Network Anomaly Detection Using Probabilistic Sketch Data Structures

EN.601.714 - Advanced Computer Networks

Omar Ahmed

Overview of Presentation

- Define the network anomaly detection problem & classes of solutions

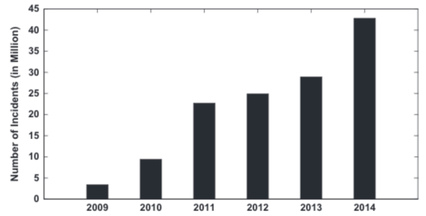


Fig. 1. Growth of information security incidents (The Global State of Information Security Survey, 2015).

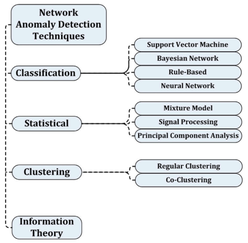


Fig. 3. Taxonomy of network anomaly detection techniques.

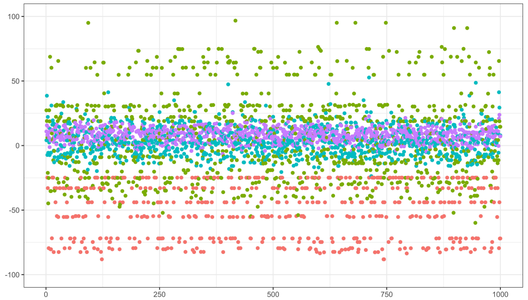
- Discuss my approach using sketching data-structures

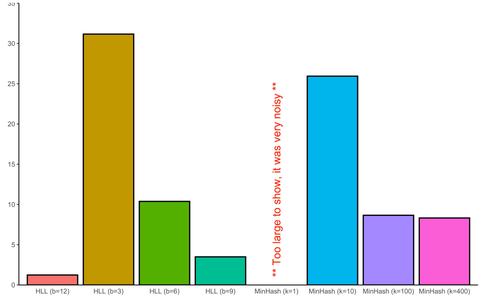
hash()

0b0000110
0b0010110
0b1010110

0b1000110

- Discuss the current progress of my solution





Network Anomaly Detection

- Lets start with the definition of anomaly:[1]
  - *"An anomaly is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism"*
    - Douglas Hawkins in *Identification of Outliers*

- Network anomalies typically fall into two broad categories[2]
  - **Performance-related:** broadcast storms, congestion, etc.
  - **Security-related:** Malicious users traffic, DoS attacks, Probe, etc.

i) **point** ii) **contextual** iii) **collective**

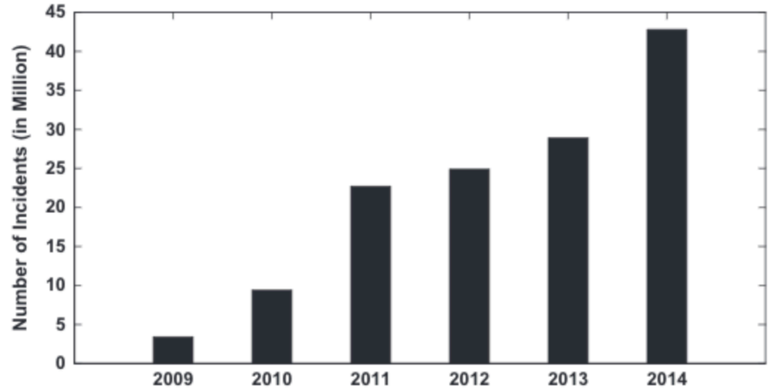


Fig. 1. Growth of information security incidents (The Global State of Information Security Survey, 2015).

Motivation: The faster and more accurately we can detect anomalies, the quicker we can respond

[1] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19-31.

[2] Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2013). Network anomaly detection: methods, systems and tools. *Ieee communications surveys & tutorials*, *16*(1), 303-336.

