2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W) | 978-1-6654-0262-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/DSN-W54100.2022.00037

# A Comprehensive Dynamic Quality Assessment Method for Cyber Threat Intelligence

Menghan Wang
*School of Cyber Science and Technology*
*Northwestern Polytechnical University*
Xi'an, China
wmh@mail.nwpu.edu.cn

Libin Yang
*School of Cyber Science and Technology*
*Northwestern Polytechnical University*
Xi'an, China
libiny@nwpu.edu.cn

Wei Lou
*Department of Computing*
*The Hong Kong Polytechnic University*
Hong Kong
csweilou@comp.polyu.edu.hk

***Abstract*—Extraordinary growth of the Internet poses a great challenge for defending worldwide evolution of cyber attacks. Introducing cyber threat intelligence (CTI) is a promising approach for alleviating malicious attacks, which heavily relies on the quality of CTI themselves. However, most of current studies develop CTI quality assessment from the perspective of source or content separately, regardless of their availability in practical. In this paper, a dynamic method named CTIC to comprehensively assess CTI quality is proposed. Specifically, we propose a novel CTI feed assessing scheme by modeling the interactions of feeds as a correlation graph. An iterative algorithm is elaborated to depict the feed quality precisely. We design a CTI content assessing scheme together with a machine learning algorithm to score the availability of content from multi-dimensions. Experimental results on real data confirm our proposed mechanism can quantitatively as well as effectively assess CTI quality.**

***Index Terms*—Cyber threat intelligence, quality assessment, graph, machine learning, dynamic**

## I. Introduction

Now the Internet is facing an unprecedented increase together with a worldwide evolution of cyber attacks. Especially with the relatively frequent emergence of Advanced Persistent Threats (APTs), traditional security defenses are unable to keep up with the increasing sophistication of attack tools and methodologies. Cyber threat intelligence (CTI) is a promising approach for alleviating malicious attacks, by providing additional information to depict a full picture of the fast-evolving cyber threat situation. CTI is evidence-based knowledge, including context, mechanisms, indicators, implications, and actionable advice [1]. Recently, many institutes and companies have been reported to provide various threat intelligence services, including IBM X-Force, AlienVault, etc.

The success of the CTI heavily relies on its effectiveness, including timeliness, completeness, accuracy and so on. To assess CTI quality before utilizing it, some recent works gave a structured definition of intelligence [2], [3] before assessing it. Quite amount of works have paid more attention to study content assessment with machine learning in term of time, location and so on [2], [4], [5]. Some works utilized graph mining techniques to assess the quality of intelligence sources [6]. Remarkably, most of the previous literature develops CTI quality assessment ontology from the perspective of CTI source or content separately, and none of these works have addressed their assessment methods based on real dataset.

To better assess the quality of CTI, we propose CTIC, a comprehensive CTI quality assessing mechanism with time dynamic, which fully synthesizes various criteria tightly associating with CTI quality, e.g., trustworthy of CTI source, availability of CTI content. Specifically, we make the following key contributions with our work:

- We propose a novel CTI feed assessing scheme by modeling the interactions of feeds as a correlation graph. An iterative algorithm is elaborated to depict the CTI feed quality precisely.
- We design a CTI content assessing scheme based on a machine learning algorithm to score the availability of individual CTI content from multi-dimensions, i.e., verifiability, richness and timeliness.
- We evaluate our mechanism based on real-world IoCs. The results demonstrate it can quantitatively as well as effectively evaluate the quality of CTI.

The rest of the paper is organized as follows. In Section II, we describe our methodology for CTI assessing. Section III demonstrates experimental results on real datasets. Section IV reviews related work, and Section V concludes this paper.

## II. Methodology

This section describes our comprehensive dynamic assessment method named CTIC, including three steps of feed assessment, content assessment and comprehensive assessment.

### A. Intelligence Feed Assessment

A typical intelligence structure should contain information such as source, time, category, tag, geo, value, and so on, expressed in the form of (Source, Time, Category, Tag, Geo, Value), where Source describes where the intelligence comes from. Time is the timestamp. Category can be classified as spam, scanner, malware, botnet, phishing, etc. Tag describes the threat description label, such as IDC, attacks email, attacks mail, etc. The geographical location is given in Geo. Value is the IP or Domain. Here we give an example of structured intelligence: (*StopForumSpam*, 2020-11-25T18:57:10Z, spam, IDC, Seoul Korea, 164.52.42.6).

