore

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2024 doi:10.20944/preprints202401.0633.v1

the vulnerability of the mathematical model of the reputation systems utilized to judge credibility of web content.

Attacks on reputation management systems are directed at the system's objective. The capacity to ap- propriately portray a participant's ranking based on their interactions with other participants is the aim of reputation systems. However, under these assaults, the system is unable to generate correct ranks, conse- quently rewarding dishonest individuals and discouraging the truthful ones. The attack model is carried out by insiders who are authenticated to the system. This paper identifies the popular attacks on content credibil- ity systems, that are Sybil, Slandering and Whitewash attacks. The existing reputation based approaches of content credibility do not identify these attacks, and are easily vulnerable to these attacks. The existing con- tent credibility systems are based upon simple summation [13], pagerank [14] and normal distribution(NDR) [15] based reputation calculation structures. More recently an iterative filtering(IF) based reputation algo- rithm [16] has been proposed to address the issue of unfair ratings. This algorithm is based upon updating of weight by calculating the distance of a user's rating from the aggregate. This article proposes the defense algorithms to counter these attacks as part of the content credibility framework that is based upon Bayesian algorithm utilizing beta probability density function [5]. Finally efficacy of the proposed mechanism is demonstrated through simulation experiments in different scenarios. The article contains sections of Prob- lem, Research Objectives, Literature Review, followed by Experiments evaluating different attack scenarios and finally the Conclusion.

2. Problem Formulation

Let $U = \{u_1, u_2, u_3, \cdots, u_n\}$ has interactions I where $I = \{i_1, i_2, i_3, \cdots, i_n\}$ are the interactions that U had, I is categorized as positive or negative i.e $i = alpha/beta$.Reputation rank is calculated through the [5] algo- rithm,as $R = \{r_1, r_2, r_3, \cdots, r_n\}$,in such a way that element of U are in sorted manner according to R.Attacker A manipulates the R of U such that elements of U are not in sorted manner now.$A = \{a_1, a_2, a_3, \cdots, a_n\}$ be- comes subset of U thereby increasing rank of u_k in case of Sybil attack or decreasing rank of u_k in case of Slander attack. The problem scenario is graphically presented in the Figure 1.

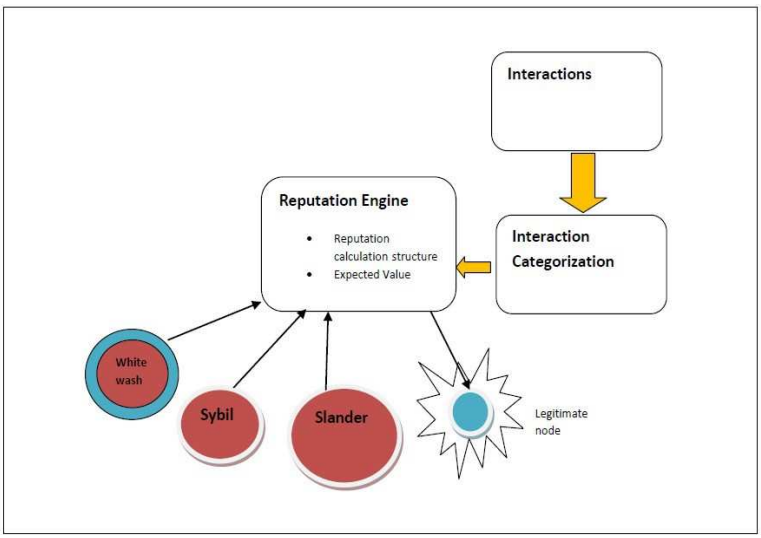

Figure 1. The Attack Model Scenario.

2.1. Research Contribution

The Content credibility systems based upon reputation systems are vulnerable to reputation attacks. Thus content credibility systems require a defense module that can prevent such attacks. Following objec- tives are addressed by authors in this article.

- To protect the reputation ranking from the Sybil attacks.
- To design prevention technique against the Slandering attack whereby the attacker

 doi:10.20944/preprints202401.0633.v1

intends to nega- tively impact the reputation rank of the target node.

- To prevent Whitewash attack that is launched as either sybil or slandering attack. The attacker gains a good rank initially and after sometime it starts attacking as a sybil or slander.

3. Literature Review

3.1. Threat Model

An attacker in the context of reputation based credibility systems is someone who has intention to reduce the reputation score of an agent, user, person, expert commenting or recommending a piece of infor- mation. Attacks under discussion in this paper are categorized as active attacks. For example an entity gain higher rank and is thus regarded as subject expert with greater reputation rank. In order to tarnish the rank, an adversary introduces many fake identities in the system. The purpose of the act is to increase the rating of the attacker while decreasing the rating of the genuine user. This attack can be targeted to decrease the rank of the genuine expert by producing negative reviews, or ratings. Another type of attacker starts of with passive behavior, after gaining some reputation through positive behavior, it starts attacking behavior by pro- ducing negative comments/reviews for the target thereby decreasing its reputation.These attackers exploit the calculation structure of the reputation systems. The calculation structures adopted by content credibility systems are based upon page rank, normal distribution(NDR) and simple summation. Mathematically NDR is presented as,

$$NDR = \sum_{l=1}^{k} (l * LW) \tag{1}$$

where LW is the weighing factor. PageRank based reputation is mathematically represented as,

$$PR(u) = \sum_{v}^{k} \frac{PR(v)}{L(v)} \tag{2}$$

where page rank value of page u is dependent upon the pagerank value of each page in set B divided by the total number of pages. The reputation model in [13] is simple summation of the total number of followings, tweet and likes. While the mathematical representation of [5] is

$$E \quad M-1 \quad \frac{p+1}{} = \sum_{i=1} \frac{p+n+2}{} \tag{3}$$

where E represents the expected value(reputation) of a node in a particular context.The p represents the positive interactions and n represents negative interactions.The proposed defense mechanism is based upon this calculation structure. In the following sections authors have elaborated the three types of attacks in the context of reputation calculation structures.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2024 doi:10.20944/preprints202401.0633.v1