Formal Problem Definition & Classes of Solutions

- A network anomaly detection approach can be thought of in these terms: [2]
  - $S = (M, D)$ where *S* is the anomaly detection engine.
    - *M* is the model of normal behavior (e.g. normal packet traces)
    - *D* is proximity measure to compute how close we are to "normal"

- An overview of current types of approaches: [1]

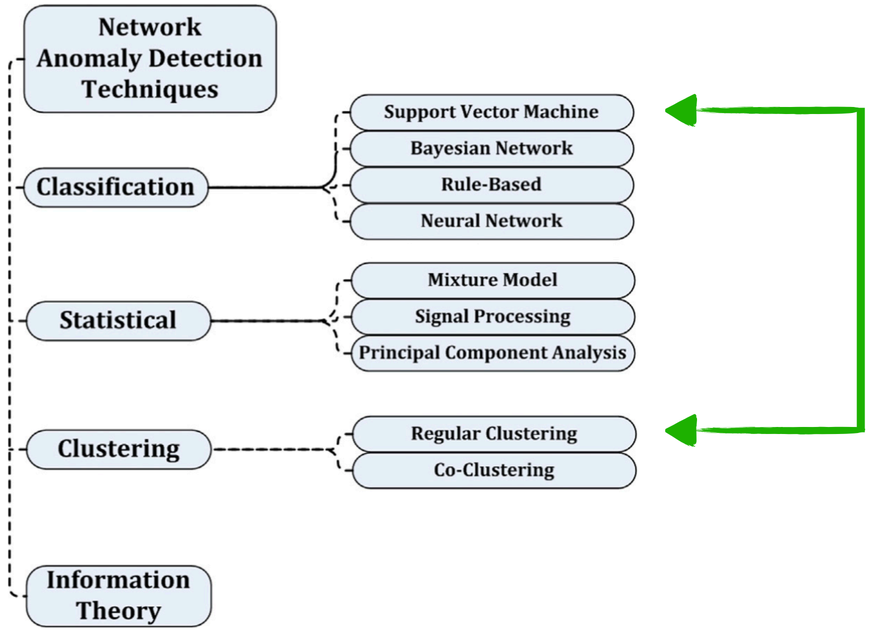


Fig. 3. Taxonomy of network anomaly detection techniques.

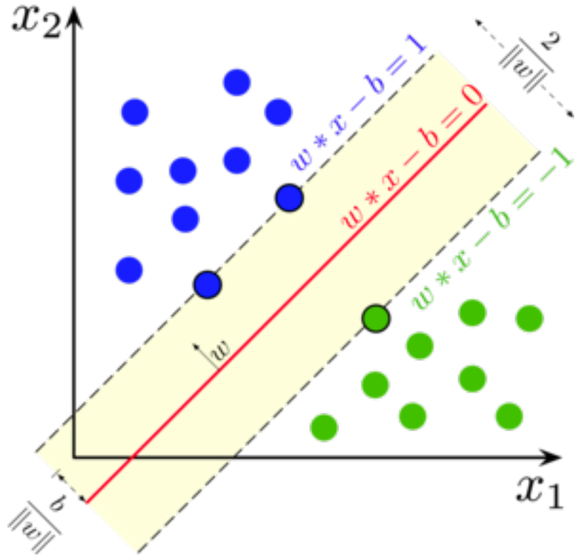
Lets take a quick look at some of these methods

[1] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19-31.

[2] Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2013). Network anomaly detection: methods, systems and tools. *Ieee communications surveys & tutorials*, *16*(1), 303-336.

Overview of Related Methods

- **Classification Method - Support Vector Machine:**
  - Presented in [3] where they used an unsupervised version of SVM detect anomalous network packets.
  - Data consisted of TCP connection information from KDDCup 1999 dataset, and features included duration, bytes transferred, flags, etc.
  - Results consisted of ROC curve showing the SVM performed well in the classification of TCP records.



[3] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security* (pp. 77-101). Springer, Boston, MA.

Overview of Related Methods

Clustering - Regular Clustering:

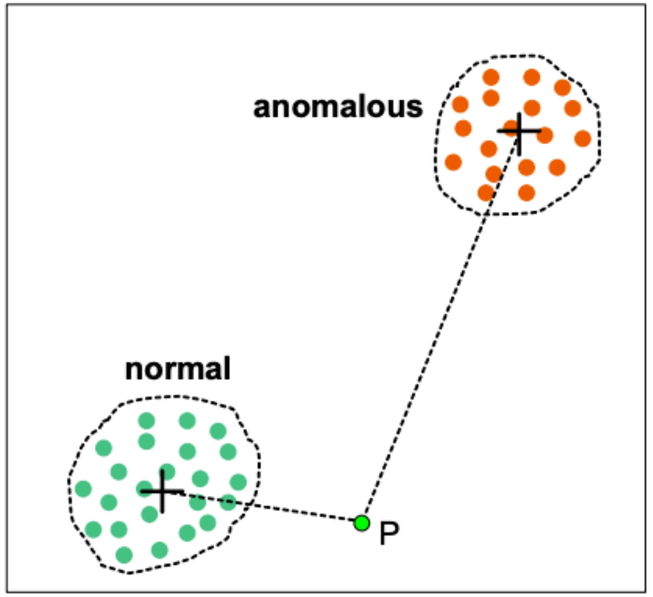


Fig. 2. Classification for $K = 2$

- Presented in [5] and they focus on determining time intervals that have anomalous traffic.
- Convert flow records into **a set of features** like number of packets, number of bytes, and number of different source-destination pairs, and then cluster.
- Use distance from centroid to classify packets from a time interval

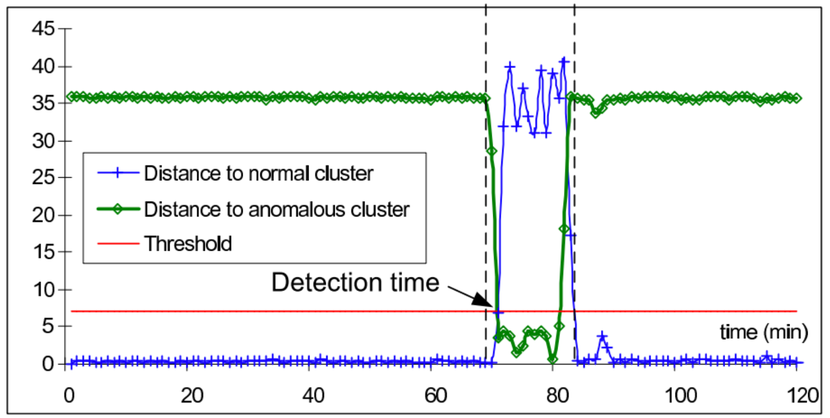


Fig. 5. Ping flood detection with generated traffic

[5] Münz, G., Li, S., & Carle, G. (2007, September). Traffic anomaly detection using k-means clustering. In *GI/ITG Workshop MMBnet* (pp. 13-14).

Overview of Presentation

- Define the network anomaly detection problem & classes of solutions

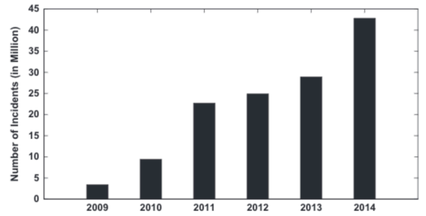


Fig. 1. Growth of information security incidents (The Global State of Information Security Survey, 2015).

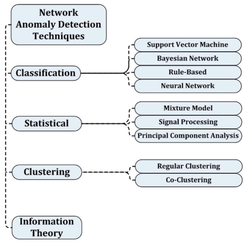


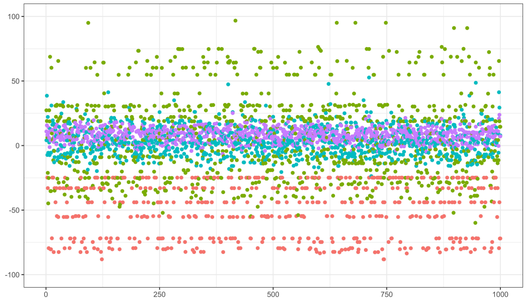
Fig. 3. Taxonomy of network anomaly detection techniques.

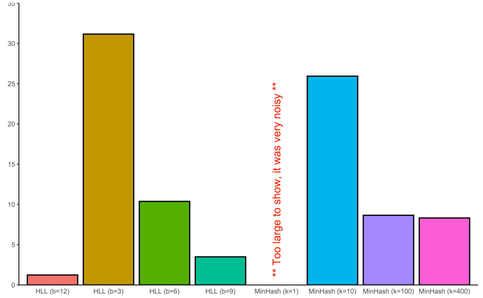
- Discuss my approach using sketching data-structures


```
0b0000110
 0b0010110
  0b1010110

0b1000110
```

- Discuss the current progress of my solution





Quick Overview of My Approach

- Goal: Develop an approach that can detect anomalies …

  (i) accurately (ii) in online fashion (ii) that can attempt to classify the type of anomaly

How do we want to approach this problem?

