International Journal of Computer Applications Technology and Research
Volume 5– Issue 2, 99 - 103, 2016, ISSN:- 2319–8656

Distributed Digital Artifacts on the Semantic Web

Susan P Kurian
Department of Computer Science and Engineering
Mangalam college of Engineering
Kottayam, India

Vishnu S Sekhar
Department of Computer Science and Engineering
Mangalam college of Engineering
Kottayam, India

Abstract: Distributed digital artifacts incorporate cryptographic hash values to URI called trusty URIs in a distributed environment building good in quality, verifiable and unchangeable web resources to prevent the rising man in the middle attack. The greatest challenge of a centralized system is that it gives users no possibility to check whether data have been modified and the communication is limited to a single server. As a solution for this, is the distributed digital artifact system, where resources are distributed among different domains to enable inter-domain communication. Due to the emerging developments in web, attacks have increased rapidly, among which man in the middle attack (MIMA) is a serious issue, where user security is at its threat. This work tries to prevent MIMA to an extent, by providing self reference and trusty URIs even when presented in a distributed environment. Any manipulation to the data is efficiently identified and any further access to that data is blocked by informing user that the uniform location has been changed. System uses self-reference to contain trusty URI for each resource, lineage algorithm for generating seed and SHA-512 hash generation algorithm to ensure security. It is implemented on the semantic web, which is an extension to the world wide web, using RDF (Resource Description Framework) to identify the resource. Hence the framework was developed to overcome existing challenges by making the digital artifacts on the semantic web distributed to enable communication between different domains across the network securely and thereby preventing MIMA.

Keywords: ***Digital artifacts, man in the middle attack(MIMA), semantic web, RDF, trusty URI***.

1. INTRODUCTION

With the ascend of credit cards, contactless payments & crypto currencies people have been predicting the end for physical money for nearly 60 years. Over the past decades, researchers have confirmed that there is only 9% of physical money with men and the rest is invested via internet, as technology has made work easier, which can be done from anywhere at any time. And here comes the relevance of this system to provide security for data in web, which is one among the greatest challenges currently raised. The solution for this is the distributed digital artifact system, which prevents the relevant man in the middle attack to an extent by ensuring verifiability and reliability.

The system consists of a coordinator process, to manage the domain which is assumed to be trusted. Seed generator is used to connect server in a domain which want to part of the semantic web publication, through which index of reference tree is built in multiple domain. Hash value will be calculated and Base 64 encoding is done. It is then published on the interface once the RDF encoding has been generated. On users request for service at the server, the server in turn connect to other server which has the required resource and the document is delivered to the client if the right access is satisfied followed by Base 64 decoding.

The rapid development of online payments, e-commerce sites, netbanking etc. made human work much easier, where they can do everything sitting at home on a button click. Due to these emerging developments website attacks have increased rapidly, where user security is at its threat. Among websites attacks man in the middle attack (MIMA) is a serious issue, where a malicious actor places himself into a conversation between 2 parties to access the information that they have send to each other. This work tries to prevent MIMA to an extent, by providing self reference and trusty URIs even when presented in a distributed environment. Any manipulation to the data is efficiently identified and any further access to that data is blocked by informing user that the uniform location has been changed.

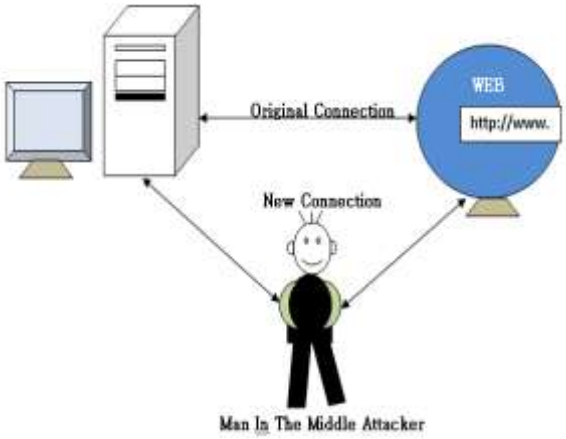

Fig.1 Man In the Middle Attack

2. RELATED WORKS

Lots of research are going on in the task of making digital artifacts on the web verifiable and reliable[1], [3], [5] ,[6]. In [1] authors suggest a module wise approach to make documents on the web correct, unmodified and always made available. The system makes use of trusty URIs[2] including hash values to identify modified input, which returns a totally changed value, even when slightly changed.