3.2. Types of Attacks

3.2.1. Self Promotion or Sybil Attack

Sybil refers to fake identities. Sybil can launch an attack by demeaning a reputed identity by false accusation. Initially Sybil gains some reputation in the system, by fabricating its behavior. An extreme kind of Sybil attack would be, where multiple Sybil identities are created. One scenario of a Sybil attack is when Sybils form a group to give false positive feedback about a node, just to increase the reputation [17] of that particular node. In a different scenario, a Sybil account could generate fraudulent accounts to harm a legitimate user's reputation or rating. Today, online bullying, etc., are all examples of this kind of attack. These attacks have recently been discussed in relation to social media, particularly twitter. [18],[19]. Regression modelling has been used by the authors of this work to make predictions about the user profile.

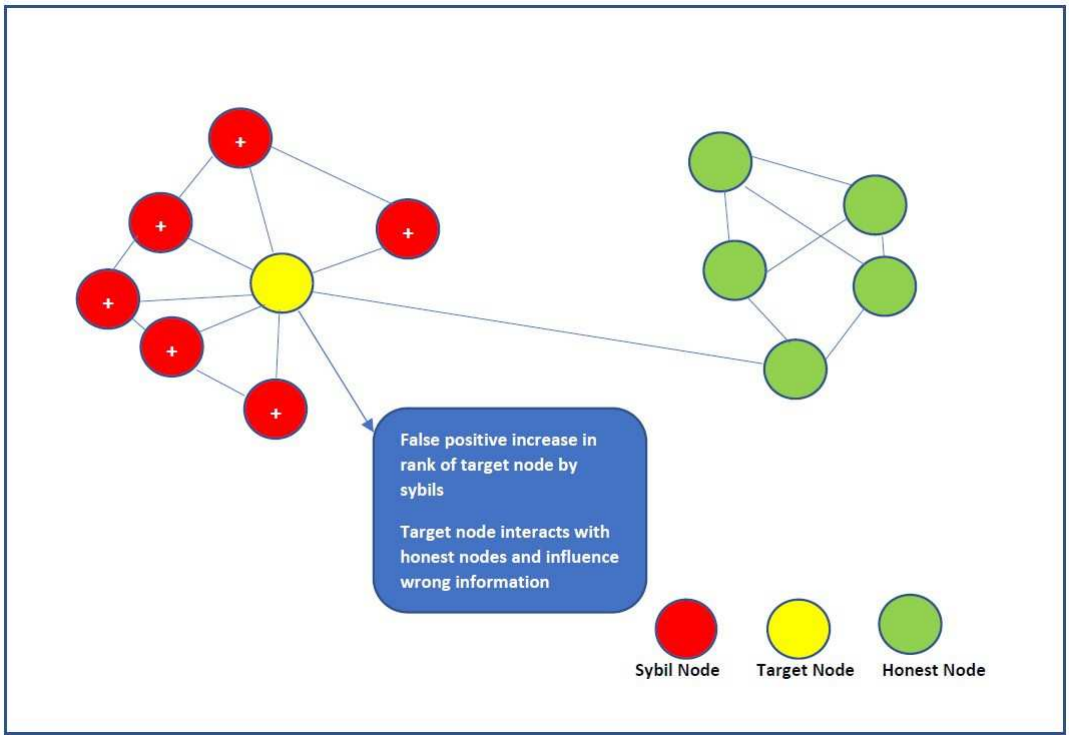

Figure 2. The Sybil Attack.

More recently authors utilized game theoretic [20] approach to mitigate the Sybil malicious behavior by addressing issue of dynamic update of trust value and trust threshold. The Figure 2 elaborates the Sybil attack scenario, whereby false positive increase in the rank of the target node promotes malicious activity.

3.2.1. White Washing

In white wash attack , the node tries to whitewash its bad behavior . The node initially gains good reputation through positive interactions. Once trust is established it behaves maliciously . This attack is coupled with Sybil or Slander attacks.

Literature reveals systems including Sybil Guard[21], SumUp[22], Sybil Limit[23] fighting against these attacks. These systems are not suitable for the content credibility systems.

Another study on preventing Sybil attacks in expert ranking systems, the SumUp algorithm is used by MHITS[24]. In this system, the nodes are eliminated by the SumUp approach prior to the ranking procedure. SumUp is an online content rating system that resists Sybil attacks by relying on the user trust network. It makes advantage of the idea of max-flow. The peer prediction approach, which has been the subject of research in the field of defence mechanisms, is related to receiving

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2024 doi:10.20944/preprints202401.0633.v1

honest feedback from users [25, 26]. In a Nash equilibrium fashion, they offer an appropriate payment scheme to nodes providing accurate reports about others. Theoretically, these systems have attempted to combat these attacks by observing behaviour related to cost incurred in providing opinions.

3.2.3. Slandering

Positive users may receive negative comments from malicious users, harming the reputation of deserv- ing individuals. The influence of a single defamatory node is minimal; however, when nodes band together to harm a node's good reputation, it can have an effect. The recognised slander node is penalised by standard defence measures. As a precautionary step, attaching nodes to genuine transactions or interactions is also possible. The slander assault scenario is further described in the figure **??**. The slander node defames by encouraging fictitious negative interactions with the honest node.

