Poster: High Performance Data Leak Detection

Xiaokui Shu, Jing Zhang, Danfeng (Daphne) Yao, Wu-chun Feng
Computer Science Department
Virginia Tech
Blacksburg, Virginia
{*subx,danfeng,jing85,feng*}*@cs.vt.edu*

***Keywords*-Data leak detection, deep packet inspection, algorithm, parallelism, dynamic programming.**

Data leaks on personal computers and organization networks may be caused by mistakes of users (e.g., accidentally sending sensitive documents to a wrong recipient), stolen laptops, malicious insiders, and infection with malicious software. Preventing data leaks typically require a suite of security technologies, including firewalls and intrusion detection/prevention systems at the network, as well as file encryption, strong authentication, and high system assurance tools on the hosts. The multiple points of security measures are complementary to each other and this paper addresses the problem of network-based data leak detection, specifically, how to accurately and efficiently inspect traffic content for patterns of sensitive information. The data may be transformed during the transmission. Network-based data leak detection (DLD) imposes unique research challenges, including:

- *Data Characteristics*: to detect long and non-repeated data leaks (e.g., documents, source code and binaries). Data of any kind could be leaked. While a single piece of short sensitive data with obvious patterns – such as SSN – can be handled by Aho-Corasick algorithm [1] and its extensions, long and non-repeated data is not well tackled by existing methods.
- *Accuracy*: to detect under noisy traffic condition, specifically the leaked data may be transformed during the transmission (e.g., extra or deleted characters due to an application). The transformation renders exact pattern matching useless. Traffic inspection cannot handle leak cases where strong encryption is used.
- *Scalability*: to process a large volume of traffic and sensitive data patterns efficiently.

An exact string matching approach may fail to detect leaks, as it has low tolerance for unknown noises in the network traffic. For example, characters inserted into the packet payload (e.g., due to formatting) may destroy the match. The scalability of such an exact-match approach is also extremely low, especially when there are many long and unique sensitive data patterns to match. Regular expression based comparison supports wild cards and thus can tolerate some traffic noises in the detection. However, it is not scalable and impractical to deploy. This limitation is due to the high complexity of all possible matching patterns and associated performance issues [2]. Thus, conventional string matching algorithms used by intrusion detection system (e.g., Snort [3]) are inadequate for detecting data leaks.

Set intersection of n-grams has been used to detect similar documents on the web [4], shared malicious traffic patterns [5], malware [6], as well as data leaks [7]. One performs set intersection on the two sets of n-grams to determine whether there are any sensitive data n-grams appearing in the traffic. The advantage of using n-grams in the detection is that n-grams capture local features of a string, enabling the comparison to tolerate discrepancies. However, the set intersection operation is orderless; the order of matching n-grams in the two sets does not matter, and set-based detection may generate undesirable false alarms.

Our solution to high performance network-based data leak detection is a new local sequence alignment algorithm and a powerful sampling method for detecting the occurrences of sensitive data (which may be partly altered) in outbound traffic. Our algorithms inherit the local feature preservation concept from set-based detection and extend the power of it to perform more accurate detection with a significant lower false positive rate. The use of alignment in analyzing traffic makes the detection robust and tolerant to noises and certain types of data transformation such as insertion. It allows one to accurately identify the existence of sensitive data pieces in the noisy traffic.

Another significant algorithmic contribution of this work is to achieve detection speedup through sampling while it maintains a very high detection precision. Our novel sampling algorithm has a useful property, referred to by us as comparable sampling. It preserves the similarity of two sequences even after they are sampled (and much shorter). More formally, if string a is a subsequence of string b, then their sampled strings, a' and b', also have the subsequence relation, i.e., a' is a subsequence of b'. Naive solutions such as random sampling do not satisfy this requirement, as random sampling may generate two completely different new strings. We provide both the specific solutions and the general requirements for this unique sampling procedure. The key to our solution is to preserve the local context of strings while sampling. Such a powerful sampling algorithm has general applications beyond network security. We then align the sampled traffic stream with the sampled sensitive data to yield a sensitivity score.

Coupled with our sampling algorithm, our newly designed alignment algorithm not only takes in sampled items, but also carefully keeps track of the information on null regions, which correspond to unselected items during sampling. Both algorithms are designed with the parallel consideration, and we present two non-trivial parallel versions of our algorithms, one with multiple threads on CPU and one for GPU. We demonstrate the strong scalability of our design and the prototypes achieve very good performance.

