MeshTrust: A CDN-centric Trust Model for Reputation Management on Video Traffic

Xiang Tian[1,2,3], Yujia Zhu[2,3], Zhao Li[1,2,3],
Chao Zheng[2,3], Qingyun Liu[2,3], and Yong Sun[2,3]

[1] School of Cyber Security, University of Chinese Academy of Sciences
[2] National Engineering Laboratory for Information Security Technologies
[3] Institute of Information Engineering, Chinese Academy of Sciences
{tianxiang, zhuyujia, lizhao, zhengchao, liuqingyun, sunyong}@iie.ac.cn

Abstract. Video applications today are more often deploying content delivery networks (CDNs) for content delivery. However, by decoupling the owner of the content and the organization serving it, CDNs could be abused by attackers to commit network crimes. Traditional flow-level measurements for generating reputation of IPs and domain names for video applications are insufficient. In this paper, we present *MeshTrust*, a novel approach that assessing reputation of service providers on video traffic automatically. We tackle the challenge from two aspects: the multi-tenancy structure representation and CDN-centric trust model. First, by mining behavioral and semantic characteristics, a *Mesh Graph* consisting of video websites, CDN nodes and their relations is constructed. Second, we introduce a novel CDN-centric trust model which transforms *Mesh Graph* into *Trust Graph* based on extended network embedding methods. Based on the labeled nodes in *Trust Graph*, a reputation score can be easily calculated and applied to real-time reputation management on video traffic. Our experiments show that *MeshTrust* can differentiate normal and illegal video websites with accuracy approximately 95% in a real cloud environment.

Keywords: CDN · reputation management · network embedding · DNS · trust model · video traffic analysis · .

1 Introduction

In the past few years, the Internet has witnessed an explosion of video streaming applications. The explosive growth of video websites and traffic is largely due to the popularity of content delivery networks (CDNs).

Besides being used for obvious benign purposes, video websites are also popular for malicious or illegal use [4]. e.g., websites are increasingly playing a role for the management of disseminating illegal content. Going beyond being popular among adolescents, video content has now evolved into a much-debated public concern because of excessive or maladaptive use. For decades, the public have been consistently concerned about the potentially harmful influences of

exposure to pornographic and violent content, being targeted for harassment, cyber-bullying, sexual solicitation, and Internet addiction [12]. Therefore, it is of great significance to label the reputation for video services.

The traditional mechanisms for generating trust and protecting content security by finding the IPs and domain names attached to the website, using the historical data in traffic log cannot be used for tangled networks [2]. This is because the deployment of a horizontally scalable website tending to use public infrastructure, especially video websites distribute video to CDNs. The spreading of video content by CDN service have been brought new challenges to website reputation evaluation problems: (*i*) dynamic changes of IPs and domains, and (*ii*) the relationship between website and domains is no longer a one-to-one relationship. But a multi-tenant, multi-CDN graph structure is formed by CDN as a common software service [8].

However, the problem still exits and we are trying to identify the website reputation through the reputation of rented CDN nodes. Our intuition is that the CDN nodes which hosting similar content or providing similar services are likely to be in a homophilic state. In this work, we propose *CCTrust*, a model for evaluating reputation of CDN nodes, and *MeshTrust*, a mechanism that is able to build a CDN-centric model for reputation management. First, we need measurement video websites and CDNs from distributed vantage points in order to characterize lease relationship between them. Second, realizing the reputation of CDN nodes based on network embedding. Third, *MeshTrust* can assign appropriate reputation score of video websites.

This paper provides the following contribution:

1) We consider the real situation of multi-CDN. Based on this, an automatic detection mechanism, *MeshTrust* is proposed, which solving the problem of reputation evaluation in multi-tenant scenario. To the best of our knowledge, we are the first to detect illegal or malicious websites in cloud scenes using CDN trust model.

2) We design a novel model, *CCTrust*, constructing *Mesh Graph* from network video traffic and transforming the heterogeneous graph into homogeneous *Trust Graph*.

3) We extend graph embedding algorithm, transforming the sparse graph from high dimension to low dimension to form a labeled *Trust Graph*. On this basis, a reputation score can be easily calculated and applied to real-time reputation management

4) We verify the effectiveness of our mechanism in a real environment. The results show that *MeshTrust* can evaluate websites reputation with accuracy 95%. Meanwhile, our mechanism is lightweight and can be applied to real-time detection.