2325-6664/22/$31.00 ©2022 IEEE
DOI 10.1109/DSN-W54100.2022.00037

We consider a weighed directed graph (digraph) $G = (V, E)$ to model the interactions of CTI feeds, where $V$ is the set of feeds, $E$ is the set of directed edges between feeds. There are $N$ nodes/feeds, i.e., $|V| = N$. An edge $< i, j >$ indicates node $i$ has cited $j$. Each edge in $E$ is associated with a weight $w_{i,j}$, which represents the number of citations. The graph is asymmetrical, i.e., it is possible that $w_{i,j} \neq w_{j,i}$.

Based on structured CTI, a threat intelligence feed relationship graph can be constructed according to the above description. Next, we propose the Citation-based feed assessment algorithm in Algorithm 1, which is inspired by the HITS algorithm, but introduces the weight of each edge additionally, in order to consider the times of citation. We have the intuition that feeds with high originality should be considered as higher-quality feeds.

Every feed $v$ has two attributes, $hub$ and $au$, used to describe the citations, where $hub$ denotes link authority, and $au$ denotes content authority. Citing intelligence from other feeds will increase $hub$, and being cited helps increase $au$. For feed $s$, its $au$ and $hub$ enhance each other.

**Algorithm 1** Citation-based feed assessment algorithm

1: **Input:** CTI feed relationship graph $G$;
2: Initialize the content authority vector $A$ and link authority vector $H$ to 1;
3: In the $k^{th}$ iteration, $au$ of feed $s$ is calculated by $A(s) = \sum_{i=1}^{n} H(i) w_{i,s}$;
4: After the new vector $A$ is obtained in step 2, $hub$ of the feed $s$ is calculated by $H(s) = \sum_{i=1}^{n} A(i) w_{i,s}$;
5: Normalize the vector A and vector H, and then iterate until convergence;
6: Calculate the feed quality score $S_{feed} = au \times 100$;
7: **Output:** The $au$, $hub$ and $S_{feed}$ of each feed;

In the step 3 of Algorithm 1, $n$ is the number of feeds pointing to $s$, $i$ is one of them, $H(i)$ represents the $hub$ of feed $i$, and $w_{i,s}$ denotes the weight of edge $< i, s >$. The $au$ of the feed is equal to the sum of the product of the $hub$ of all feeds pointing to it and the weight of the corresponding edge. In the step 4, $n$ is the number of feeds pointed to by $n$, $i$ is one of the feeds pointed to by $s$, $A(i)$ denotes the $au$ of feed $i$, and $w_{i,s}$ denotes the weight of edge $< s, i >$. The $hub$ of the feed is equal to the sum of the product of the $au$ of all feeds it points to and the weight of corresponding edge. The quality of an intelligence feed is determined by its $hub$ and $au$ jointly. Finally, an intelligence feed with a high $au$ is considered to be of high originality, and quality score $S_{feed}$ of each feed are obtained.

### B. Intelligence Content Assessment

The intelligence content describes the attacker's attack time, location, and method [7]. We construct indicators from three dimensions of multi-source verification, content richness, and timeliness to directly measure CTI quality.

*a) Multi-source verification:* We consider content consistency of multiple threat sources and similarity between CTI is used for comparison. The greater the similarity, the higher the consistency, and the higher its quality [2].

For CTI $v_t$, it has multi-source verification CTI set $V = \{v_1, v_2, ..., v_i, ..., v_m\}$. The size of set $v$ is defined as $N_{verify}$, representing the number of intelligence that can be used for verification. The similarity between $v_t$ and $v_i$ is calculated using four characteristics of source, timestamp, threat category and description tag, as shown below.

$$\begin{aligned} s(v_t, v_i) &= \theta_1 \times s_{source} + \theta_2 \times s_{category} \\ &+ \theta_3 \times s_{time} + (1 - \theta_1 - \theta_2 - \theta_3) \times s_{tag} \end{aligned} \tag{1}$$

$$s_{time}(v_t, v_i) = 1 - \frac{|t(v_t, v_i)| - min(|t(v_t, v_i)|)}{max(|t(v_t, v_i)|) - min(|t(v_t, v_i)|)} \tag{2}$$

