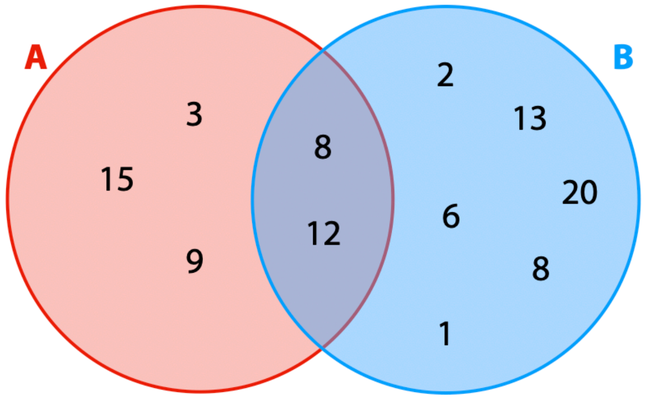
Network Anomaly Detection Using Probabilistic Sketch Data Structures

EN.601.714 - Advanced Computer Networks

Omar Ahmed

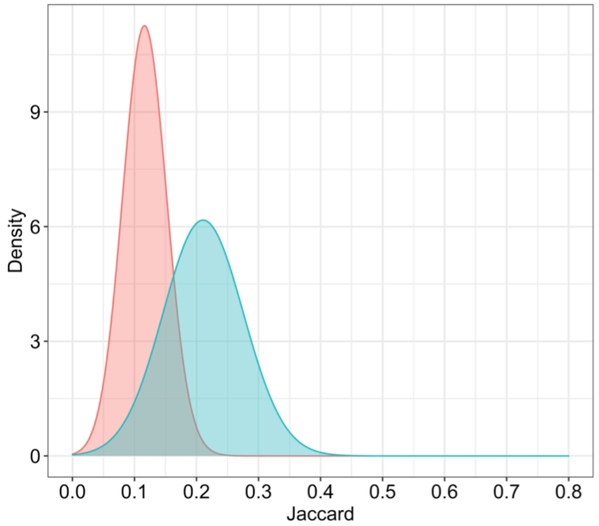
Overview of Presentation

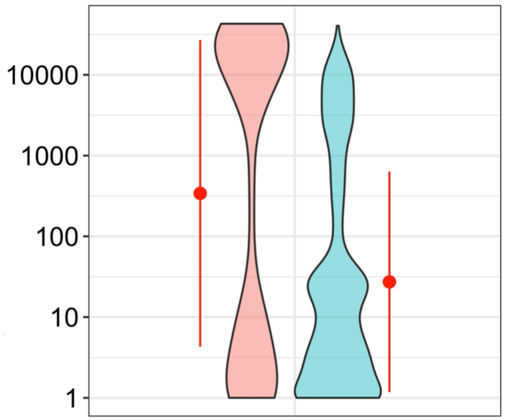
- Overview of my approach for network anomaly detection



$$J(\text{📄}, \text{📄}) = J_n$$
$$J(\text{📄}, \text{📄}) = J_a$$

- Results covering data-structure testing, feature analysis, and jaccard analysis





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, etc.

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

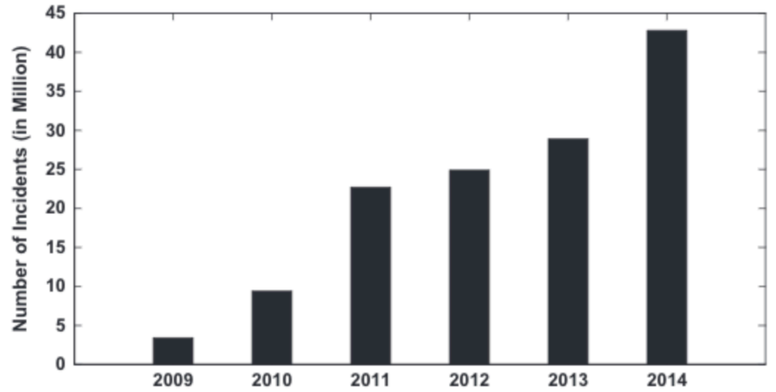


Fig. 1. Growth of information security incidents (The Global State of Information Security Survey, 2015).

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.

Quick Overview of My Approach (UPDATED)

- Goal: Develop an approach that can detect security anomalies in US Air Force LAN …

  (i) accurately (ii) estimate the % of anomalous activity in time window

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 network anomalies accurately.

[3] 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.

[4] 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.