Our detection method can be made privacy-preserving, specifically, to realize the detection of data leak without revealing any of the original sensitive data. By avoiding the exposure of sensitive data during the detection procedure, operations could be conveniently outsourced to a third party service provider. Furthermore, even if the detection program is compromised, the attacker does not have access to any sensitive data. We realize the privacy-preserving property through the use of a special oneway computation technique and a corresponding protocol described in [8].

Our contributions are summarized as follows and the full details are presented in [9].

1) We present an alignment-based data leak detection model for accurately detecting leak incidents in noisy network traffic.
   We describe a sample-and-align approach, where sampling reduces the size of inputs and alignment identifies traffic content that is similar to (known) sensitive data patterns. Specifically. we invent a novel family of sequence sampling algorithms with a unique property referred to as *comparable sampling*. Unlike random sampling, comparable sampling preserves the similarity of two original sequences during sampling, allowing sampled sequences to be comparable. Alignment over two sampled sequences poses a new technical challenge in terms of accuracy. We fill in the gap and design a new local alignment algorithm that keeps track of null regions in sampled sequences and makes intelligent inferences of item similarity through dynamic programming. Our new algorithms, in particular the sampling technique, are useful beyond the specific data leak problem.
2) We provide several nontrivial prototypes of our detection system, including two high performance parallel versions. We employ multi-level parallelism in our prototypes, from TCP or application stream level, to score table filling in dynamic programming alignment.
   We perform extensive experimental evaluations on the accuracy (namely, false positive and false negative rates) and scalability of our systems with large real-world datasets. Our sample-and-align method achieves the same level of detection accuracy and sensitivity as the non-sampled alignment. For example, averaged sensitivity is 97.5% for sample-and-align and 99.6% for non-sampled alignment. Our method using short n-grams ($n = 3$) is less sensitive to background noise than traditional set-based detections with even longer n-grams ($n = 8$). We also demonstrate that our method detects transformed data leaks much better than set-based detections, especially for English word length modification. Overall speaking, the SNR of detecting real data leaks vs. deceptive noise data is over 10dB for our method and is about 3dB for traditional set-based detection with same parameters.
   For performance evaluation, multiple experiments show that 3% sampling rate gives very accurate detection, which accelerates the detection about 1000 times compared to non-sampled alignment. We also demonstrate high multithreading scalability using our prototypes. Our GPU prototype realizes 100Mbps throughput for 1000 pieces (1000 bytes each) sensitive data.

References

[1] A. V. Aho and M. J. Corasick, "Efficient string matching: An aid to bibliographic search," *Commun. ACM*, vol. 18, no. 6, pp. 333–340, 1975.

[2] S. Kumar, B. Chandrasekaran, J. S. Turner, and G. Varghese, "Curing regular expressions matching algorithms from insomnia, amnesia, and acalculia," in *ANCS*, R. Yavatkar, D. Grunwald, and K. K. Ramakrishnan, Eds. ACM, 2007, pp. 155–164.

[3] M. Roesch, "Snort: Lightweight intrusion detection for networks," in *LISA*. USENIX, 1999, pp. 229–238.

[4] A. Z. Broder, "Identifying and filtering near-duplicate documents," in *COM '00: Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*. London, UK: Springer-Verlag, 2000, pp. 1–10.

[5] M. Cai, K. Hwang, Y.-K. Kwok, S. Song, and Y. Chen, "Collaborative internet worm containment," *IEEE Security and Privacy*, vol. 3, no. 3, pp. 25–33, 2005.

[6] J. Jang, D. Brumley, and S. Venkataraman, "Bitshred: feature hashing malware for scalable triage and semantic analysis," in *Proceedings of the 18th ACM conference on Computer and communications security*, ser. CCS '11. New York, NY, USA: ACM, 2011, pp. 309–320. [Online]. Available: http://doi.acm.org/10.1145/2046707.2046742

[7] K. Li, Z. Zhong, and L. Ramaswamy, "Privacy-aware collaborative spam filtering," *IEEE Transactions on Parallel and Distributed systems*, vol. 20, no. 5, May 2009.

[8] X. Shu and D. Yao, "Data leak detection as a service," in *International Conference on Security and Privacy in Communication Networks*, 2012.

[9] X. Shu, J. Zhang, D. Yao, and W. chun Feng, "Quantitative deep packet inspection with application to high performance data leak detection," Computer Science, Virginia Tech, Tech. Rep., 2013.