The rest of the paper is organized as follows. Sect.2 presents problem and preliminary. Sect.3 presents our *CCTrust* Model. In Sect.4, we provide details of our mechanism. Experiments on real world datasets are shown in Sect.5. We comment on related works in Sect.6. Sect.7 gives a conclusion.

2 Problem and Preliminary

In this paper, we aim at assessing reputation of video websites by identifying CDN nodes behind specific websites, rather than directly classifying the website's IPs and domains. Before presenting our framework, we introduce multi-CDN. And our problem statement is given.

2.1 Multi-CDN

To further ensure the availability of content, customers may use multi-CDN deployments. Other reasons to utilize more than one CDN provider may be cost efficiency, e.g., different prices to serve content at different times or due to traffic volume contracts [8]. Video websites deploy multi-CDN, which mixed multiple CDNs to satisfy various requirements. In an attempt to minimize single points of failure and eke out even more performance advantages many video websites have started to deploy multi-CDN for higher availability, better performance, increased capacity, and better security. Maybe one CDN is suitable for video content while another is suited to hosting image or picture content. By tapping into both, the content owner is getting a more global friendly solution.

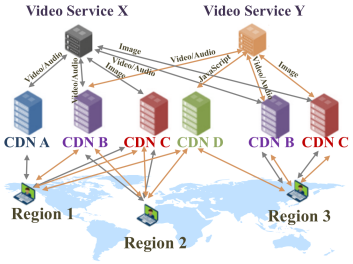


Fig. 1. Multi-CDN

In Fig. 1, CDN provider A, B, C, D host content for video service X and Y. For better performance and security in disparate geographic region, CDN providers host different types of resources. Multi-CDN results in a complicated subscription relationships for video websites. Instead of helping users to choose CDN providers, we utilize the complex relationship between websites and CDNs to build a graph. This is also an innovation of our work.

2.2 Problem Statement

MeshTrust is a novel mechanism for reputation management on video traffic. We focus on multi-tenancy structure in our trust model. First *MeshTrust* builds a *Mesh Graph* consisting of video websites, CDN nodes and their relations by mining behavioral and semantic characteristics from large-scale passive video traffic.

Second, *MeshTrust* assesses reputation of video websites by a novel CDN-centric trust model (*CCTrust*). *CCTrust* transforms *Mesh Graph* into *Trust Graph* and extends network embedding methods to learn a mapping function. *CCTrust* outputs the reputation evaluating results of CDN nodes, which is an input of online assessment for websites reputation.

In this paper, we analyze the CDN nodes from the perspective of domains and IPs. We refer to the first sub-domain after the Top-Level Domain(TLD) as Second Level Domain(2LD). It generally refers to the organization that owns the domain name. Fully Qualified Domain Name(FQDN) is the domain name complete with all the labels that unambiguously identify a resource [3].

A Graph is $G = (V, E)$, where $v \in V$ is a node and $e \in E$ is an edge. G is associated with a node type mapping function $f_v : V \rightarrow \tau^v$ and an edge type mapping function $f_e : E \rightarrow \tau^e$. τ^v and τ^e denote the set of node types and edge types, respectively [5].

Definition 1. ***Mesh Graph***. *Mesh Graph is an interaction information graph, which is defined as* $G_m = (W, C, R, \lambda)$, *where* $W = \{w_1, ..., w_n\}$ *represents* n *websites nodes,* $C = \{c_1, ..., c_m\}$ *represents* m *CDN nodes and* $R = r_{ij}$ *represents edges from* W *to* C, *edges means lease relationship. If there exists an edge from* w_i *to* c_j, $r_{ij} = 1$. *Otherwise,* $r_{ij} = 0$. λ *is the representation of CDN nodes. In our work,* $\lambda = 2LD, FQDN, IP$.

Definition 2. ***Trust Graph***. *Trust Graph* $G_t = (C, E, W, \lambda, L)$ *is an undirected graph where* C *is a node set which is the same in Mesh Graph, the node represents the CDN nodes.* E *is defined as an edge set, the edge indicates that there are leased relation between CDN nodes with the same video websites.* W *is the edge weight corresponding to the node similarity.* λ *is the CDN nodes representation.* L *is a label set of all CDN nodes. The purpose of Trust Graph with network embedding is finding* L.

