Rapid Screening of Transformed Data Leaks with Efficient Algorithms and Parallel Computing*

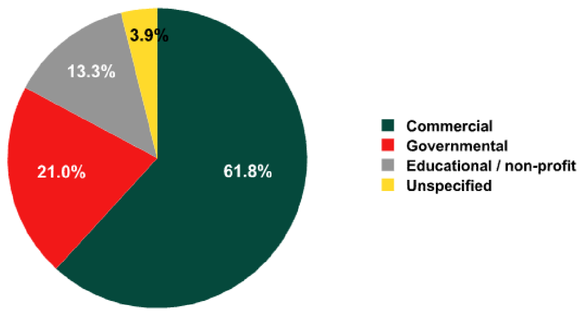
Xiaokui Shu [subx@cs.vt.edu], Jing Zhang, Danfeng (Daphne) Yao [danfeng@cs.vt.edu] and Wu-chun Feng
Department of Computer Science, Virginia Tech, Blacksburg, VA, 24060

Motivation

Detecting data exfiltration in enterprise environments

- 822 million records leaked in 2013
- 765% increases, compared to 2010
- Enterprise is the most significant victim of data exfiltration

(data from InfoWatch)



Our techniques will provide a accurate, fast and scalable data leak detection solution in enterprise environments.

Challenges

- *Data Characteristics:* to detect long and non-repeated data leaks (e.g., documents, source code and binaries).
- *Accuracy:* to detect under noisy traffic conditions, specifically when the leaked data is transformed.
- *Scalability:* to process a large volume of traffic and sensitive data patterns efficiently

Limitations of Existing Algorithms

1. Cannot handle long and non-repeated data efficiently
   Exact string matching method is restricted by wild card implementation complexity.
2. Can be easily deceived by simple data transformations
   Conventional n-gram mechanisms cannot use short n-grams because of high false positives.

Our Solution

Comparable Sampling + Sample-oblivious Alignment

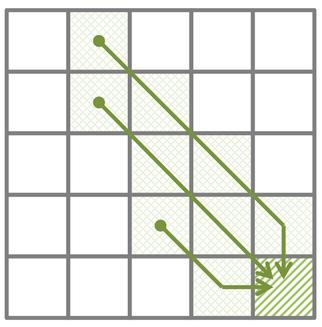
Comparable sampling: where if string *a* is similar to a substring of string *b* according to a similarity measure M, then their sampled subsequences (*a'* and *b'*) are also similar according to M.

```
1, 9, 4, 5, 3, 5, 9, 7, 6, 6, 3, 3, 7, 1
1, 9, 4, 5, 3, 5, 9, 7, 6, 6, 3, 3, 7, 1

A comparable sampling may give:
1, -, 4, -, 3, 5, -, -, -, -, 3, 3, -, 1
1, -, 4, -, 3, 5, -, -, -, -, 3, 3, -, 1

A random sampling may give:
1, -, 4, -, 3, 5, -, 7, -, 6, -, -, 7, 1
-, 9, -, 5, -, 5, -, 7, -, 6, 3, -, -, 1
```

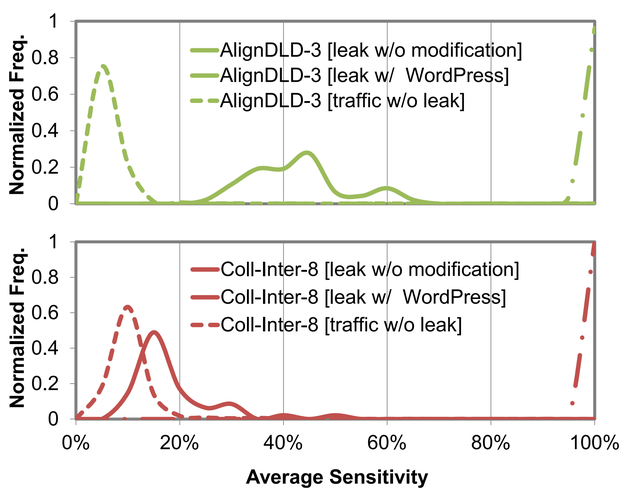
- Sliding window
- Local selection function
- Local environment understanding
- Comparable sampling