Nymble [2] is a system in which servers blacklist misbehaving users and blocks them. Websites use a seed for each nymble for blacklisting users, which in turn links future

nymbles from same user. It's a comprehensive credential system which maintains the privacy of blacklisted users.

Tobias Kuhn and Michel Dumontier in [3], a mechanism to incorporate cryptographic hash values in URIs was proposed . It was used to make the entire reference trees verifiable. The modular architecture used improves reliability and efficiency of tools.

Semantic web security and privacy system [4] deals with policy based security and privacy management on the Semantic Web. It supports protecting sensitive resources and information revelation. It describes policy, their interactions, specification, conflict detection and validation.

In [5] a decentralized approach to circulate, access and storing of data is considered. It propose a web based bottom-up process allowing researchers to publish, retrieve data in a reliable and trustworthy conduct.

Data lineage [7], [8] is used for checking data correctness. It describes data origin, how its extracted and its modification over time.

3. SYSTEM ARCHITECTURE

In centralized digital artifact system, when users request for service it will be fetched from the RDF stored in the central server and delivered. In semantic web which uses self-reference the verification occurs between a single central server and different URNs resulting in just a reference tree as output.

But in distributed digital artifact system, resources are distributed among different domains and each domain can communicate with each other. Here resources will not be stored in central server, rather will be distributed, and requests from users will be passed between different domains for processing. Here cross site verification is possible between different domains resulting in a complete forest as output. In distributed environment RDF is automatically generated which ensures efficiency of the system whereas in the other its externally generated which is a drawback consuming more time.

The system consists of different domains, which will be managed by coordinator process. Seed generator is used to generate a number for unique identification of multiple domains in a distributed environment. The resources will be distributed among multiple domains where they can communicate with each other. Hash value of a particular cited document will be calculated [8] and Base 64 encoding [1] is done. It can then be published on the interface once the RDF encoding has been generated. On users request for service at the server, the server in turn connect to other server which has the required resource and the document is delivered to the client if the right access is satisfied followed by Base 64 decoding.

Trusty URIs [1] will end with a hash value encoded in Bae64 notation, which can be a typical ASCII character (A-Z or a-z), any digit (0-9), a hyphen (-) or an underscore (_). All trusty URI will end with no less than 25 Base64 characters. The *artifact code*, whose first character represent type and second character version representing module identifiers which are followed by *data part*, which holds a hash part.

http://localhost:8080/r1.RA5AbXdpz5DcaYXCh9l3eI9ruBosiL5XDU3rxBbBaUO70

From the above example, localhost:8080/ represents self-references, resources that holds within, their own trusty URI. The whole thing that follows r1. is the artifact code. Its first character R recognize the module indicating its type and second character A specifies the version. The left behind 43 characters represent the real hash value.

The system consists of a server process which manage the various domains. The RDF generator process is responsible for generating metadata for the uploaded data in the session. Centralized and distributed storage of document/data is controlled by a storage process. Hashing make sure that data is integrated and encoding is employed for secure traversal of hash value. Client process verify the integrity of the document on the arrival at client side.