Definition 3. ***Reputation Score***. *The reputation score enable dynamic domain name blacklists to counter illegal or malicious website much more effectively [1]. Given a specific video website* v, $Score_v$ *is reputation scores for this video website.* $Score_c$ *is reputation scores for a given CDN node* c.

Our main goal is to construct a trust graph for reputation management on video traffic, which can be applied to effectively differentiate normal and illegal applications and activities. Thus the goal of the evaluation reputation problem can be expressed as how to determine L in *Trust Graph* $G_t = (C, E, W, \lambda, L)$.

3 The CCTrust Model

We consider identification method of the CDN nodes reputation, based on video website content hosted on these nodes. Combining the historical data in traffic, *CCTrust* model aims to solving the credibility evaluation of CDN nodes. Starting from the *Mesh Graph* of video websites, our model: (i) transforms *Mesh Graph*

to *Trust Graph*, (*ii*) achieves classification of CDN nodes through applying the network embedding algorithm which could extract the structure feature of a graph, and (*iii*) evaluates the reputation of CDN nodes.

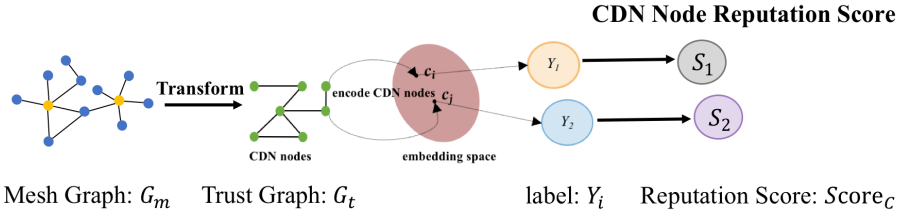


Fig. 2. CCTrust Model Overview

3.1 Graph Transformation

Mesh Graph $G_m = (W, C, R, \lambda)$ is a heterogeneous graph, which $|\tau^v| = 2$ and $|\tau^e| = 1$. All nodes belong to two types, while all edges belong to a single type. Given an *Mesh Graph*, *CCTrust* transform the Graph to *Trust Graph* $G_t = (C, E, W, \lambda, L)$. *Trust Graph* is a homogeneous graph, which $|\tau^v| = |\tau^e| = 1$. All nodes and edges in G_t belong to one single type.

3.2 Network Embedding and Classification

Considering the sparsity of the *Trust Graph*, *CCTrust* choose network embedding to finding L. **Network embedding** is an important method for learning the low-dimensional representation of vertices in a network. It transforms network information into low-dimensional dense real vectors and is used for input of existing machine learning algorithms to capture and retain the network structure. GCN(Graph Convolutional Networks) pertain to deep learning based graph embedding without random walk paths [5]. The first GCN introduced for learning representations at the node level was in [10], where they utilized GCNs for the semi-supervised node classification problem.

Reputation of nodes based network embedding learns a mapping function $f : C \to \mathbb{R}^d$, where $d \ll |W|$. The mapping function should preserve the graph structure information. After that, all CDN nodes can be classified in a low-dimensional latent space. The input data of the network embedding is *Trust Graph* $G_t = (C, E, W, \lambda, L)$. The output result is the node feature matrix with $N \times F$ dimensions, F is the dimension to feature vector. Our intuition is that once two CDN nodes serve the same website, the two nodes are similar. Based on the fact, we using **Jaccard Distance** to determine the node similarity. In addition, considering the sparsity of the graph, we apply add-one smoothing

method to smooth similarity. The similarity between two CDN nodes c_i and c_j can be calculated as:

$$sim_{ij} = \begin{cases} \frac{|V_i \cap V_j|}{|V_i \cup V_j|+n} & |V_i \cap V_j| = \varnothing \\ \frac{|V_i \cap V_j|+(n-1)}{|V_i \cup V_j|+n} & |V_i \cap V_j| = |V_i \cup V_j| \\ \frac{|V_i \cap V_j|}{|V_i \cup V_j|} & otherwise \end{cases} \tag{1}$$

CCTrust classifies CDN nodes, determining what is the reputation score of the nodes. The node feature matrix and node label matrix as the input data of GCN. We produce the node label matrix using the label of video website hosted by CDN node. Intuitively, relations between the two nodes directly connected is equivalent to the adjacency matrix of the original network modeling. However, the relationships tend to be very sparse in the network. It is necessary to further depict the global and local similarity to consider nodes which is not directly connected, whereas have respectable common neighbor nodes or same structure. Thus GCN is used for evaluating the CDN nodes reputation.