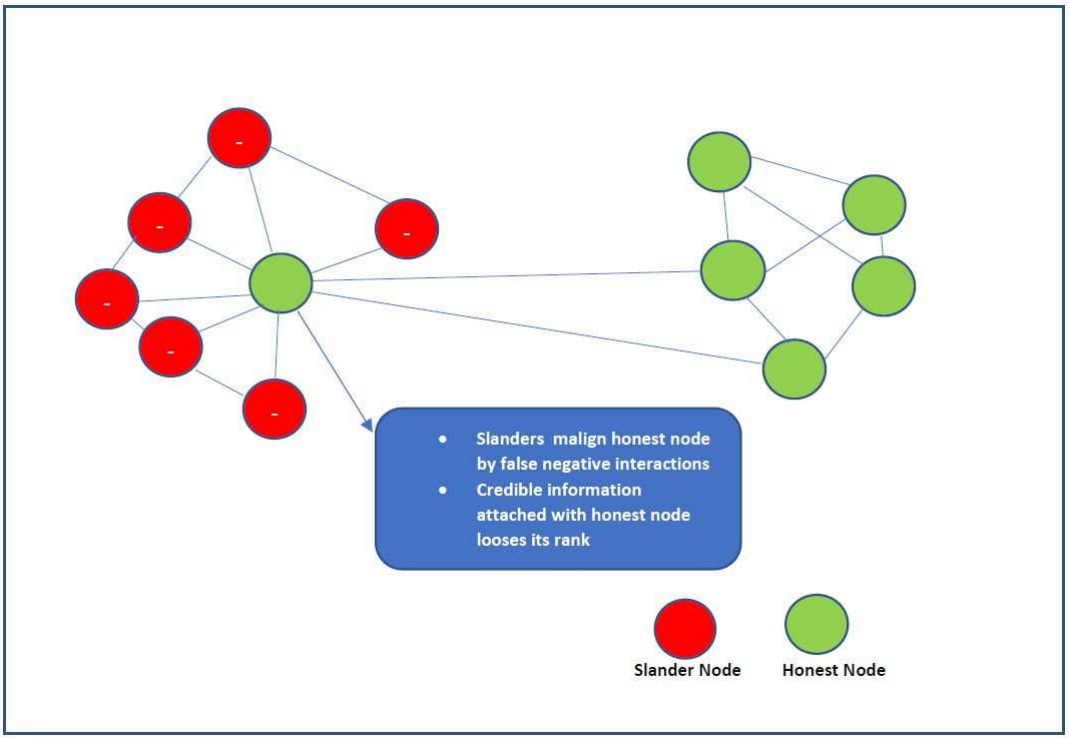

Figure 3. The Slander Attack.

4. Defense Mechanism

The scheme presented in this article utilizes Bayesain based reputation model[5] . The Bayesian repu- tation system utilizes beta probability density function and takes into account positive and negative interac- tions.

$$f(p \mid \alpha, \beta) = \frac{\Gamma(i.\alpha + i.\beta)}{\Gamma(i.\alpha)\Gamma(i.\beta)} p^{(1-i.\alpha-1)}(1-p)^{(i.\beta-1)} \quad (4)$$

The probability expectation value is given by, where alpha and beta are two events

$$E(p) = \frac{\alpha}{\alpha + \beta} \quad (5)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2024 doi:10.20944/preprints202401.0633.v1

Suppose there are two outcomes of interaction i named as α and β then mathematically we can write it as

$$i.\alpha,\ i.\beta \tag{6}$$

If the observed number of outcomes are denoted by n then it can be expressed as ,

$$i_n.\alpha,\ i_n.\beta \tag{7}$$

$$\alpha = i_n.\alpha + 1 \tag{8}$$

$$\beta = i_n.\beta + 1 \tag{9}$$

Substituting in the equation4 following mathematical expression can be derived.

$$f(p \mid \alpha, \beta) = \frac{\Gamma(in.\alpha + in.\beta)}{\Gamma(i_n.\alpha)\Gamma(i_n.\beta)} p^{(1-i}n.\alpha - 1)(1-p)^{(i}n.\beta - 1) \tag{10}$$

$$E(p) = \frac{i_n.\alpha + 1}{i_n.\alpha + i_n.\beta + 2} \tag{11}$$

The probability expectation value from the above equation is interpreted as the relative frequency of the outcome $i_n.\alpha$ is uncertain in future and presented by the value calculated by the equation 11. The reputation value is symbolised by the expectation value. Every user in the network is ranked using the reputation value. The input provided by the other members is used to determine the reputation value. The suggested method divides interactions into positive and negative categories before computing the probability distribution of two events, i.e., the two categories in this method. As a result, if a Sybil attack is initiated using false identification, not all of the network's legitimate users will be complimentary of them; in other words, they may experience isolation. This is true whether the feedback is in the form of opinion, text or both.

By incorporating a time factor into the reputation value, the white wash attack can be thwarted. Older feedback is therefore given less weight than more recent feedback values. A threshold can aid in identifying the presence of harmful feedback in the instance of a slandering attack where the user is providing falsely bad feedback. Since a genuine user will stop interacting after unfavourable interactions. The next sections discusses attack-resistant reputation algorithms.

4.1. White Wash Defense Algorithm

The reputation value incorporates a temporal element to combat the whitewash attack, devaluing older feedback in favour of more recent feedback. The dishonesty detector based on historical data[27] also discusses a related notion. The algorithm suggests that the model will only include recent interaction if the value of the variable *t* is zero. Else other interactions in the history are also counted. These can be restricted by mentioning the numbers 2,3,etc showing number of interaction from the history included in the calculations. Algorithm 1lines 1 to 6 address defense against this attack, where *N* is a list containing the number of interactions.