$$s_{tag}(v_t, v_i) = \frac{X_t \cdot X_i}{|X_t| \times |X_i|} \tag{3}$$

In Equation (1), the weights are set to 0.25 in our work. $s_{source}$ and $s_{category}$ are determined by comparing whether the two pieces of CTI have the same source or category. We compute the normalized time delta to estimate the similarity $s_{time}$ in Equation (2). Besides, in Equation (3), the similarity between two description tags is compared by using cosine similarity. We use $N_{sup}$ to denote the collections of positive verified intelligence. With the help of $\kappa$, we get $N_{sup}$. If $s(v_t, v_i) \geq \kappa$, $v_i$ expresses support for $v_t$, and opposition otherwise. The optimal value of $\kappa$ is evaluated by experiments. Then we get support ratio as $R_{sup} = \frac{N_{sup}}{N_{verify}}$.

*b) Content richness:* CTI with high quality can provide rich contextual information. Therefore, an indicator named tag content richness is defined, and the value is the number of items in the tag.

*c) Timeliness:* Attackers will constantly replace some characteristic information to hide their tracks. Thus, we use the time delta between the timestamp and a certain fixed moment to measure the timeliness of each piece of CTI.

After preprocessing, the extracted features are inputs of a KNN classifier, where CTI samples are divided into five categories according to content quality and every category has a corresponding score. This way, each piece of intelligence has its content quality level and content quality score $S_{content}$.

### C. Intelligence Comprehensive Assessment

The key aspects of this model need to be that it takes into account each factor according to its importance for a speacial case, and the quality score can be updated if new evidence arrives, to account for the dynamically evolving cybersecurity environment.

For each CTI, its comprehensive quality score can be assessed by adding a weighted sum of its basic scores, i.e., its feed-based score and content-based score, to comprehensive score of the previous CTI, which is from the same feed [8] as this one. The ageing factor [9] multiplied to historial score is used to put less emphasis on past events and highlight more

recent events. This is to account for the fact that, while the past should not be dismissed completely, CTI quality should be judged more on what it is able to deliver in the present. Equation (4) details the comprehensive assessment, where $S(n)$ denotes the score of the $n^{th}$ intelligence sample from feed $S$. $D$ is the ageing factor. $s_i(n)$ represents the $i^{th}$ basic quality score of the $n^{th}$ sample, specifically, $s_1(1)$ is the first sample's feed-based score , $s_2(1)$ is its content-based score, and $w_i$ is the weight of each factor $s_i(n)$.

$$S(n) = \frac{D \times S(n-1) + \sum_{i=1}^{2} w_i \cdot s_i(n)}{D+1} \tag{4}$$

Note that it is a special case of comprehensive assessing when $D = 0$, which is exactly a static assessment combining intelligence feed and content. We first assume $D = 0.5$, and then determine the actual value based on the performance of the model under different values. Here we apply the coefficient of variation to choose $w_i$, which is a normalized measure of the degree of dispersion of a probability distribution in probability theory and statistics.

Through the above assessing mechanism, we score the quality of each sample. And they can be divided into five levels. Each 20% of the score value is a level.

## III. Experiments

### A. Experiment Setup

A series of evaluations are implemented in Python. We picked valid 21,448 CTI samples from the original data set, which contains millions of samples within a window of four months.

### B. Feed Assessment Results

We first construct a CTI feed relationship graph in Figure 1(a). As we can observe, there are 14 nodes and 39 edges, whose size and width are diverse. It implicates that 14 feeds have 39 kinds of citations. The size of a node is determined by in-degree and out-degree, i.e., the more the node points to others or is pointed to, the greater the degrees, the bigger the node. The width of an edge is determined by the weight, i.e., the more citations, the greater the weight, the wider the edge. Especially, the biggest size of node *Emergingthreats* means its degree is the highest and it has the most citations. The widest edge from *StopForumSpam* to *Emergingthreats* indicates that the citations between them are the most.

Then, $au$ of the 14 CTI feeds can be obtained. We note that the $au$ of source *Emergingthreats* is the highest, whose content authority is relatively high. It reveals that the intelligence generated by this source is more frequently cited by other CTI feeds. This finding is consistent with our previous analysis of the relationship graph. Such a high-authority CTI feed is considered a high-quality intelligence feed. According to the statistics, the average of $au$ is 0.13, the variance is 0.05, and the dispersion coefficient is 1.77, indicating that the data has a large degree of dispersion and a large range of fluctuations. This confirms that there are indeed differences in the quality of CTI feeds. The $au$ of each feed is multiplied by 100 for feed-based quality score $S_{feed}$.