A two-layer GCN is adapted to CDN node classification on *Trust Graph* with a symmetric adjacency matrix A_C (weighted). The model applies a softmax classifier on the output features:

$$Z = softmax(\hat{A_C} ReLU(\hat{A_C} X {W_C}^{(0)}) {W_C}^{(1)}) \tag{2}$$

where $\hat{A_C} = {\tilde{D_C}}^{-\frac{1}{2}} \tilde{A_C} {\tilde{D_C}}^{-\frac{1}{2}}$. The loss function is defined as the cross-entropy error over all labeled examples:

$$\zeta = \sum_{c \in y_C} \sum_{l=1}^{|L|} Y_{cl} ln Z_{cl} \tag{3}$$

where y_C is the set of node indices that have labels.

3.3 CDN Nodes Reputation Evaluation

In order to evaluate the reputation of CDN nodes based on classified results. Given a CDN node c_i, the probability of each label represented by p_{il}. According to the predict type, *CCTrust* cacultes the reputation of a specific CDN node c_i. The sum total of the probability of each category label is the reputation of the CDN node.

$$Score_{c_i} = \sum_{l=1}^{|L|} p_{il} \tag{4}$$

4 The MeshTrust Mechanism

In this section we describe how *MeshTrust* works. Fig. 3 shows a high-level overview of this mechanism. *MeshTrust* is composed of off-line and on-line procedures: (i) **off-line**: identifying CDN nodes and websites, measurement relationship between them, constructing *Mesh Graph*, and evaluating reputation of

CDN nodes by *CCTrust* model. (*ii*) **on-line**: identifying CDN nodes of a specific website, and evaluating the website reputation combined the CDN node reputation results.

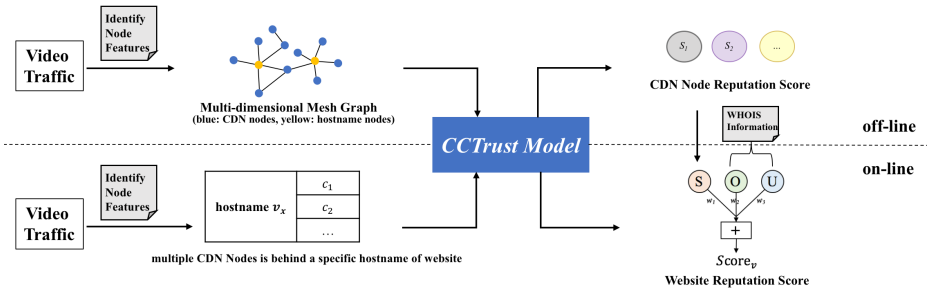


Fig. 3. MeshTrust Mechanism Overview

4.1 Mining Nodes from Passive Video Traffic

Research about CDN measurement mostly starting from identifying CDN. The main method of identifying CDN nodes: (*i*) **based on CNAME**: Satellite exploits the reverse PTR records and the WHOIS organization to identify CDN nodes[15]. Singh performs a reverse DNS lookup on the server IP address, develop regular expressions to identify specific CDNs based on the returned hostname [16]. (*ii*) **based on HTTP header**: McDonald identifies CDN population by HTTP hearder [14]. Based on these studies, we use CNAME and HTTP headers to identify CDN nodes.

While there are a large number of CDN domains in passive traffic, we could mine more CDN subdomains corresponding to providers. Monitoring the traffic passively, then parsing the DNS and HTTP traffic. The candidate CDN domains are obtained based on the multi-features identification whether the video resource is hosted on CDN node. Multi-features refer to behavioral and semantic characteristics: (*i*) **behavioral features**: the same content resource corresponds to multiple server-side IPs, (*ii*) **semantic features**: CNAME has CDN keywords or contains the name of the CDN providers.

For the candidate CDN domains, it is necessary to verify the validity of the domain and provider pair through the WHOIS information. The pair determines whether the domain belongs to the CDN provider. For domains that have blank WHOIS information, we can utilize the search engine searching for information. According to match whether the provider name appears in the search result to verify the validity. Finally, we obtain a CDN identification feature library which including CDN provider domain name set, CDN domain general feature words(e,g., cachecdn), and HTTP header(e.g., a special header: "cdn cache server" arises in Server field means it is ChinaNetCenter CDN nodes).