4.2. Slandering Defense Algorithm

In this attack [28] a malicious entity is identified by the frequency of false negative feedback over a period of time. A genuine user will not have more interactions, after the negative encounters. The attack resistant reputation algorithm 1 assumes a node 's' as a slander node. The interactions are firstly categorized as positive*P* and negative*N*. Then number of negative interactions are checked against a threshold. If it falls beyond the threshold, the node is filtered. In the assumption node 's' is filtered. Algorithm 1 lines 8 -17 categorizes the interaction from the list *N* as positive or negative. If the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2024 doi:10.20944/preprints202401.0633.v1

negative interactions are beyond a certain threshold that node is added to the list of slandering nodes depicted by s.

Algorithm 1 Defense Algorithm

1: Get time t node i interacting node j
2: $t \leftarrow value$
3: **if** $t = 0$ **then**
4: Last interaction between node i,j
5: **else**
6: $N \leftarrow interactions$
7: **end if**
8: **while** $N \neq 0$ **do**
9: **if** $interaction == negative$ **then**
10: $Ni++$
11: $Nj++$
12: **else**
13: $Pi++$
14: $Pj++$
15: **end if**
16: **if** $Nj > threshold$ **then**
17: $s \leftarrow j$ where s is list of slandering nodes
18: **else**
19: $Ev(j) \leftarrow p + 1/p + n + 2$
20: $Ev(i) \leftarrow p + 1/p + n + 2$
21: $Ev(i, j) \leftarrow Ev(i) * Ev(j)$
22: **end if**
23: **end while**
24: $ExpectedValue \rightarrow Ev(i)$

4.3. Sybil Defense Algorithm

The reputation value is utilized to rank the entities. The algorithm determines the user's reputation value by comparing it to the reputation value of the user with whom they are interacting, i.e. j when they provide feedback, ratings, or reviews. The reputation of node i is calculated based on the weights of the nodes it had interactions. The proposed mechanism to address the Sybil attack is shown in lines 18 to 22 in algorithm 1,indicates the expected value of a node that is established by weighing it in accordance with an interaction node's expected value i.e. j.

The proposed defense algorithm counters these three attacks on every interaction of a node.

5. Experiment and Results

5.1. Experiment Setup

The experiments are carried out on three varied scenarios . The scenarios are related to the dataset that lists the interactions of 40 nodes. The interactions are rated on the scale of 1-5, such that 1 is the lowest and 5 is considered as the highest rating. These rating are the interaction feedback values after utilizing the service. This dataset is developed as part of study carried out with the students of UET

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2024 doi:10.20944/preprints202401.0633.v1

Peshawar.The study was unbiased carried out by a third party. The performance evaluation of the defense mechanism of the content credibility system is carried out by authors utilizing Rate of change in rank, Precision, Recall,FMeasure as the performance metrics[29, 30].

5.2. Attack Model

5.2.1. Scenario 1

A rank list is generated from the dataset with the basic reputation rank algorithm. In order to prevent the Sybil attack a rank list is generated according to the Sybil algorithm as discussed above. The top three nodes are node 17,2, 32. In order to verify if Sybil can be prevented the dataset is manipulated and new nodes are introduced with the intention to rise the reputation rank of a particular node i.e. node 13. The results however, show that node 13 does not attain any significant advantage, thus the attack is effectively prevented. In this particular case the newly introduced nodes had no interaction with other nodes so their rank is zero. However, in another case the reputation rank of these newly introduced nodes is fabricated so as to find the implications of the attack. In such case the rank of node 13 got improved from the previous case. Launching the same attack on other content credibility systems i.e pagerank the rank of target nodes gets incremented since more connections are now developed . To authors knowledge this issue is not addressed in any of the versions of the pagerank. Similarly NDR based system could not prevent it since more feedback are now added thereby increasing the overall of average of the reputation value of the target node. Another baseline [13] that is based upon summation reputation calculation, whereby the number of followers, likes are treated as input to calculate the reputation rank easily suffer this attack. The recent randomized IF algorithm has similar results to the proposed technique, however due to its iterative nature the algorithmic complexity is high that increases with increase in number of users and ratings. They also have the assumption that randomly selected users, lead to filtering of unfair rater. The assumption however does not always hold true. The Table 1 shows the ranking before and after attack. After clear observation it was found that node13 gained reputation and its rank improved to 4th place from 7th place in the ranking.The authors calculated the Precision that measures the number of nodes that gained new improved rank in Sybil attack. Thus in this scenario Precision is 0.94 , while Recall is 0.90

5.2.2. Scenario 2

Scenario2 demonstrates the Slandering attack. The attackers intend to decrease the reputation rank of the target node. Thus in this scenario newly introduced nodes give poor feedback so as to malign the target node. The analysis of the results as shown in Figure 4 reveal that ranking after slander attack is quite similar to the ranking after sybil attack. However, few nodes lost their rank and went down in the rank list.The target node 13 didn't lose its rank, the reason being the nodes that are referred by node 13 also face change in rank, thereby in overall calculation node 13 still retains its position, however its reputation value drops considerably. The Precision calculation of this scenario yields value of 0.86 and recall is also 0.8 approx.