- My approach: Use sketching data-structures that are designed for cardinality, and set-operations
  - The HyperLogLog data-structure [6] has been used for cardinality and set comparisons in various settings like with websites with Google, and genome comparisons.
  - Similarly with MinHash [7], it was used for comparing webpages in the AltaVista search engine.

Key Idea: These data-structures could allow us to detect anomalies in real-time accurately.

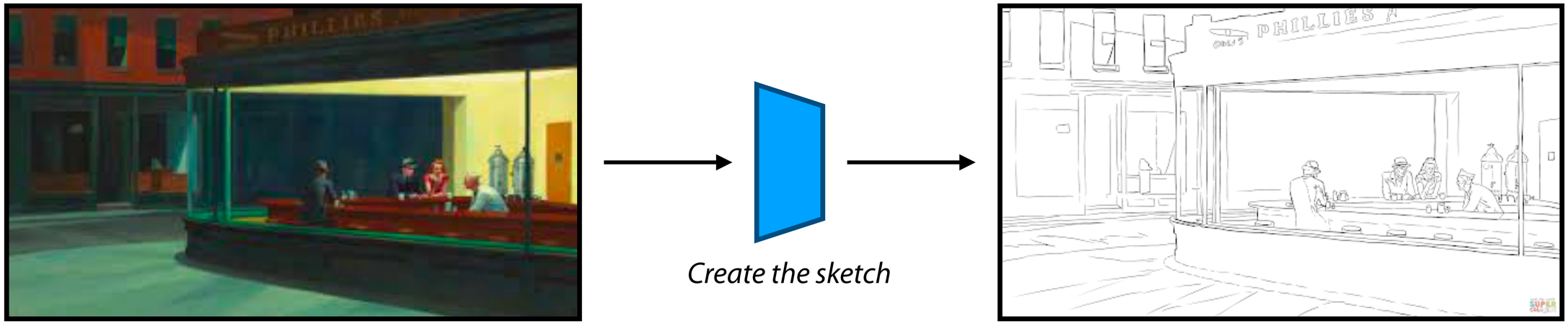
[6] Flajolet, P., Fusy, É., Gandouet, O., & Meunier, F. (2007, June). Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. In *Discrete Mathematics and Theoretical Computer Science* (pp. 137-156). Discrete Mathematics and Theoretical Computer Science.

[7] Broder, A. Z. (1997, June). On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)* (pp. 21-29). IEEE.

What is a sketch?

- Say you are Google, and you want to know how many distinct src IPs you see in a day …
  - Straightforward, but not practical: `int* all_src_ips = new int[very_big_number];`
  - Maybe we can randomly sample, however we could be biased by frequent items.

- A sketch is a smaller, yet informative set of data about a large dataset.



What is a sketch?

- Say you are Google, and you want to know how many distinct src IPs you see in a day …
  - Straightforward, but not practical: `int* all_src_ips = new int[very_big_number];`
  - Maybe we can randomly sample, however we could be biased by frequent items.

- A sketch is a smaller, yet informative set of data about a large dataset.

Types of Different Sketching Data-Structures

Approximate Set Membership	*Cardinality Estimation*	*Frequency Estimation*
Bloom Filter	MinHash	Count-Min Sketch
Cuckoo Filter	HyperLogLog	Count Sketch

MinHash Data Structure

- What is the intuition behind MinHash?
  - Let's assume we pick 5 numbers between 0 and 1 …

  - If we can keep track of minimum hash, we can calculate the cardinality!

$$E[X] = \frac{1}{min\_hash}$$

- How does is it used in practice usually?
  - Using just smallest hash can be noisy, so there are other options …

Averaging the randomness to improve the confidence

- Track k smallest hashes
- Use k hash functions, and track smallest hash from each hash functions.
- Split hash range into k buckets, and keep track of smallest in each bucket

