POSTER: An Online Prefix-Preserving IP Address Anonymization Algorithm for Passive Measurement Systems

Kai Cao[1], Yunchun Li[1(✉)], Hailong Yang[1], Jiqiang Tang[2], and Xiaoxiang Zou[2]

[1] Sino-German Joint Software Institute, School of Computer Science and Engineering, Beihang University, Beijing, China
lych@buaa.edu.cn
[2] National Computer Network Emergency Response Technical Team/Coordination Center (CNCERT/CC), Beijing, China

Abstract. To strike a balance between usefulness of network traces and privacy protection, offline prefix-preserving anonymization has been studied extensively to anoymize IP addresses while preserving their prefix nature. In this paper, a novel Dynamic Subtree-scheduling Packet Anonymization scheme called DS-PAn is developed for measurement systems based on the prefix-preserving algorithm Crypto-PAn. DS-PAn makes online anoymization practical to be operated at a high rate, while using less memory compared to precomputed Crypto-PAn. Performance evaluations validate that the proposed algorithm outperforms the conventional anonymization mechanism in terms of computation speed as well as memory requirement.

Keywords: IP address anonymization · Dynamic subtree-scheduling · Crypto-PAn

1 Introduction

Network traces are valuable data for network researchers. Sensitive header fields need to be sanitized before the trace is made public. Prefix-preserving IP address anonymization is implemented in TCPDpriv[1] and Crypto-PAn[2], and seems to be suit for offline way. However, when online anonymization is required with a case that traffic traces are anonymized as soon as they are collected in a measurement node, the performance of offline anonymization algorithm should be improved. In this paper, we present a novel IP address anonymization algorithm based on Crypto-PAn and it is able to anonymize IP address at line speed with moderate memory requirement.

2 Crypto-PAn

The anonymization is a one to one mapping from original IP addresses to anonymized ones. Let f_i be a function from $\{0,1\}^i$ to $\{0,1\}$, for $i = 1,2,\cdots,31$ and f_0 is a constant function, and f_i is defined as

$$f_i(a_1a_2\cdots a_i) \coloneqq L\big(R(P(a_1a_2\cdots a_i),k)\big)$$

B. Thuraisingham et al. (Eds.): SecureComm 2015, LNICST 164, pp. 581–584, 2015.
DOI: 10.1007/978-3-319-28865-9_35

where L returns the "least significant bit" and R represents Rijndael cryptographic computation and P is a padding function. Then given the original address $a = a^1a^2 \cdots a^n$, the anonymization function could be defined as:

$$F(a) \coloneqq a'_1a'_2 \cdots a'_n$$

where $a'_i = a_i \oplus f_{i-1}(a_1a_2 \cdots a_{i-1}),\ i = 1,2 \cdots n.$

Since the input of f_i is a bit sequence whose length varies from 0 to 31, and the output is a 0 or 1, the results of f_i can be organized as a perfect binary tree, as shown by Figure1, where a black node represents 1 and a white node represents 0.

3 The Proposed Algorithm

3.1 DS-PAn Algorithm

For the proposed anonymization scheme DS-PAn, the anonymization tree is divided into two parts: the first k levels of the anonymization tree (level 0 to level k-1) stay unchanged, and the remaining part of anonymization tree is comprised of 2^k subtrees, as shown in Figure1.

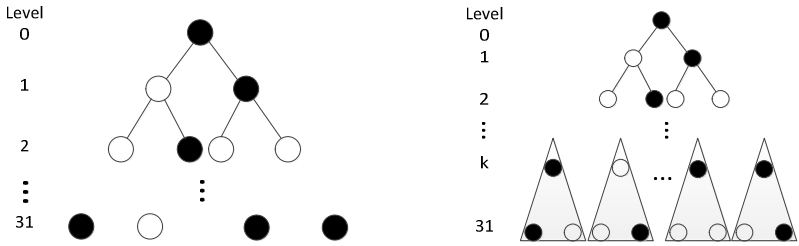

Fig. 1. Anonymization trees of Crypto-PAn and DS-PAn

When k is 24, for example, the the first 24 bits of the original IP address is anonymized as precomputed Crypto-PAn, and the remaining 8 bits are anonymized according to one of the 2^{24} subtrees. If the desired subtree has been computed and stored in memory, it is accessed directly, otherwise the corresponding subtree has to be calculated and stored to memory for later use.