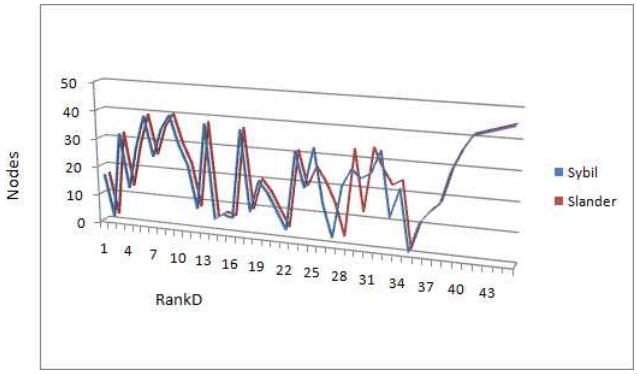

Figure 4. Sybil , Slander RankD comparison.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2024 doi:10.20944/preprints202401.0633.v1

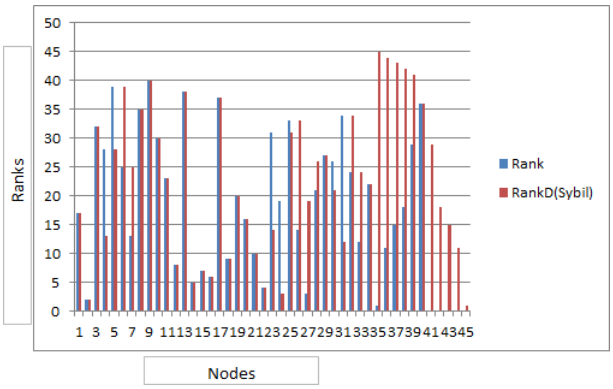

Figure 5. Comparison of ranking before the Sybil attack and after Sybil attack.

5.2.3. Scenario 3

The Scenario 3 is targeted towards effect of the Whitewash attack. The algorithm proposes that with t=0 all interactions in history are used and for t=1..n a specific number of history interactions are utilized in calculating final reputation rank. In order to evaluate this, a simulation is created where change in the reputation rank of node is recorded according to the history of interactions, it can clearly be analysed from the results that the rank is changing with change in history of interactions. This behavior is compared against a competing reputation structure based upon Hidden Markov Model(HMM)[31]. The Table 2 and the Figure 7 shows comparative result of change in reputation values in case of HMM based model and the proposed technique by varying numbers of interactions.The basic thing to compare between the two techniques is the in terms of the response to change in behavior by the two systems. Thus, the authors tested a particular situation involving a node that had a history of 10 interactions with other nodes. We obtained expected value of 0.4 when time factor t=0 was considered, and expected value of 0.3 was obtained when time factor t=1 was considered. The time factor t=0 means usage of all interactions of the history while t=1 means utilization of only recent interactions. The change in HMM based model is steeper meaning it can easily capture the rapidly changing behaviors. However the scenario of whitewash attack does not assume such rapid changes as node starts behaving either as slander or sybil after gaining some reputation in the system, thus employing HMM based model to capture this would be an expensive choice. Also HMM is limited by the duration of states and if combined with learning technique it offers a complex solution. Comparatively beta probability based reputation system equipped with the proposed mechanism to fight against whitewash attack is simpler in terms of time complexity.

Table 1. Ranking without and RankingD with defense.

Rank	Nodes	RankD(Sybil)	Nodes	Nodes	RankD(Slander)
1	17	1	17	1	17
2	2	2	2	2	2
3	32	3	32	3	32
4	28	4	13	4	13
5	39	5	28	5	28
6	25	6	39	6	39
7	13	7	25	7	25
8	35	8	35	8	35
9	40	9	40	9	40
10	30	10	30	10	30
11	23	11	23	11	23
12	8	12	8	12	8
13	38	13	38	13	38
14	5	14	5	14	5
15	7	15	7	15	7

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2024 doi:10.20944/preprints202401.0633.v1

16	6	16	6	16	6
17	37	17	37	17	37
18	9	18	9	18	9
19	20	19	20	19	20
20	16	20	16	20	16
21	10	21	10	21	10
22	4	22	4	22	4
23	31	23	14	23	31
24	19	24	3	24	19
25	33	25	31	25	26
26	14	26	33	26	21
27	3	27	19	27	14
28	21	28	26	28	3
29	27	29	27	29	33
30	26	30	21	30	12
31	34	31	12	31	34
32	24	32	34	32	27
33	12	33	24	33	22
34	22	34	22	34	24
35	1	35	45	35	1
36	11	36	44	36	11
37	15	37	43	37	15
38	18	38	42	38	18
39	29	39	41	39	29
40	36	40	36	40	36
41	0	41	29	41	41
42	0	42	18	42	42
43	0	43	15	43	43
44	0	44	11	44	44
45	0	45	1	45	45

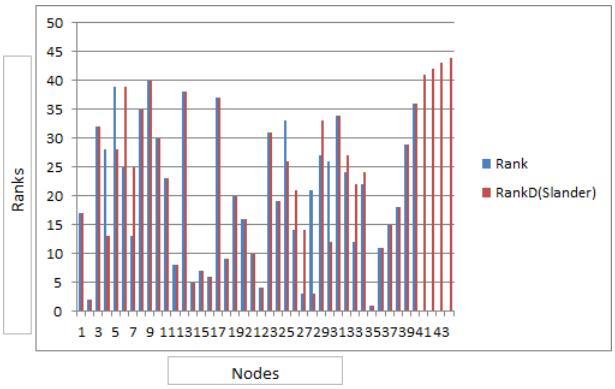

Figure 6. Comparison of ranking before the Slander attack and after Slander attack.

Table 2. Reputation Value (observation probabilities) With Varying History.