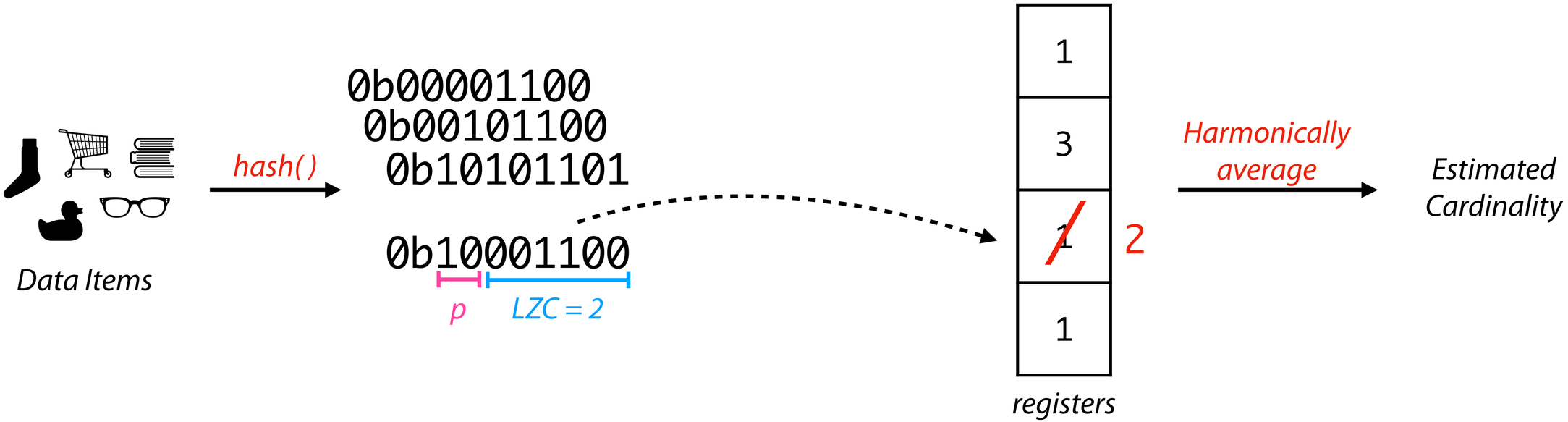
HyperLogLog Data Structure

- What is the intuition behind HyperLogLog?
  - Let's assume we have a stream of binary numbers …

`LZC(00110011) = 2` `LZC(00011110) = 3`

  - What's the probability of a LZC of x? $(\frac{1}{2})^x$

- How is this intuition used?



Overview of Presentation

- Define the network anomaly detection problem & classes of solutions

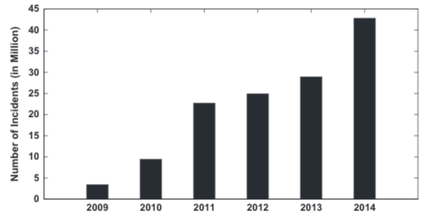


Fig. 1. Growth of information security incidents (The Global State of Information Security Survey, 2015).

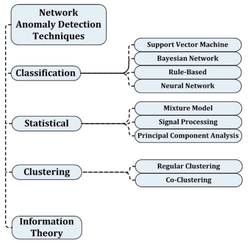


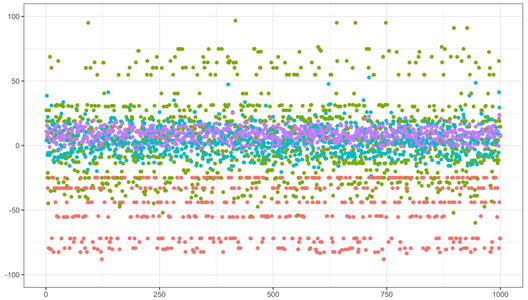
Fig. 3. Taxonomy of network anomaly detection techniques.

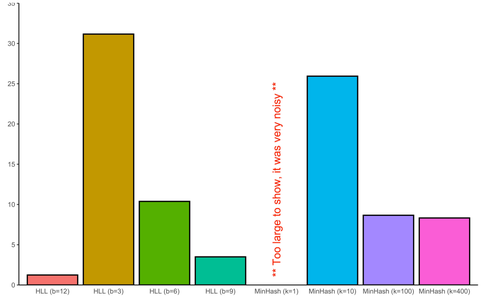
- Discuss my approach using sketching data-structures


```
0b0000110
 0b0010110
  0b1010110

0b1000110
```