Baseline Approach & Motivation for Sketches

- Baseline approach: Sampling records from a stream of data, like 1 in every 256 records [5]
  - We could be biased by frequent items in the stream, **so what else can we do?**

- My approach: Using sketches which are small, yet informative subsets of the data
  - We can get around the frequent-item bias, and we get other benefits …
  - Sketches allow us to track data-streams without having O(n) space complexity …
    - **Examples**:

(1) Heavy Hitters on Amazon

Item:	Monitor	Air Fryer	Echo	Speaker	…..
Views:	100	45	87	112	

Space Complexity: O(n)

Expensive & Keeps Growing!

(2) Daily Users on Amazon

User Set:	bob1	bluejay07	ravens01	bob2	…..

Same as above!

[5] Tune, P., & Veitch, D. (2011, April). Sampling vs sketching: An information theoretic comparison. In *2011 Proceedings IEEE INFOCOM* (pp. 2105-2113). IEEE.

Baseline Approach & Motivation for Sketches

- Baseline approach: Sampling records from a stream of data, like 1 in every 256 records [5]
  - We could be biased by frequent items in the stream, **so what else can we do?**

- My approach: Using sketches which are small, yet informative subsets of the data
  - We can get around the frequent-item bias, and we get other benefits …
  - Sketches allow us to track data-streams without having O(n) space complexity …

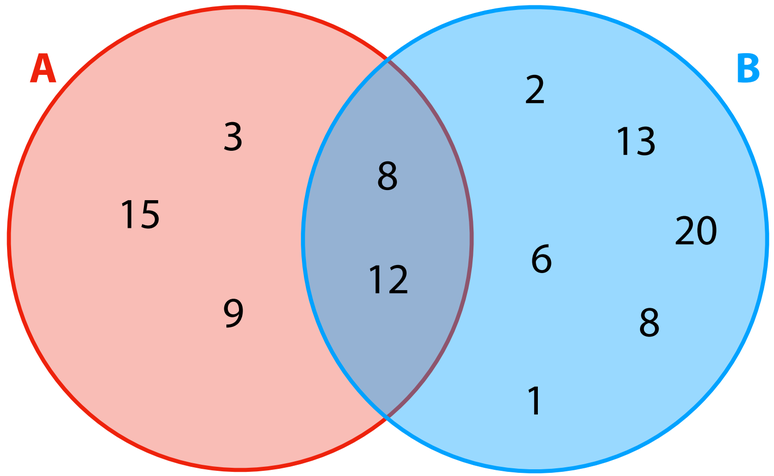
Key Takeaway:
If you sacrifice a bit of "accuracy", you can solve these large problems in smaller space with sketches

[5] Tune, P., & Veitch, D. (2011, April). Sampling vs sketching: An information theoretic comparison. In *2011 Proceedings IEEE INFOCOM* (pp. 2105-2113). IEEE.

MinHash Data Structure

- What is the intuition behind MinHash?
  - If hashes are roughly uniform, we can use the smallest hash to calculate cardinality …

- How can we use MinHash to compare sets?
  - We can compute the Jaccard similarity measure: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{2}{11}$$

HyperLogLog Data Structure

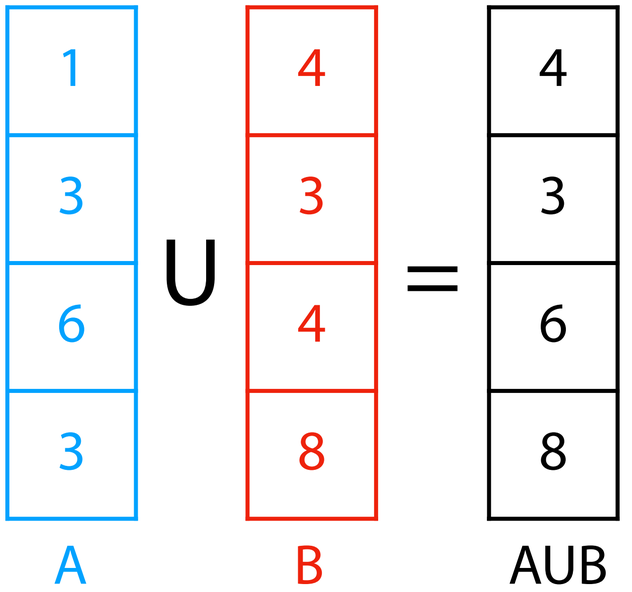
- What is the intuition behind HyperLogLog?
  - Using the leading-zero count can give intuition of set cardinality

LZC(00110011) = 2 $\qquad P(LZC = x) = \left(1/2\right)^x$

- How can we use HyperLogLog to compare sets?
  - Again we can use Jaccard, but use inclusion-exclusion principle