Interaction History	Reputation Value	Expected Value (HMM)
All History	0.40	0.45
Latest	0.3	0.20
Latest 3	0.40	0.40
Latest 5	0.42	0.50
Latest 7	0.50	0.60

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2024 doi:10.20944/preprints202401.0633.v1

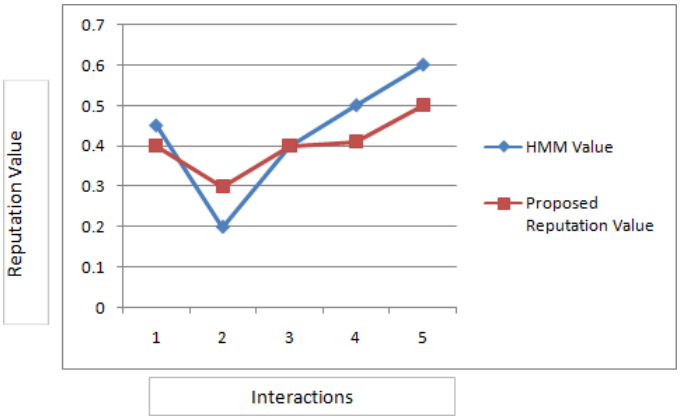

Figure 7. Reputation Rank in case of Whitewash Attack.

5.3. Analysis

Three different scenarios are discussed in this article. The results reveal that performance of defense scheme in case of Sybil attack is better compared to the case of Slandering attack thus, showing that the nodes gaining ranks are monitored in a better manner. The Figures 5 and 6 report the comparison of ranks of nodes before the attack and after the attacks, while the defense algorithms are in practice. It is to be noted that the defense mechanism is preventive in nature. The precision values verify this in both Scenario1 and

Table 3. Time Complexity Comparison- Of Reputation Calculation Structures.

IF Algorithm	HMM	PageRank	Proposed
$O(n^2)$	$O(K^2N)$	$O(n^2)$	O(n)

Table 4. Related Work Efficiency Evaluation.

Schemes	Evaluation Metrics	Results
SybilGuard	Probability of honest node acceptance	87%
SybilLimit	Number of sybil nodes accepted per attack edge	O(log n)
M. Qurishi et.al	Accuracy	86%
Bhupender Kumar et.al	Cost benefit, Utility	Optimum Value

Scenario2. Scenario3 evaluates defense against white wash attack, giving us an insight that it can be a good candidate for domains that do not experience rapid dynamism and a likeable solution for the domains where nodes/agents do not change their behavior rapidly, instead sustain it. The results also reveal that datasets that are dense in nature can have behavior similar to above whereby decreasing or increasing rank of a node can have ripple effect on many others. While in case of sparse datasets where the connectivity is loose, prevention against these attacks can have promising outcomes. In scenario 1 although the attacking nodes had no reputation rank but they still succeeded to raise rank of target node due to density of connections. Thus we can predict that attacker could gain advantage if the attacking nodes have already gained reputation. The comparative Figure 8 shows the Precision, Recall and FMeasure in case of Scenario1 and Scenario2. The FMeasure clearly shows that the proposed mechanism is good enough in prevention against these attacks. Algorithmic comparison to the IF algorithm reveals a major difference in terms of time complexity. The proposed algorithms in this research has complexity of O(n) while the complexity of IF [16] based algorithm is $O(n^2)$. The time

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 January 2024 doi:10.20944/preprints202401.0633.v1

complexity comparison of different reputation models is shown in the Table3. The IF algorithm iteratively operates until the mean average error of a user no more changes. The weights of raters are adjusted according to their distance from the average ratings. The proposed algorithm has mean average error to negligible value. Thus the distinctive feature for comparison is the time complexity of the algorithm. We have also carried out a comparative analysis of previous related work whose target was to detect and prevent Sybil attacks in case of social networks. These schemes utilized different evaluation metrics and reported the results accordingly as shown in the Table 4. However they have proposed schemes in case of social networks . This article has proposed such attacks on reputation based systems that are widely employed in different domains to judge the credibility and quality of information. Also we have evaluated different kinds of attacks that are carried out by fake or Sybil nodes. The previous work as shown in the Table 4 proposes identification of the sybils in the system, we have however assumed that manipulation of Sybils that are not identified can be prevented by securing the calculation structure of the system. We propose that the mathematical model of reputation systems that are used to calculate, ratings, rankings are equipped so as the malicious activity is filtered out. The existing web content credibility systems utilizing reputation [13] do not have defense module to prevent such attacks, that can effect the credibility evaluation of the system. Thus comparative analysis to these cannot be reported. In the context of countering threats that propagate misinformation, this study extends from prior research [32-45] and focuses on the development of a robust defense mechanism

6. Conclusions

Content credibility systems can help users identify credible and correct information. Users utilize ratings, and reputation scores to judge the credibility of content. However such ratings, ranks can suffer these attacks thereby presenting wrong information. This article is focused on the information in social network blogs and related domains. The reputation based credibility systems suffer reputation attacks that can promote spread of wrong information. These attacks can bring down the credible information thereby inducing fake and incorrect information. Sybil, Slandering and Whitewash attacks are explored in this context. One approach is to identify the Sybils and prevent the honest nodes to interact with them. This study however, has proposed another approach whereby malicious act of Sybils could be prevented by the mathematical calculation structure of the reputation systems. The authors have proposed a defense module of the content credibility system that is based upon Bayesian reputation model. The defense module serves to prevent the attacks. The efficacy of the module is discussed in the context of different scenarios. The experimental results revealed a precision of 0.88 when attack is imposed on the system. Similarly the White wash attacks are compared with another solution based upon HMM algorithm also shows choice of the proposed technique appropriate in terms of cost in the given problem space. As a future enhancement the authors aim to simulate further attacks, in different scenarios considering different features of social network users.