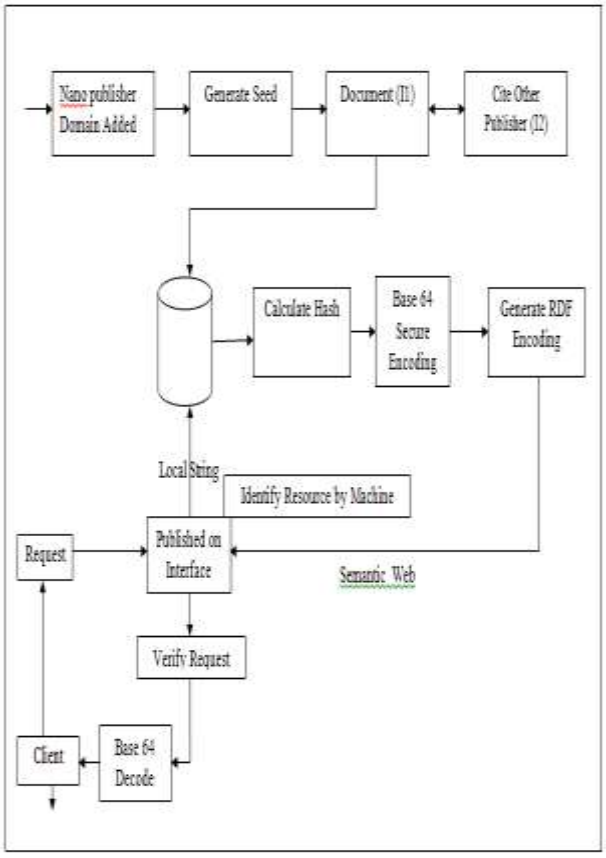

Fig.2 System Architecture

The modules of the proposed system can be broadly classified into the following namely,

a. Seed Generation
b. Distributed Communication
c. File Content Access
d. RDF Access
e. RDF Transferral

f. Client Request Processing

a. Seed Generation

Seed is a sequence of randomly generated number, providing unique id. A lineage algorithm [7] is employed. Whenever a domain is registered with the distributed system,its corresponding storage id is created which will be further used for its unique identification. Each domain will be linked to a seed, using which one domain will be connected to the other. Seed generator is used to connect server in a domain which want to part of the semantic web publication, through which index of reference tree is built in multiple domain.

On each user request, domain verifies seed to identify the site of the requested document to be delivered. Distributed network connects its different domains to each other where users can publish, retrieve and replicate documents distributed through the network.

b. Distributed Communication

Here each domain is free to communicate with each other, since the index of reference tree is built in multiple domain. A document can cite other publisher in a distributed environment. Each domain can post their publication to another domain or even to themselves. The domain to which posted can either approve or reject the document. But it requires no validation if posted to themselves. If approved its RDF [10] is automatically generated, and the document is published on the interface, accessible to all others across the network and if rejected its corresponding entry will be deleted. A domain himself acting as an attacker can sabotage the entrusted document given upon trust. But even presented in a distributed environment enabling inter-domain communication, the system ensures security to the document making digital artifacts on the web verified and trustworthy using *trusty URIs* and prevents *MIMA* attacks.

c. File Content Access

At FA, using SHA-512 hash generation algorithm [8] hash value is calculated, to which after appending two zero-bits are converted to Base64 notation generating trusty URN and complete trusty URI.

d. RDF Access

At RA, supports multiple graphs which works on RDF content. It allows self-references, resources that contain their own trusty URI. For Unicode characters a SHA-512 is generated in UTF-8 encoding, append two zero bits and is finally converted to Base64 notation.

e. RDF Transferral

At RB, trusty URI represents single RDF graph. Similar to RA, hash value is calculated for Unicode using SHA-512 in UTF-8 encoding and is transformed to Base64 notation.

f. Client Request Processing

The user request for service (finding, querying, filtering) at the server. The server in turn connect to other server which has the required resource. The connection requesting server has the hash index to verify that they are also in trusted domain. If the right access satisfied, Base64 decoding employed and the document is delivered to the client.. Any modification deny further access to that URL, returning an error message informing uniform location has been changed.

Integration or verification of trusty uri is made with the help of RDF meta data, which is machine understandable data. Whenever a domain uploaded the data, its corresponding hash value is included in the rdf tag with another metadata like seed, storage location etc. On browser's request for the document, the server respond with trusty uri which contain the hash value in the Base64 encoded form. When the document is loaded the client process calculate the hash value and perform matching function to do accept/reject.