### C. Content Assessment Results

We quantify the feature indicators to obtain feature vectors and then preprocess the data as the input of the classifier. The classification algorithm we use is KNN and the label is the level to which the sample's reputation score belongs, as detailed in Section II. Classification results demonstrate that the samples are divided into five categories significantly, which implies that the intelligence samples are divided into five levels based on their content quality. This finding helps to prove that the intelligence samples do have a difference in content quality level and our content assessment method is effective.

The multiclass classification performance evaluation criteria we adopt are precision, recall rate, F1-score and accuracy, which are commonly used in the field of machine learning. Experimental results show that the performance criteria are all over 0.9.

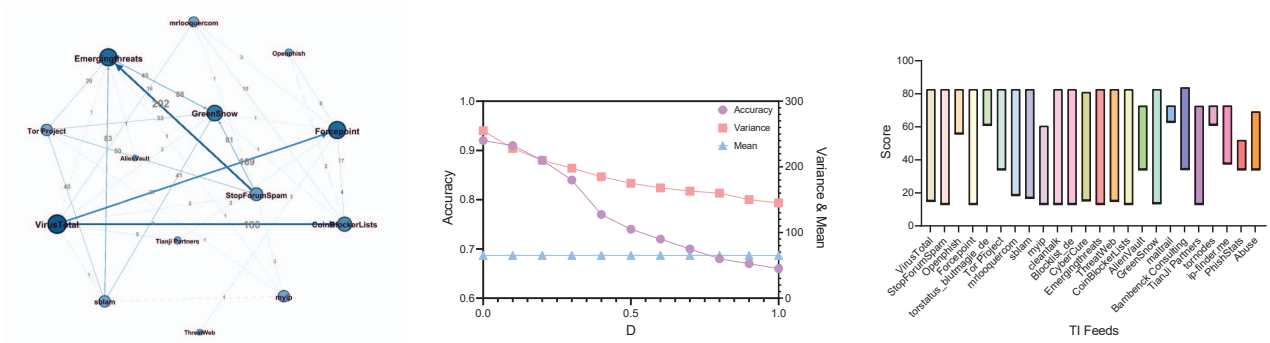
After classification, each sample has its content quality level and score $S_{content}$. For the comparison, we take the "reputation" marked in the original dataset as a benchmark and get the result that the matching accuracy is 0.923. It indicates that the content assessment we present is of effectiveness.

### D. Comprehensive Assessment Results

We examine the comprehensive Assessment mechanism performance under different parameters. To start with, we evaluate the static quality assessment without ageing factor. Experiments show that when the weights are set as $w_1 = 0.001, w_2 = 0.999$, the matching accuracy of the static score and marked score is the highest, with the value 0.925. This ratio decreases with the increase in the proportion of feed quality factors. Following this, we introduce the ageing factor $D$. Using the coefficient of variation method, the weights of the two factors are $w_1 = 0.83, w_2 = 0.17$. Given $w_1$ and $w_2$, we vary $D$ from 0 to 1. When $D = 0.1$, the matching accuracy is 0.908. The larger the $D$, the lower the matching accuracy, as shown in Fig. 1(b). It can be considered that when the original platform assessing intelligence, it did not consider the influence of feeds and time dynamics, but only assessed the current content.

We then test the impact of $D$ on variance and mean of comprehensive quality, as shown in Fig. 1(b). It can be seen that as $D$ increases, the variance decreases, which indicates the the fluctuation of comprehensive quality is smaller and the adaptability of our mechanism to the dynamic changes of the network is worse. Different values of $D$ have no obvious impact on the mean. But it still shows that when $D$ is 0.6, the mean is the highest, therefore, we set $D = 0.6$ in our work.

The distribution range of comprehensive quality of 24 feeds are shown in Fig. 1(c). We can see there are gaps in the quality of CTI from different feeds, which proves the importance of comparing them. Meanwhile, the CTI quality of some feeds such as *VirusTotal* fluctuates significantly, which reveals that their CTI quality is unstable and it's essential to carefully screen. In contrast, sources like *Openphish* have relatively stable and high CTI quality. It indicates that their




(a) Threat intelligence feed relationship graph. (b) The influence of parameter D. (c) Score distribution.

Fig. 1: Evaluation Results.

CTI quality is usually stable and higher. Hence, these sources can be trusted.