Table 5. Comparative Analysis in Terms of Precision , Recall and FMeasure.

Metrics	Scenario1	Scenario2
Precision	0.94	0.836
Recall	0.9	0.8
F Measure	0.919	0.817

 doi:10.20944/preprints202401.0633.v1

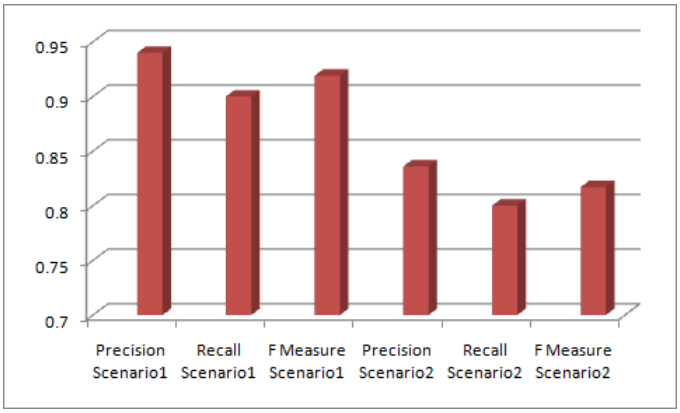

Figure 8. Comparative Analysis in Terms of Precision, Recall and FMeasure.

Appendix A. Example of appendix

Authors that need to include an appendix should place it after the References section. Multiple ap- pendices are allowed and they should be labeled in the order in which they appear in the text. Each of the appendices shall have its heading that follows the style detailed in Section 2.2. Appendices shall be labeled as Appendix A, Appendix B, Appendix C, etc.

Appendix B. Another appendix

References

1. Berghel, H. (2017). Lies, damn lies, and fake news. *Computer*, *50*(2), 80–85.
2. Fairbanks, J., Fitch, N., Knauf, N., Briscoe, E. (2018). Credibility assessment in the news: Do we need to read//*Proc. of the MIS2 Workshop held in conjuction with 11th Intâl Conf. on Web Search and Data Mining*.
3. Benlahbib, A., et al. (2020). Amazonrep: A reputation system to support amazonâs customers purchase decision making process based on mining product reviews//*2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*. IEEE.
4. Liu, X., Nielek, R., Adamska, P., Wierzbicki, A., Aberer, K. (2015). Towards a highly effective and robust web credibility evaluation system. *Decision Support Systems*, *79*, 99–108.
5. Mahmood, S., Ghani, A., Daud, A., Shamshirband, S. (2019). Reputation-based approach toward web content credibility analysis. *IEEE Access*, *7*, 139957–139969.
6. Alqwadri, A., Azzeh, M., Almasalha, F. (2021). Application of machine learning for online reputation systems. *International Journal of Automation and Computing*, *18*(3), 492–502.
7. Alshammari, S. T., Alsubhi, K., Aljahdali, H. M. A., Alghamdi, A. M. (2021). Trust management systems in cloud services environment: Taxonomy of reputation attacks and defense mechanisms. *IEEE Access*, *9*, 161488–161506.
8. Alshammari, S. T., Albeshri, A., Alsubhi, K. (2021). Building a trust model system to avoid cloud services reputation attacks. *Egyptian Informatics Journal*.
9. Ahmed, A. I. A., Ab Hamid, S. H., Gani, A., Khan, M. K., et al. (2019). Trust and reputation for internet of things: Fundamentals, taxonomy, and open research challenges. *Journal of Network and Computer Applications*, *145*, 102409.
10. Nedunchezhian, P., Mahalingam, M. (2022). Sybilsort algorithm-a friend request decision tracking recommender system in online social networks. *Applied Intelligence*, *52*(4), 3995–4014.
11. Chen, Z., Zhao, H. V. (2021). Optimal attacking strategy against online reputation systems with con- sideration of the message-based persuasion phenomenon//*ICASSP 2021-2021 IEEE International Con- ference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.
12. Jethava, G., Rao, U. P. (2022). User behavior-based and graph-based hybrid approach for detection of sybil attack in online social networks. *Computers and Electrical Engineering*, *99*, 107753.
13. Alrubaian, M., Al-Qurishi, M., Al-Rakhami, M., Hassan, M. M., Alamri, A. (2017). Reputation-based credibility analysis of twitter social network users. *Concurrency and Computation: Practice and Expe- rience*, *29*(7), e3873.
14. Lodigiani, C., Melchiori, M. (2016). A pagerank-based reputation model for vgi data. *Procedia Com- puter Science*, *98*, 566–571.