3.1 The Seed Generation Algorithm

A lineage algorithm [7], [9] is used to generate a seed which is a randomly generated number for unique identification. Whenever a domain is registered with the distributed system, its corresponding storage id is created which will be further used for its unique identification. Each domain will be linked to a seed using which one domain will be connected to other. On each user request, domain verifies seed to identify the location of the requested content to be delivered. Every first post in each seed will be treated as *parent seed* which will be followed by *child seeds*. Each user request will processed from parent seed to childrens. The parent seed is searched using bubble sort, with a complexity of *O(n)* whereas childrens use quick sort with *O(nlogn)* complexity. The search is completed with an overall complexity of *O(nlogn)* which improves the performance.

4. EXPERIMENTAL EVALUATION

Experiments are conducted on Intel Core i3 processor with CPU of 2.40GHz. In order to measure several parameters of the system different data sets were experimented.

Distributed digital artifact system shows high performance than other systems in terms of MIMA detection rate and MIMA prevention rate.

A. MIMA Detection Rate

Man in the middle attack is a type of cyber attack where a malicious actor tries to get information that two parties send to each other. Since humans totally dependent on the internet, MIMA attacks have tremendously increased and preventing them is very essential.

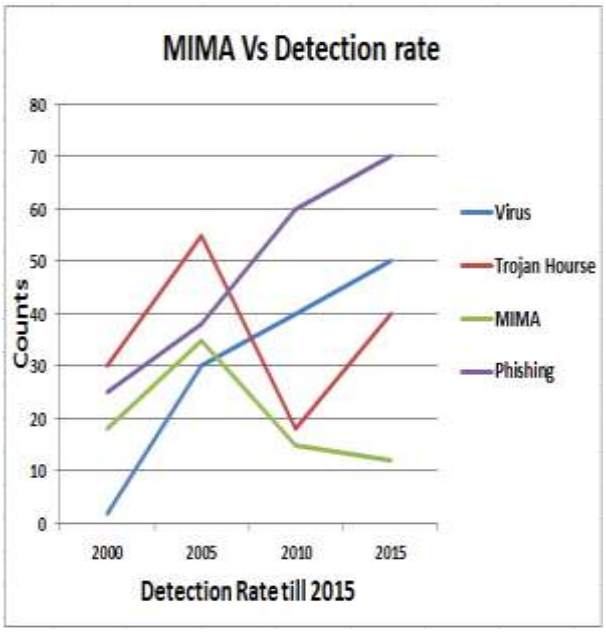

Fig.3 MIMA detection rate

Fig.3 shows that as years pass by the attacks rapidly increase. It illustrates a comparison between different attacks like virus, trojen hourse, phishing and man in the middle attacks and it shows that as years go man in the middle attack(MIMA) is on its hike and detecting MIMA is very difficult i.e, its detection rate has rapidly decreased which shows the relevance of this work.