- Dynamic programing
- Local alignment
- Accurate unselected items inference
- Extra fields for null-region information
- Deep substructures for local problems

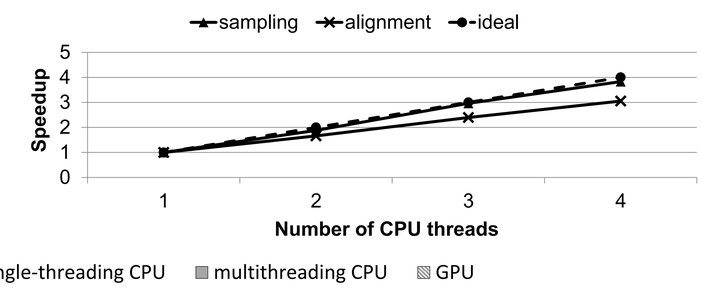
Sampling oblivion: the ability to infer the similarity of raw streams through sampled streams in an alignment; it makes the alignment transparent to the sampling procedure.

Evaluations

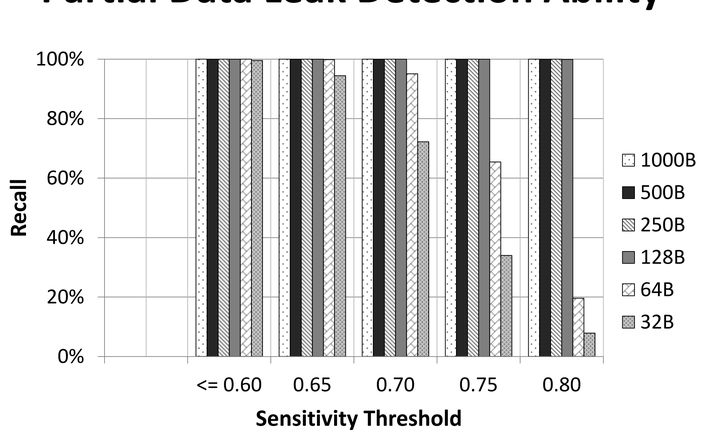


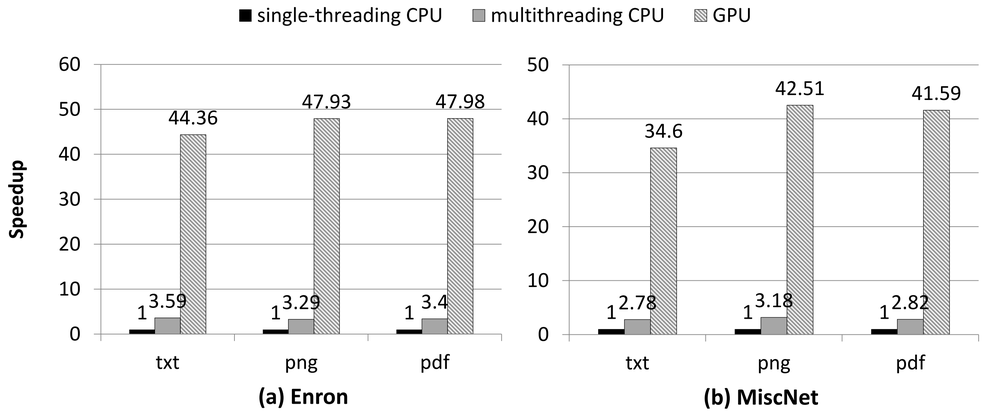
Transformed Leak Detection Ability

Our GPGPU prototype demonstrates a 50 times speedup over normal single threaded CPU version. Our system reaches a throughput at 110Mbps against 1MB sensitive data.



Partial Data Leak Detection Ability





Conclusion

We designed a pair of algorithms to perform accurate, fast and scalable data leak detection in enterprise networks. We prototyped our system on GPGPU and evaluated its accuracy and performance against real-world leaking scenarios.

*This work has been supported in part by Security and Software Engineering Research Center (S²ERC), a NSF sponsored multi-university Industry/University Cooperative Research Center (I/UCRC).

VirginiaTech *Invent the Future*

Yao Group on Cyber Security
http://people.cs.vt.edu/danfeng/

acm SIGSAC **CODASPY 2015**