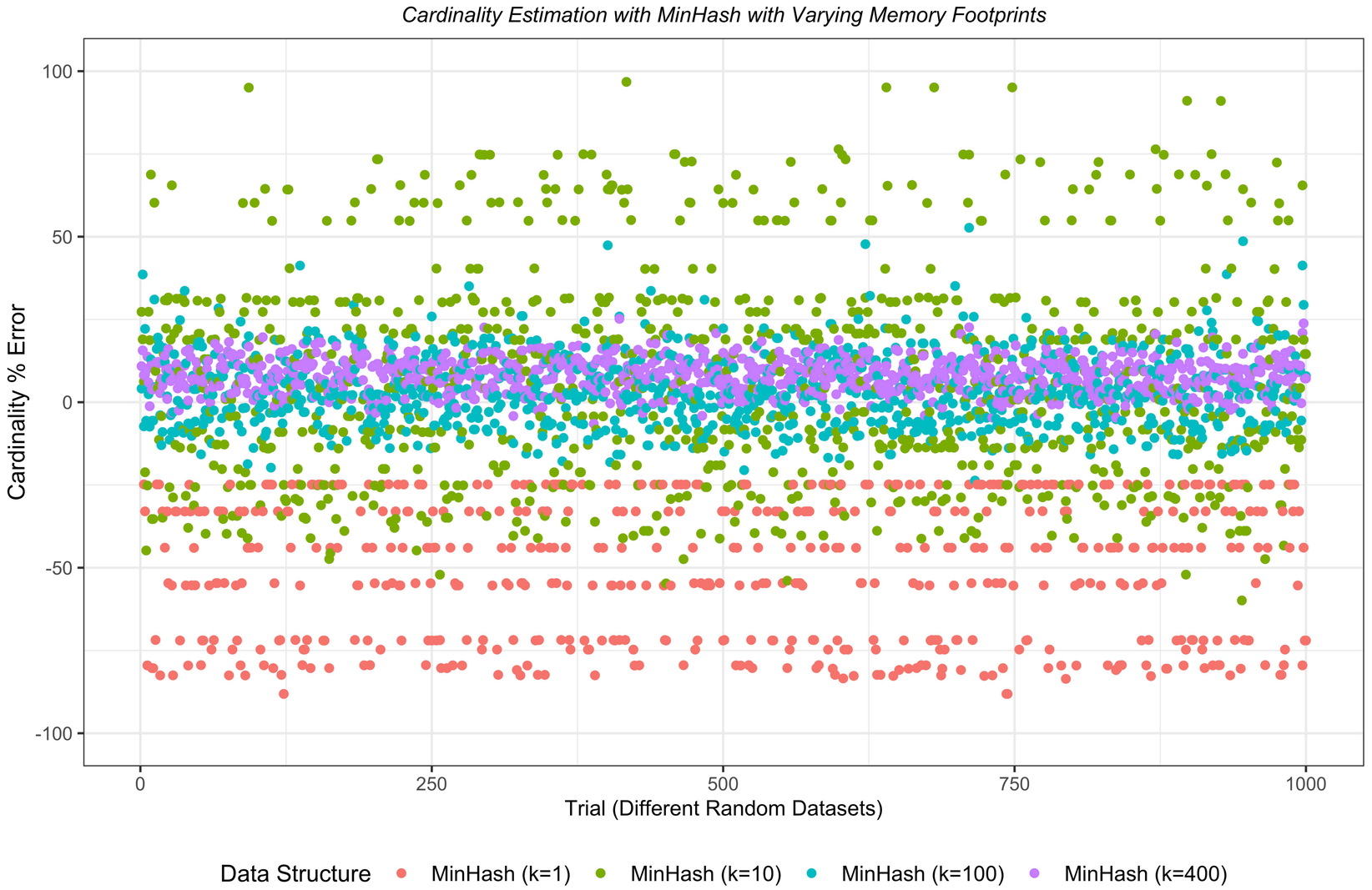
- Discuss the current progress of my solution