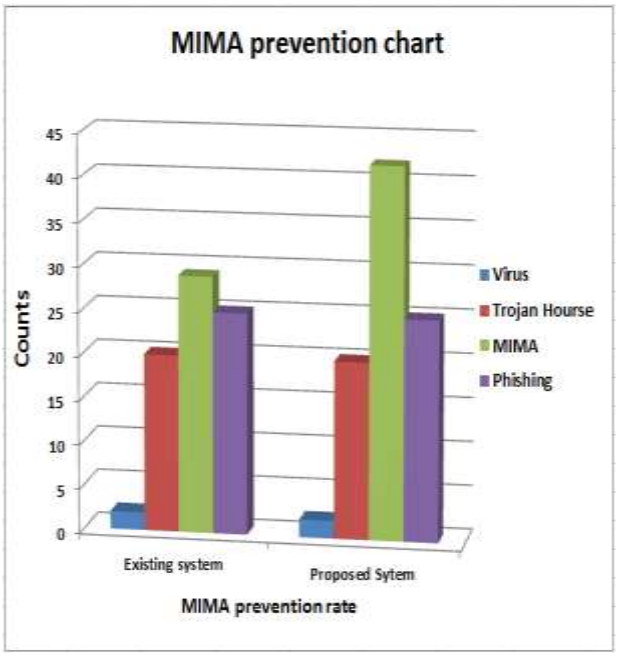

Fig.4 MIMA prevention rate

Fig.4 shows shows that MIMA is evidently prevented using this system compared to the existing. It offers security to the data in the semantic web using reference links. On the web, attacker constantly watches user practices and is vigilant of web applications. They always try to impose attacks on the network, by even slightly manipulating any content on the web. The user unknowing of the attack access the data which seems to be same as original and gets exposed to these attacks. Since fisching, online payments, e-commerce sites, netbanking etc. gained wide proliferation nowadays, these type of website attacks are very emerging and has become a serious issue.

5. CONCLUSION

The Distributed digital artifact system for MIMA is where resources are distributed among different domains and each domain can communicate with each other. Unlike centralized system, distributed system gives users possibility to check whether the data have been modified. The relevant man in the middle attack is prevented to an extent by ensuring verifiability and reliability. The system ensures that data published within the system interface cannot be accessed anywhere outside the system, with the use of reference trees providing security at an overall level. Any manipulation to the data is efficiently identified and any further access to that data is blocked by informing user that the uniform location has been changed. Here only man in the middle attack is considered. This can be extended to more attacks.

6. REFERENCES

[1] Tobias Kuhn and Michel Dumontier, “Making Digital Artifacts on the Web Verifiable and Reliable”, IEEE Transactions on Knowledge and Data Engineering, Vol NO 99 YEAR 2015

[2] Patrick P. Tsang, Apu Kapadia, Member, IEEE, Cory Cornelius, and Sean W. Smith, “Nymble: Blocking Misbehaving Users in Anonymizing Networks”, IEEE Transactions ON Dependable and Secure Computing

[3] Tobias Kuhn and Michel Dumontier, “Trusty URIs: Verifiable, Immutable, and Permanent Digital Artifacts for Linked Data”, in Proceedings of the 11th Extended Semantic Web Conference (ESWC 2014), ser.Lecture Notes in Computer Science. Springer, 2014

[4] N K Prasanna Anjaneyulu anna, Shaik Nazeer, “Semantic Web Security and Privacy”, Journal of Theoretical and Applied Information Technology

[5] Tobias Kuhn, Christine Chichester, Michael Krauthammer and Michel Dumontier, “Publishing without Publishers:a Decentralized Approach to Dissemination,Retrieval, and Archiving of Data”, arXiv preprint arXiv:1411.2749, 2014

[6] Momi Maity, Neha Verma, Rupali Wadikar, Sayali Shevkar, Prof. V.K. Bhusari, “Providing Security to Web Applications in Anonymizing Networks Using Nymble” International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 1, January 2014 ISSN: 2277 128X

[7] Mingwu Zhang, Xiangyu Zhan, Sunil Prabhakar, “Cost Effective Forward Tracing Data Lineage”, Computer Science Technical Reports. Paper 1669, 2007

[8] S.FarrelL, C.Dannewitz, D.Kutscher, B.Ohlman, A.Keranen, P. Hallam-Baker, “Naming Things with

Hashes", Internet Engineering Task Force (IETF),April 2013

[9] Robert Ikeda and Jennifer Widom, "Data Lineage: A Survey", frmikeda,widomg@cs.stanford.edu

[10] C. Sayers and A. Karp, "Computing the digest of an RDF graph", Mobile and Media Systems Laboratory, HP Laboratories, Palo Alto, USA, Tech. Rep. HPL-2003-235(R.1), 2004.

[11] M. Bellare, O. Goldreich, and S. Goldwasser, "Incremental cryptography: The case of hashing and signing", in Advances in Cryptology — CRYPTO'94. Springer, 1994, pp. 216–233.

[12] R. D. Peng, "Reproducible research in computational science", vol. 334, no. 6060, p. 1226, 2011.