We monitor DNS and HTTP video traffic passively. Eventually, we analyze the relationship between the video website and the CDN nodes.

4.2 Mesh Graph Construction

To better utilize content-multihoming, multi-CDN enable content providers to realize custom and dynamic routing policies to direct traffic to the different CDNs hosting their content [8]. The relationship of hosting is intricate. We construct the *Mesh Graph* of video website and CDN nodes, based on multi-domain integration. In a word, our mechanism builds the *Mesh Graph* automatically by merging CDN nodes from the dimensions of providers(such as Akamai), 2LD(such as akamai.net), FQDN(such as cdn.cdn-baidu.net), and IP.

4.3 CCTrust Model Construction

The main tasks of *CCTrust* model in reputation evaluation mechanism: transform *Mesh Graph* to *Trust Graph*, identify and evaluate the reputation score of CDN nodes. *CCTrust* transforms the heterogeneous *Mesh Graph* to homogeneous *Trust Graph* which only consist of CDN nodes firstly. The reputation result $Score_c$ based on classifying the CDN nodes in the new low-dimensional space to gain the probability that each node belongs to each category label(e.g., benign, porn, malicious).

4.4 Websites Reputation Evaluation

When a specific website occurs in video passive traffic, *MeshTrust* identifying the CDN nodes provide service for the website. In addition, we consider two other factors: WHOIS organization, the update frequency of WHOIS information, which influence the degree of trust on. WHOIS organization is represented by O. The update frequency is represented by U. Our key insight in these reputation factors is that as benign domain names: (i) updates WHOIS information more often, while most illegal or malicous domain names almost change WHOIS data rarely, even never change, (ii) always has a clear and specific organization information, while illegal domains dont have. The reputation score $Score_v$ of the video website can be calculated as follows, m is the number of domains belonging to the website v, n is the number of CDN nodes belonging to the domain j:

$$Score_v = w_1 \sum_{j=1}^{m} \sum_{i=1}^{n} c_j Score_{c_{ij}} + w_2 \sum_{j=1}^{m} o_j + w_3 \sum_{j=1}^{m} u_j \tag{5}$$

After determining the reputation of these CDN nodes. WHOIS organization and update frequency are supplementary, together calculating the reputation score $Score_v$.

5 Experimental Results

In order to evaluate the availability of *MeshTrust* for identifying illegal video websites, we verify the evaluation effectiveness on the reputation of CDN nodes and websites. *MeshTrust* (*i*) surveys and maps the dependency relation of video websites and CDNs, (*ii*) building *CCTrust* model, and (*iii*) evaluating the reputation of the video websites.

5.1 Input Data

We adopt the observation configuration of passive traffic, and verify the effect of our mechanism according to the known website category. We judge whether it is video traffic by file suffix. We sniffed 24-hours long traffic in February 28, 2018 and June 23, 2018 within a large ISP. We focuse on 2,000 video websites, which contain the Baidu rank[1] of China video websites TOP 700, Alexa rank[2] of other countries video websites TOP 500(we remove the websites from the list of publicly available illegal websites), and 800 porn video websites which are produced from public blacklist. We extract the traffic which is relevant to these specific websites by matching the domain names. We labeled the websites derived from Baidu rank and Alexa rank are benign, these sites continue to appear in the ranking after a long period of observation.

5.2 Characterizing CDNs of Video Websites

The CDN service lease relationship is measured from the dimension of geographical location. Besides passive video traffic, we sniff traffic with measurement points covering six countries: China, American, Australia, Japan, South Korea and Singapore. At each measurement point, our mechanism automatically simulated user behaviour to visit video URL. The measurement of CDN service shows that there are 388 mainstream benign video websites rent CDNs. Meanwhile, the number of porn websites which exist lease relationship with CDNs is 66. We focus on analyzing the measurement results of video websites which rent CDNs in China.

Measurement the Tenancy of Chinese Website. In the measurement result of Chinese website, the top 100 video websites all use CDN, and 33% of video websites rely on more than one CDN provider. About 10% of video websites rely on three or more CDN, e.g., iQIYI mainly use self-built CDN in China, while using Akamai CDN in American and Australia primarily. The video resource of Youku is hosted by AliCloud, but the picture resources are mostly hosted by Akamai and ChinaNetCenter. These further indicates that video websites tend to rely on the parallel mode of multi-CDN providers for distributing service.