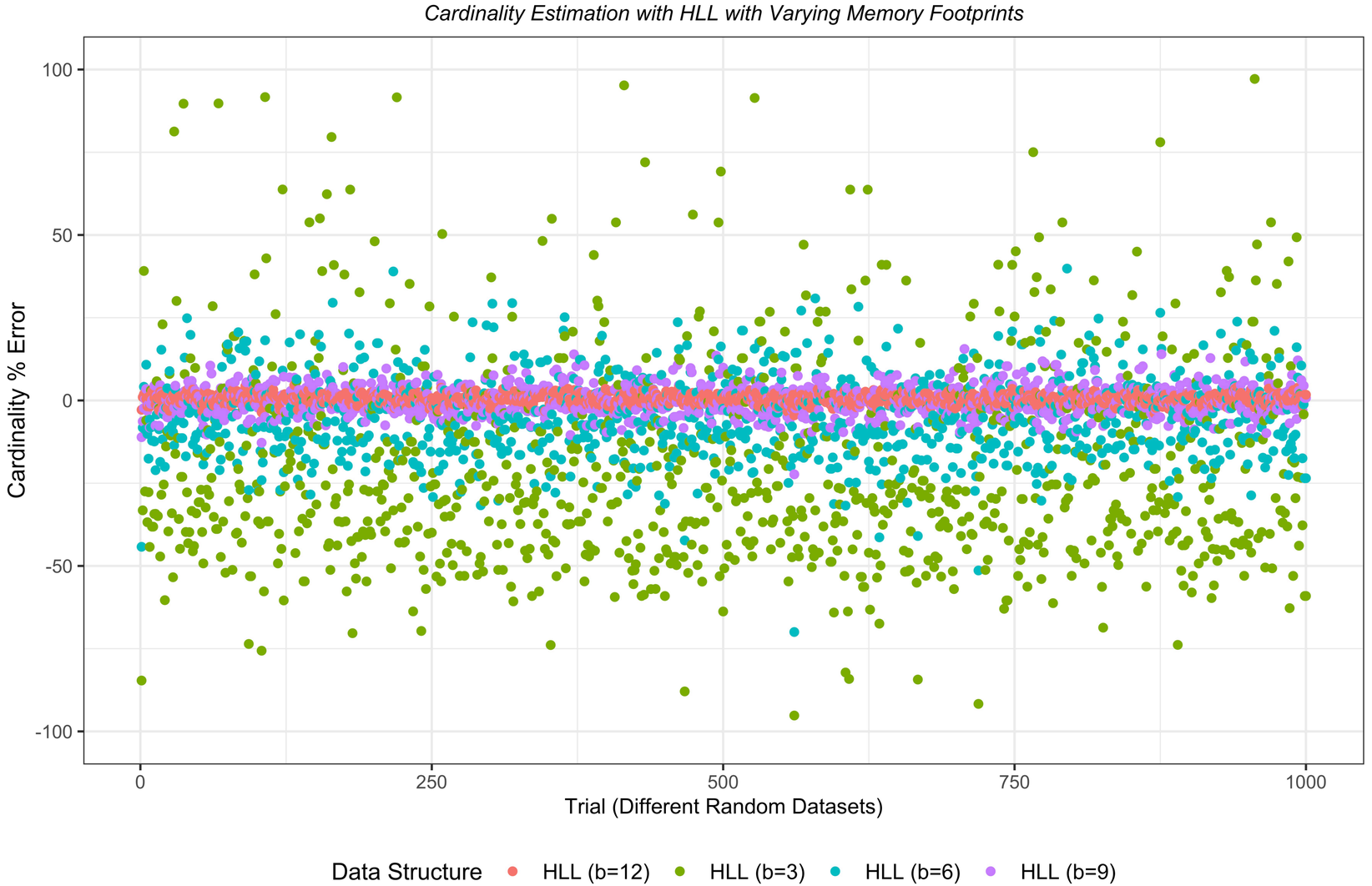
Implementing the Data Structures

- Implemented the MinHash data structure in C++



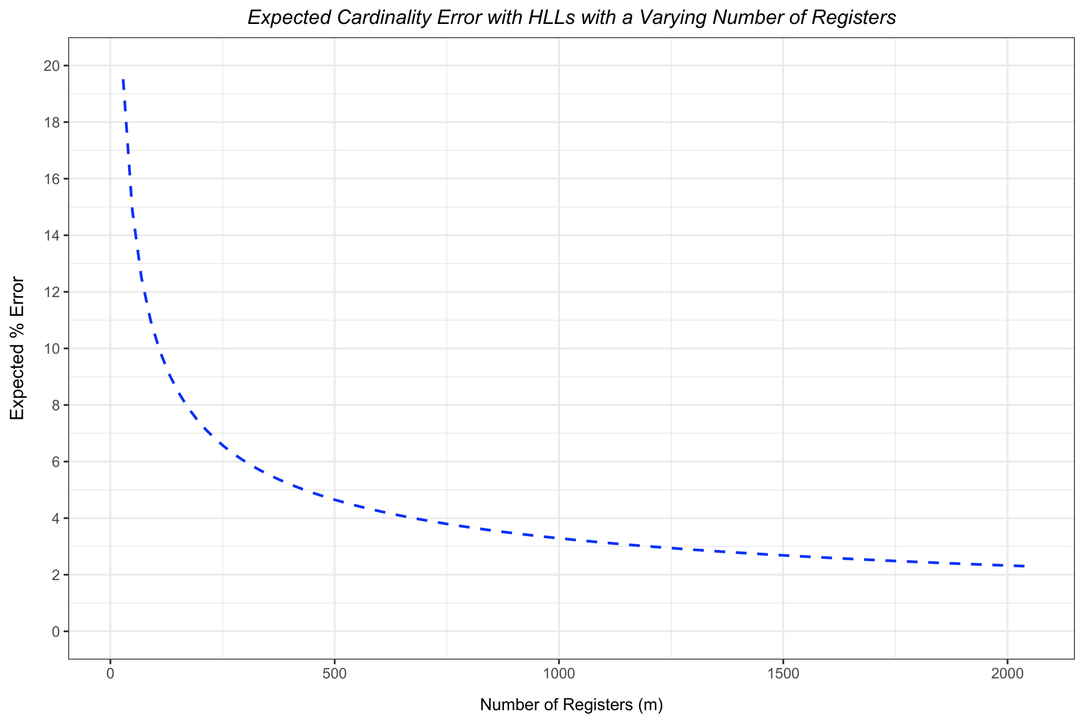
Implementing the Data Structures

- Implemented the HyperLogLog data structure in C++



Implementing the Data Structures

- Additional results regarding cardinality estimation



$$error = \frac{1.04}{\sqrt{m}}$$

Discussion of Datasets

- Datasets that I plan to work with …
  - KDD Cup 1999 Dataset
    - Widely used across anomaly detection paper, however it has limitations …
    - Contains ~5 million connection records which are labeled as normal or an specific attack
  - NSL-KDD Dataset
    - Fixes some problems of the KDD Cup 1999 dataset such as redundant records
    - Therefore, it is just a subset of the KDD Cup 1999 dataset
  - Data from "*Network Traffic Characteristics of Data Centers in the Wild*" paper
    - Lead author, Dr. Theophilus Benson, has the datasets from paper available. It could be used a normal set of data.
  - Intrusion Detection Dataset 2017
    - Motivation for authors was the lack of newer datasets, that are more realistic.
  - Other potential datasets …
    - Facebook datacenter dataset & CAIDA DDoS attack datasets were requested …

References

Papers:

[1] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19-31.