As more IP addresses are anonymized, the subtrees stored in memory will increase gradually. Subtree removal is necessay when memory limitation is reached. We refer to this strategy that subtrees are dynamically constructed and destructed during the anonymization process as subtree scheduling.

When k is larger than 24, the size of subtree is smaller, thus constructing a subtree is quicker. However, the number of subtrees grows, so managing these subtrees is more time-consuming. When k is set to a smaller number, constructing a subtree may take more time, but it is less likely that an inserting or removing action is needed.

3.2 Detail of Anonymization Tree

The detailed design of DS-PAn is demonstrated in Figure2. For simplicity, it shows a scenario in which only 3 subtrees exist in memory. If the maximum size of pointers array is 32, for example, then 5 bits is long enough for each pointer index. A pointer index either denotes the position of a pointer in pointer array or is null, which means the corresponding subtree is not in memory.

When accessing a subtree, DS-PAn first look up the pointer index using the k-bit prefix of the original IP address, if the corresponding index is not null, then the position of the pointer to the desired subtree can be reached directly in pointer array by index. Otherwise, the subtree is not in memory, it need to be computed immediately and the corresponding pointer need to be inserted into a proper position in pointer array.

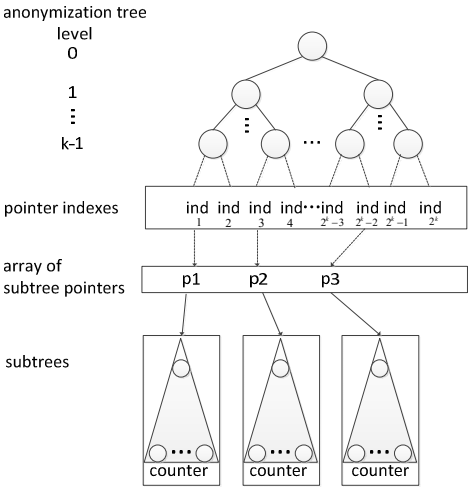

Fig. 2. Subtree-scheduling anonymization tree

Note that every subtree has a counter with it. The counter counts the time of accesses, and is used to determine which subtree should be removed when necessary.

4 Performance Evaluation

The performance of different algorithms is compared and listed in Table 1.

Table 1. Performance of Crypto-PAn and DS-PAn

	Initialization time (s)	Speed (IP addresses /s)	Memory (MB)
Crypto-PAn	0	344687	0
Crypto-PAn (precomputed)	113	3396960	512
DS-PAn (k=21)	0.056	1448964	71

5 Conclusion

In this paper, we presented a novel prefix-preserving IP address anonymization algorithm called DS-PAn which is capable of online IP address anonymization on commodity hardware. When adequately configured, DS-PAn is able to provide link-rate anonymization speed while eliminating the initialization delay and requiring small memory. The performance improvement is achieved by precomputation and the utilization of localized distribution of IP addresses in network traces, thus the security level of DS-PAn is completely the same as Crypto-PAn.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No. 61361126011) and National Hi-tech R&D program of China (863 program) (Grant No. 2015AA01A301).

References

1. Minshall, G.: Pdpfiv. http://fly.isti.cnr.it/software/tcpdpriv/
2. Xu, J., Fan, J., Ammar, M.H., et al.: Prefix-preserving ip address anonymization: measurement-based security evaluation and a new cryptography-based scheme. In: Proceedings of the 10th IEEE International Conference on Network Protocols, 2002, pp. 280–289. IEEE (2002)
3. Ramaswamy, R., Wolf, T.: High-speed prefix-preserving IP address anonymization for passive measurement systems. IEEE/ACM Transactions on Networking **15**(1), 26–39 (2007)
4. Seppänen, K.: A fast and secure method for anonymizing packet traffic and call traces. In: Proceedings of the 12th WSEAS International Conference on Communications. World Scientific and Engineering Academy and Society (WSEAS), pp. 340–346 (2008)
5. Zhang, P., Huang, X., Luo, M., et al.: Fast restorable prefix-preserving IP address anonymization for IPv4/IPv6. The Journal of China Universities of Posts and Telecommunications **17**, 93–98 (2010)