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$$J(A,B) = \frac{|A| + |B| - |A \cup B|}{|A \cup B|}$$



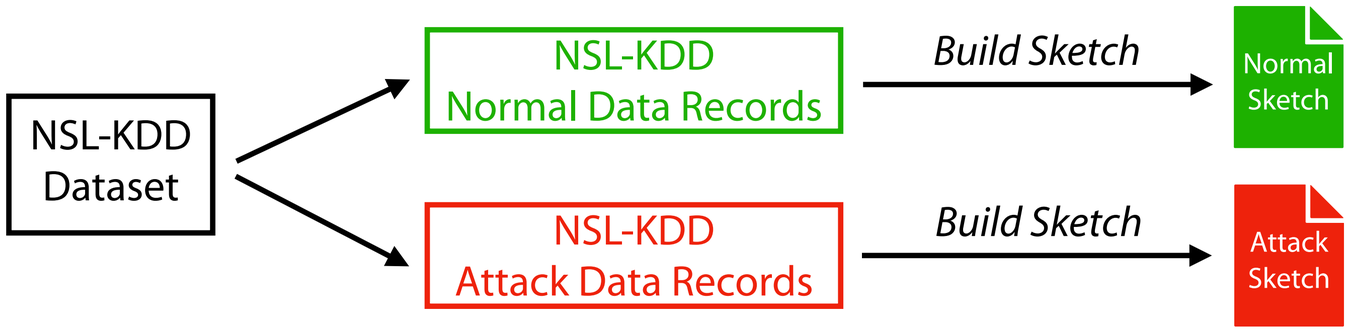
NSL-KDD Dataset

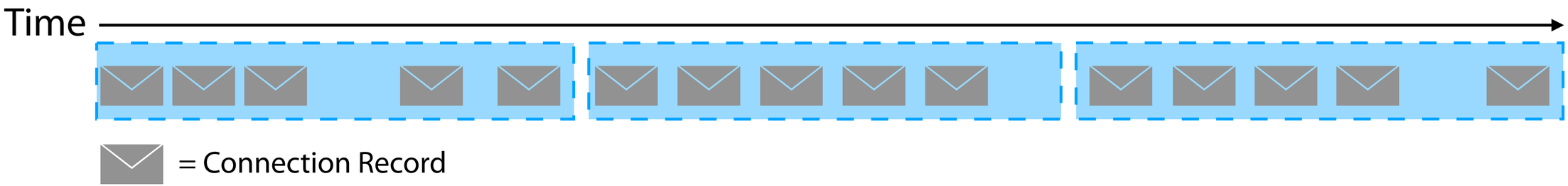
- Main Dataset Used: NSL-KDD [6]
  - It fixes some of the issues in the KDD Cup dataset …
    - Redundant records in the dataset
    - Balancing records of different difficulty
    - Makes it a more manageable size to use entire dataset
  - Contains a total of **311,027 data records** (250,436 attack and 60,591 normal)
    - Total of 41 features: 34 numeric and 7 categorial
      - e.g. duration, bytes sent, num_shell_prompts, etc.
  - Each attack is labelled and there are four broad categories of attacks …
    - DOS: denial-of-service attacks
    - R2L: unauthorized access from remote machine
    - U2R: unauthorized access to local superuser privileges
    - Probing: surveillance and other probing

[6] NSL-KDD dataset. Available at https://www.unb.ca/cic/datasets/nsl.html

Workflow of Approach

- How exactly can we use sketches to identify security anomalies?





Build Sketch

Compare to References

$J(\text{current}, \text{normal}) = J_n$

$J(\text{current}, \text{attack}) = J_a$

Classify Window

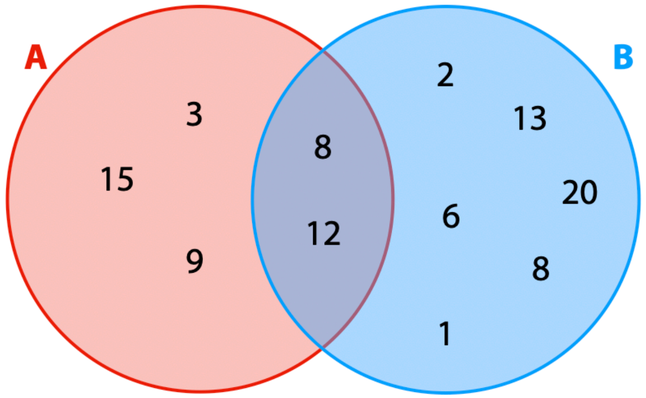
```
if Jn > Ja:
    label = "normal"
else:
    label = "attack"
```