## IV. Related Work

Based on qualitative methods [10], [11], more studies have proposed quantitative methods. Some works extract multiple characteristics as a reference for assessment, for instance, Li et al. [2] extracted 16 features from three dimensions of time, content and domain knowledge, then use an evaluation algorithm based on DBN. In [4], the authors extracted features from four dimensions of intelligence source, intelligence content, active period, and blacklist database matching degree, and designed an assessing model based on DNN algorithm and Softmax classifiers.

Some works present custom indicators, including volume, differential contribution, exclusive contribution, latency, accuracy and coverage [12], extensiveness, maintenance, false positives, verifiability, intelligence, interoperability, compliance, similarity, timeliness and completeness [9], sensitivity, originality and impact [13], etc.

Graph mining techniques have been introduced, which can express relationships between entities intuitively. As for example, [3] innovatively constructed a heterogeneous threat intelligence graph. Inspired by PageRank, FeedRank was proposed as a TI feed ranking method in [6]. The model in [5] constructed an intelligence knowledge graph, and comprehensively uses TransE and RNN models to assess the credibility of intelligence data.

Compared with previous work, we have advantages on dynamic consideration, meanwhile combining source with content and combining machine learning with graph mining.

## V. Conclusion

In this paper, we have proposed a CTI quality assessment method named CTIC, which including feed assessment, content assessment and time dynamics. Experiments based on real-world datasets show that our mechanism can effectively evaluate CTI quality and achieve stellar performance. Our insight sheds some light on CTI quality assessment, and provides not only academic research ideas but also practical implementation experiences, that will help security users share more credible CTI.

## References


[1] M. R and P. K, "Market guide for security threat intelligence services," *Gartner report (G00259127)*, 2014.

[2] L. Li, "Study on the multi-dimensional analysis model of threat intelligence credibility in cyberspace," Master's thesis, Beijing University of Posts and Telecommunications, 2018.

[3] Y. Gao, X. Li, J. Li, Y. Gao, and N. Guo, "Graph mining-based trust evaluation mechanism with multidimensional features for large-scale heterogeneous threat intelligence," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 1272–1277.

[4] H. Liu, H. Tang, M. Bo, J. Niu, T. Li, and L. Li, "A multi-source threat intelligence confidence value evaluation method based on machine learning," *Telecommunications Science*, vol. 36, no. 1, pp. 119–126, 2020.

[5] X. Cheng, "Study on trustworthy analysis on threat intelligence based on machine learning," Master's thesis, Beijing University of Posts and Telecommunications, 2019.

[6] R. Meier, C. Scherrer, D. Gugelmann, V. Lenders, and L. Vanbever, "Feedrank: A tamper-resistant method for the ranking of cyber threat intelligence feeds," in *2018 10th International Conference on Cyber Conflict (CyCon)*. IEEE, 2018, pp. 321–344.

[7] B. Fang, "Define cyberspace security," *Chinese Journal of Network and Information Security*, vol. 4, no. 1, pp. 1–5, 2018.

[8] Y. Zhu, L. Huang, G. Chen, and W. Yang, "Dynamic trust evaluation model under distributed computing environment," Ph.D. dissertation, 2011.

[9] T. Schaberreiter, V. Kupfersberger, K. Rantos, A. Spyros, A. Papanikolaou, C. Ilioudis, and G. Quirchmayr, "A quantitative evaluation of trust in the quality of cyber threat intelligence sources," in *Proceedings of the 14th International Conference on Availability, Reliability and Security*, 2019, pp. 1–10.

[10] X. Bouwman, H. Griffioen, J. Egbers, C. Doerr, B. Klievink, and M. van Eeten, "A different cup of {TI}? the added value of commercial threat intelligence," in *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020, pp. 433–450.

[11] A. de Melo e Silva, J. J. Costa Gondim, R. de Oliveira Albuquerque, and L. J. García Villalba, "A methodology to evaluate standards and platforms within cyber threat intelligence," *Future Internet*, vol. 12, no. 6, p. 108, 2020.

[12] V. G. Li, M. Dunn, P. Pearce, D. McCoy, G. M. Voelker, and S. Savage, "Reading the tea leaves: A comparative analysis of threat intelligence," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, pp. 851–867.

[13] H. Griffioen, T. Booij, and C. Doerr, "Quality evaluation of cyber threat intelligence feeds," in *International Conference on Applied Cryptography and Network Security*. Springer, 2020, pp. 277–296.