[1] http://top.chinaz.com/hangye
[2] https://www.alexa.com/topsites

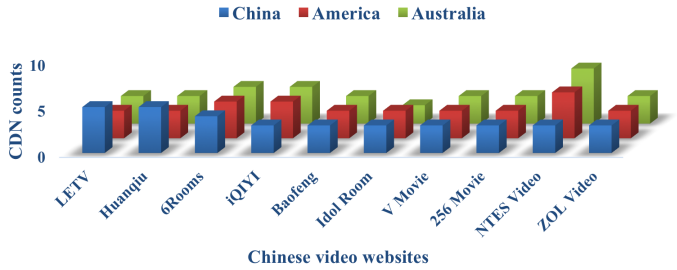
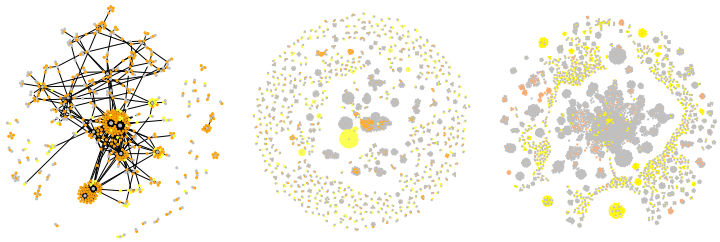


Fig. 4. Popularity of Multi-CDN

5.3 Evaluating Website Reputation based on CCTrust Model

To evaluate the accuracy of identifying result and prove the effectiveness of *MeshTrust*. We perform *CCTrust* to model the CDN nodes in *Mesh Graph*. Consequently, we evaluate the reputation of video websites. *Mesh Graph* has been constructing completely in the previous step of the mechanism.

The input data of *CCTrust* model is *Mesh Graph*, which contains CDN and website nodes. Firstly, we produce the input data of evaluateing CDN nodes reputation by transforming *Mesh Graph* to *Trust Graph*.



(a) $\lambda = 2LD$ (b) $\lambda = FQDN$ (c) $\lambda = IP$

Fig. 5. Mesh Graph of Video Websites: the porn or benign websites with orange, yellow respectively, and CDN nodes with gray

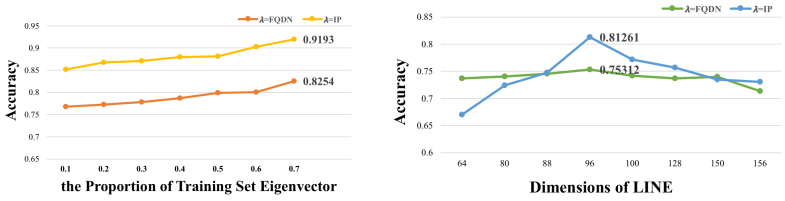
Mesh Graph. We construct the *Mesh Graph* form three dimensions: (i) $\lambda = 2LD$, we construct the video website *Mesh Graph* G_{m_1}, and label the websites porn or benign. There are 248 CDN nodes which belongs to 168 CDN providers. (ii) $\lambda = FQDN$, analogously, we bulid G_{m_2}, which contains 1580 grey FQDN nodes. (iii) $\lambda = IP$, G_{m_3} contains 4408 CDN IP nodes.

In Fig. 5, we can visually observe that some CDN nodes provide services primarily for porn websites, and the platforms of these nodes have poor self-regulation ability. While some provide services only for non-porn websites, and the platforms of these nodes are likely to have strong self-regulation ability. In addition, some nodes were mixed providing service for porn and benign websites. And it is necessary to evaluate the reputation of CDN nodes and conduct hierarchical sampling of video traffic based on the reputation scores.

Table 1. Trust Graph Datasets

Datasets	# Nodes	# Edges	Labels	Max Degree
$G_{t_1}(\lambda = 2LD)$	248	1,316	148;74;26	208
$G_{t_2}(\lambda = FQDN)$	1,604	23,282	1179;411;14	152
$G_{t_3}(\lambda = IP)$	4,408	298,336	3157;1140;111	65

Trust Graph. Three *Mesh Graph* transform to three corresponding *Trust Graph*: $(i)G_{t_1}$: $\lambda = 2LD$, including 148 benign, 74 porn and 26 CDN nodes serving the porn and benign website, a total of 248 CDN nodes, and $porn : benign = 1 : 2$. $(ii)G_{t_2}$: $\lambda = FQDN$, $porn : benign = 1 : 3$. $(iii)G_{t_3}$: $\lambda = IP$, $porn : benign = 1 : 3$.