15. Abdel-Hafez, A., Xu, Y., Jøsang, A. (2014). A normal-distribution based reputation model//*International Conference on Trust, Privacy and Security in Digital Business*. Springer.
16. Rezvani, M., Rezvani, M. (2020). A randomized reputation system in the presence of unfair ratings. *ACM Transactions on Management Information Systems (TMIS)*, *11*(1), 1–16.
17. Hoffman, K., Zage, D., Nita-Rotaru, C. (2009). A survey of attack and defense techniques for reputation systems. *ACM Computing Surveys (CSUR)*, *42*(1), 1.
18. Al-Qurishi, M., Alrubaian, M., Rahman, S. M. M., Alamri, A., Hassan, M. M. (2018). A prediction system of sybil attack in social network using deep-regression model. *Future Generation Computer Systems*, *87*, 743–753.
19. Wang, B., Jia, J., Zhang, L., Gong, N. Z. (2018). Structure-based sybil detection in social networks via local rule-based propagation. *IEEE Transactions on Network Science and Engineering*, *6*(3), 523–537.
20. Kumar, B., Bhuyan, B. (2020). Game theoretical defense mechanism against reputation based sybil attacks. *Procedia Computer Science*, *167*, 2465–2477.
21. Potey, A. B., Raut, A. B. (2013). Combating sybil attacks using sybilguard. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *2*(2).
22. Tran, D. N., Min, B., Li, J., Subramanian, L. (2009). Sybil-resilient online content voting.//*NSDI*. volume 9.
23. Yu, H., Gibbons, P. B., Kaminsky, M., Xiao, F. (2008). Sybillimit: A near-optimal social network defense against sybil attacks//*2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE.
24. Rashed, K. A., Balasoiu, C., Klamma, R. (2012). Robust expert ranking in online communities-fighting sybil attacks//*8th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*. IEEE.
25. Miller, N., Resnick, P., Zeckhauser, R. (2005). Eliciting informative feedback: The peer-prediction method. *Management Science*, *51*(9), 1359–1373.
26. Dasgupta, A., Ghosh, A. (2013). Crowdsourced judgement elicitation with endogenous proficiency//*Proceedings of the 22nd international conference on World Wide Web*. ACM.
27. Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., Zhang, Z. (2013). Expertrank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*, *54*(3), 1442–1451.
28. Agate, V., De Paola, A., Re, G. L., Morana, M. (2018). A platform for the evaluation of distributed reputation algorithms//*2018 IEEE/ACM 22nd International Symposium on Distributed Simulation and Real Time Applications (DS-RT)*. IEEE.
29. Fouss, F., Achbany, Y., Saerens, M. (2010). A probabilistic reputation model based on transaction ratings. *Information Sciences*, *180*(11), 2095–2123.
30. Liu, Y., Chitawa, U. S., Guo, G., Wang, X., Tan, Z., et al. (2017). A reputation model for aggregating ratings based on beta distribution function//*Proceedings of the 2nd International Conference on Crowd Science and Engineering*. ACM.
31. Kokkodis, M. (2019). Reputation deflation through dynamic expertise assessment in online labor markets//*The World Wide Web Conference*.
32. Lim, M., Abdullah, A., & Jhanjhi, N. Z. (2020). Data fusion-link prediction for evolutionary network with deep reinforcement learning. International Journal of Advanced Computer Science and Applications, 11(6).
33. A. Almusaylim, Z., Jhanjhi, N. Z., & Alhumam, A. (2020). Detection and mitigation of RPL rank and version number attacks in the internet of things: SRPL-RP. Sensors, 20(21), 5997.
34. Lim, M., Abdullah, A., & Jhanjhi, N. Z. (2021). Performance optimization of criminal network hidden link prediction model with deep reinforcement learning. Journal of King Saud University-Computer and Information Sciences, 33(10), 1202-1210.
35. Kok, S. H., Azween, A., & Jhanjhi, N. Z. (2020). Evaluation metric for crypto-ransomware detection using machine learning. Journal of Information Security and Applications, 55, 102646.
36. Shafiq, M., Ashraf, H., Ullah, A., Masud, M., Azeem, M., Jhanjhi, N. Z., & Humayun, M. (2021). Robust Cluster-Based Routing Protocol for IoT-Assisted Smart Devices in WSN. Computers, Materials & Continua, 67(3).
37. Hussain, K., Hussain, S. J., Jhanjhi, N. Z., & Humayun, M. (2019, April). SYN flood attack detection based on bayes estimator (SFADBE) for MANET. In 2019 International Conference on Computer and Information Sciences (ICCIS) (pp. 1-4). IEEE.
38. Lim, M., Abdullah, A., Jhanjhi, N. Z., & Supramaniam, M. (2019). Hidden link prediction in criminal networks using the deep reinforcement learning technique. Computers, 8(1), 8.
39. Kumar, T., Pandey, B., Mussavi, S. H. A., & Zaman, N. (2015). CTHS based energy efficient thermal aware image ALU design on FPGA. Wireless Personal Communications, 85, 671-696.
40. Verma, S., Kaur, S., Rawat, D. B., Xi, C., Alex, L. T., & Jhanjhi, N. Z. (2021). Intelligent framework using IoT-based WSNs for wildfire detection. IEEE Access, 9, 48185-48196.

41. Khalil, M. I., Jhanjhi, N. Z., Humayun, M., Sivanesan, S., Masud, M., & Hossain, M. S. (2021). Hybrid smart grid with sustainable energy efficient resources for smart cities. sustainable energy technologies and assessments, 46, 101211.
42. Diwaker, C., Tomar, P., Solanki, A., Nayyar, A., Jhanjhi, N. Z., Abdullah, A., & Supramaniam, M. (2019). A new model for predicting component-based software reliability using soft computing. IEEE Access, 7, 147191-147203.
43. Hussain, S. J., Ahmed, U., Liaquat, H., Mir, S., Jhanjhi, N. Z., & Humayun, M. (2019, April). IMIAD: intelligent malware identification for android platform. In 2019 International Conference on Computer and Information Sciences (ICCIS) (pp. 1-6). IEEE.
44. Sennan, S., Somula, R., Luhach, A. K., Deverajan, G. G., Alnumay, W., Jhanjhi, N. Z., ... & Sharma, P. (2021). Energy efficient optimal parent selection based routing protocol for Internet of Things using firefly optimization algorithm. Transactions on Emerging Telecommunications Technologies, 32(8), e4171.
45. Wassan, S., Chen, X., Shen, T., Waqar, M., & Jhanjhi, N. Z. (2021). Amazon product sentiment analysis using machine learning techniques. Revista Argentina de Clínica Psicológica, 30(1), 695.