Estimate Anomalous Percentage

$$percent = \frac{J_a}{J_a + J_n}$$

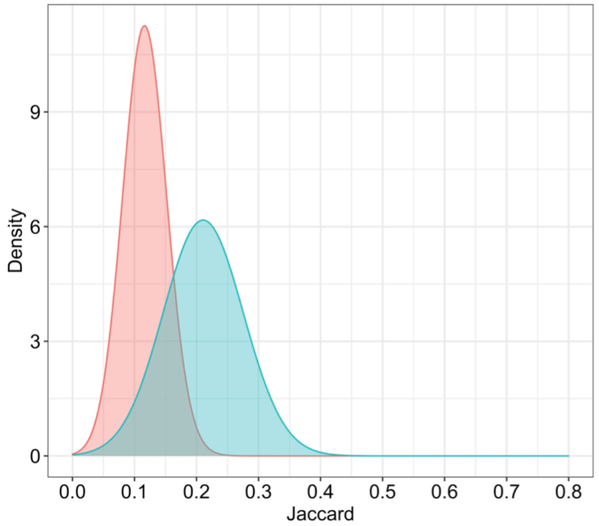
Overview of Presentation

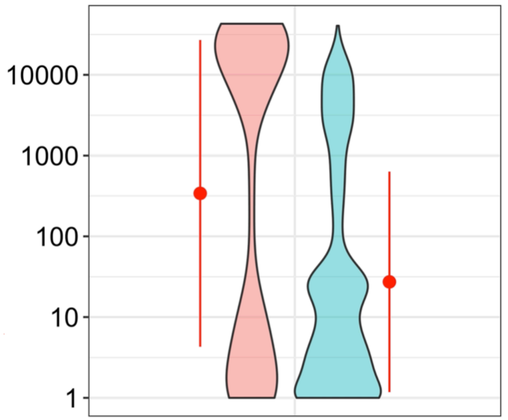
- Overview of my approach for network anomaly detection



$$J(\text{doc}_1, \text{doc}_2) = J_n$$
$$J(\text{doc}_1, \text{doc}_3) = J_a$$

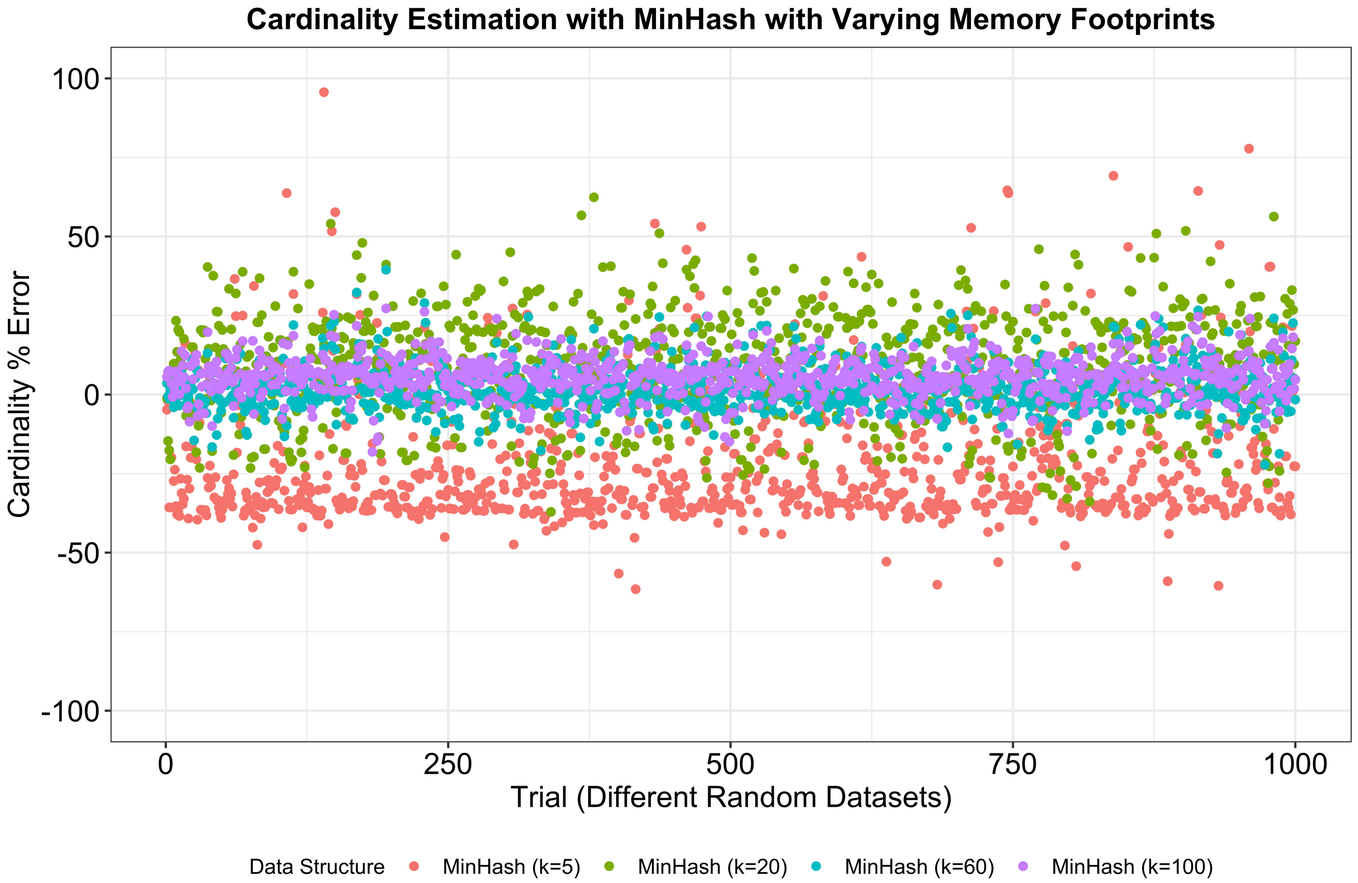
- Results covering data-structure testing, feature analysis, and jaccard analysis