[2] Bhuyan, M. H., Bhattacharyya, D. K., & Kalita, J. K. (2013). Network anomaly detection: methods, systems and tools. *Ieee communications surveys & tutorials*, *16*(1), 303-336.

[3] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. In *Applications of data mining in computer security* (pp. 77-101). Springer, Boston, MA.

[4] Shyu, M. L., Chen, S. C., Sarinnapakorn, K., & Chang, L. (2003). *A novel anomaly detection scheme based on principal component classifier*. MIAMI UNIV CORAL GABLES FL DEPT OF ELECTRICAL AND COMPUTER ENGINEERING.

[5] Münz, G., Li, S., & Carle, G. (2007, September). Traffic anomaly detection using k-means clustering. In *GI/ITG Workshop MMBnet* (pp. 13-14).

[6] Flajolet, P., Fusy, É., Gandouet, O., & Meunier, F. (2007, June). Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. In *Discrete Mathematics and Theoretical Computer Science* (pp. 137-156). Discrete Mathematics and Theoretical Computer Science.

[7] Broder, A. Z. (1997, June). On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)* (pp. 21-29). IEEE.

References

Images:

- https://towardsdatascience.com/pca-is-not-feature-selection-3344fb764ae6
- https://en.wikipedia.org/wiki/Support-vector_machine
- https://en.wikipedia.org/wiki/Nighthawks_(painting)
- http://www.supercoloring.com/coloring-pages/nighthawks-by-edward-hopper