(a) GCN(F_2) (b) GCN(F_3)

Fig. 6. The influence of training set eigenvector proportion on the accuracy of GCN(F_2) and LINE dimensions on the accuracy of GCN(F_3) identification method

Off-line Evaluation Result. We apply three methods to construct the node feature matrix of GCN :$(i)F_1$: randomly generate $n \times n$ dimensional diagonal matrix, n denotes the number of nodes, $(ii)F_2$: produce website feature matrix, that the feature vector of each node is formed with the presence of all the websites. $(iii)F_3$: obtain the output node feature matrix of network embedding, such as LINE [17]. LINE is a algorithm fall into edge reconstruction based optimization. It means the edges established based on node embedding should be as similar to those in the input graph as possible [5]. As shown in Fig. 6(a), even

when the percentage of training set is 10%, the accuracy of GCN(F_2) method is better than 77%. Therefore, our model can be used to identification CDN nodes on video traffic by training fewer samples. When the dimension of LINE is 96, GCN(F_3) could achieve the higher accuracy in Fig. 6(b). The classification result indicates that the optimum solution obtained by using F_2 matrix. The optimal accuracy could achieve **82.54%**(G_{t_2}) and **91.93%**(G_{t_3}).

Table 2. Evaluating the Website Reputation Results

Datasets	# Website	Benign	Illegal	# Max CDN Nodes	Accuracy
W_1	968	385	583	87	**95.25%**
W_2	1,287	380	907	214	**92.23%**

On-line Evaluation Result. We evaluate our model on the datasets, and the result is illustrated in Tab. 2. We monitor traffic passively to acquire CDN nodes behind specific websites. We collect W_1 and W_2 dataset to validate the reputation on account of the reputation result of G_{t_2} and G_{t_3} respectively. CCTrust performs reputation evaluation on account of W_1 and W_2. The results show that MeshTrust could evaluate the reputation of websites, identify the illegal and benign websites with accuracy **95.25%**(W_1) and **92.23%**(W_2) approximately when $w_1 = 0.7$, $w_2 = 0.15$, $w_3 = 0.15$.

6 Related Work

Several prior studies have investigated the identification of malicious domains and reputation models for websites. These relevant research based on characteristics: content characteristics (e.g., tag information of HTML pages), URL information (e.g., URL vocabulary information, URL length), DNS logs (e.g., WHOIS information, AS numbers).

Based-on Content. Canali used the abnormal content and abnormal structure of the web page as a sign to judge the maliciousness of the domains [6]. Such methods require more computing resources and network bandwidth, and generate a larger time overhead. The time required to analyze a web page also depends on the network latency and the complexity of the web content.

Based-on URL Information. Starting from the URL structure information, Le improved the accuracy of malicious domain recognition [11]. Li based on the network topology relationship with the PageRank algorithm, identifying malicious domains, which makes the false detection rate control within 2% [13]. Due to the widespread practice of HTTPS, the URL information cannot be obtained. Additionally, with the feature scale grows linearly, the characteristic set expanding.

Based-on DNS Log. Antonakakis proposed Notos, which handles DNS query responses from passive DNS traffic and extracts a set of 41 features from the observed FQDN and IP [1]. Chiba proposed Domain-Profiler to actively collect DNS logs, analyze time-change patterns, and predict whether a given domain will be used for malicious purposes [7]. In line with the global association diagram between domain and IP, Khalil proposed a path-based mechanism to derive the malicious score of each domain [9]. The above systems almost utilize the characteristics of the domain to identify the type of domains and further evaluate the reputation of the websites. Due to DNS encryption and cross-use of IPs and domains, DNS traffic is no longer suitable for detection.

Due to the web tangle: (*i*) multiple services and resources co-located on the same CDN (*ii*) more and more content providers employ multiple CDNs to serve the same content to reduce costs or to select CDNs by optimal performance, these methods above are insufficient. As far as we know, we are the first to use the CDN trust graph to solve the problem of reputation assessment for video applications in a multi-tenant scenarios.

7 Conclusion

In this paper, we propose *MeshTrust* based on *CCTrust* Model. *MeshTrust* evaluates video website reputation by *CCTrust* rather than directly using the website IPs or domains. Our mechanism is able to evaluate the websites reputation with accuracy approximately 95%. Then within the evaluation result, the back-end priority processes the associated traffic of less reputable domains. Treating video traffic differently which can greatly improve the processing content censorship performance of video traffic.

The passive traffic exploited in our work is sniffed from only one ISP with a limited range of traffic coverage. Therefore, the measured dependence of video websites and CDNs is not comprehensive, resulting in a smaller coverage of the discovered CDN nodes. We evaluate the reputation of the nodes over static data acquired in a single time point, which can be seen as a snapshot of network. Therefore, the *Mesh Graph* is static. In the future, we would deploy *MeshTrust* in a larger network environment.

8 Acknowledgments

We would like to thank hard work of MESA TEAM (*www.mesalab.cn*). This work was supported by National Key R&D Program 2016 (Grant No. 2016YFB0801300); the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDC02030600). The corresponding author is Yujia Zhu.

References

1. Antonakakis, M., Perdisci, R., Dagon, D., Lee, W., Feamster, N.: Building a dynamic reputation system for dns. In: USENIX security symposium. pp. 273–290 (2010)

2. Berger, A., DAlconzo, A., Gansterer, W.N., Pescapé, A.: Mining agile dns traffic using graph analysis for cybercrime detection. Computer Networks **100**, 28–44 (2016)
3. Bermudez, I.N., Mellia, M., Munafo, M.M., Keralapura, R., Nucci, A.: Dns to the rescue: discerning content and services in a tangled web. In: Proceedings of the 2012 Internet Measurement Conference. pp. 413–426. ACM (2012)
4. Bilge, L., Kirda, E., Kruegel, C., Balduzzi, M.: Exposure: Finding malicious domains using passive dns analysis. In: Ndss (2011)
5. Cai, H., Zheng, V.W., Chang, K.: A comprehensive survey of graph embedding: problems, techniques and applications. IEEE Transactions on Knowledge and Data Engineering (2018)
6. Canali, D., Cova, M., Vigna, G., Kruegel, C.: Prophiler: a fast filter for the large-scale detection of malicious web pages. In: Proceedings of the 20th international conference on World wide web. pp. 197–206. ACM (2011)
7. Chiba, D., Yagi, T., Akiyama, M., Shibahara, T., Yada, T., Mori, T., Goto, S.: Domainprofiler: Discovering domain names abused in future. In: 2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). pp. 491–502. IEEE (2016)
8. Hohlfeld, O., Rüth, J., Wolsing, K., Zimmermann, T.: Characterizing a meta-cdn. In: International Conference on Passive and Active Network Measurement. pp. 114–128. Springer (2018)
9. Khalil, I., Yu, T., Guan, B.: Discovering malicious domains through passive dns data graph analysis. In: Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security. pp. 663–674. ACM (2016)
10. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
11. Le, A., Markopoulou, A., Faloutsos, M.: Phishdef: Url names say it all. In: 2011 Proceedings IEEE INFOCOM. pp. 191–195. IEEE (2011)
12. Leung, L.: Predicting internet risks: a longitudinal panel study of gratifications-sought, internet addiction symptoms, and social media use among children and adolescents. Health Psychology and Behavioral Medicine: an Open Access Journal **2**(1), 424–439 (2014)
13. Li, Z., Alrwais, S., Xie, Y., Yu, F., Wang, X.: Finding the linchpins of the dark web: a study on topologically dedicated hosts on malicious web infrastructures. In: Security and Privacy (SP), 2013 IEEE Symposium on. pp. 112–126. IEEE (2013)
14. McDonald, A., Bernhard, M., Valenta, L., VanderSloot, B., Scott, W., Sullivan, N., Halderman, J.A., Ensafi, R.: 403 forbidden: A global view of cdn geoblocking. In: Proceedings of the Internet Measurement Conference 2018. pp. 218–230. ACM (2018)
15. Scott, W., Anderson, T.E., Kohno, T., Krishnamurthy, A.: Satellite: Joint analysis of cdns and network-level interference. In: USENIX Annual Technical Conference. pp. 195–208 (2016)
16. Singh, R., Dunna, A., Gill, P.: Characterizing the deployment and performance of multi-cdns. In: Proceedings of the Internet Measurement Conference 2018. pp. 168–174. ACM (2018)
17. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web. pp. 1067–1077. International World Wide Web Conferences Steering Committee (2015)