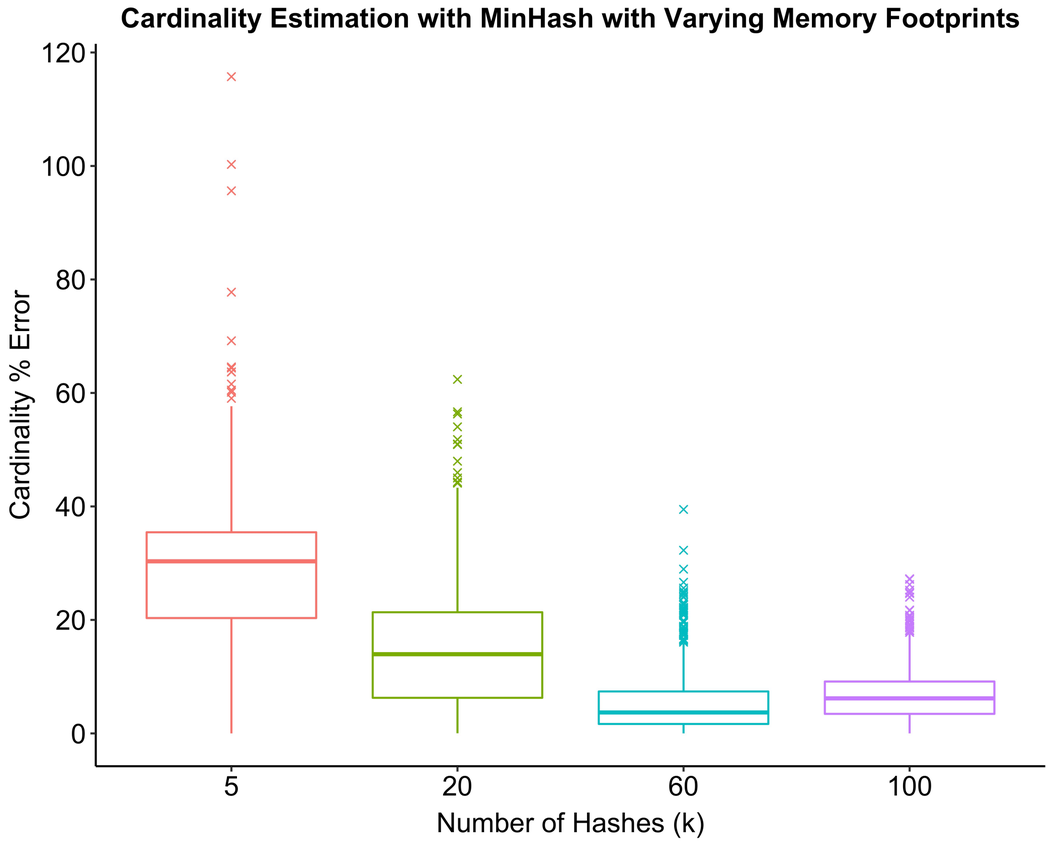




Set Cardinality Estimation with MinHash

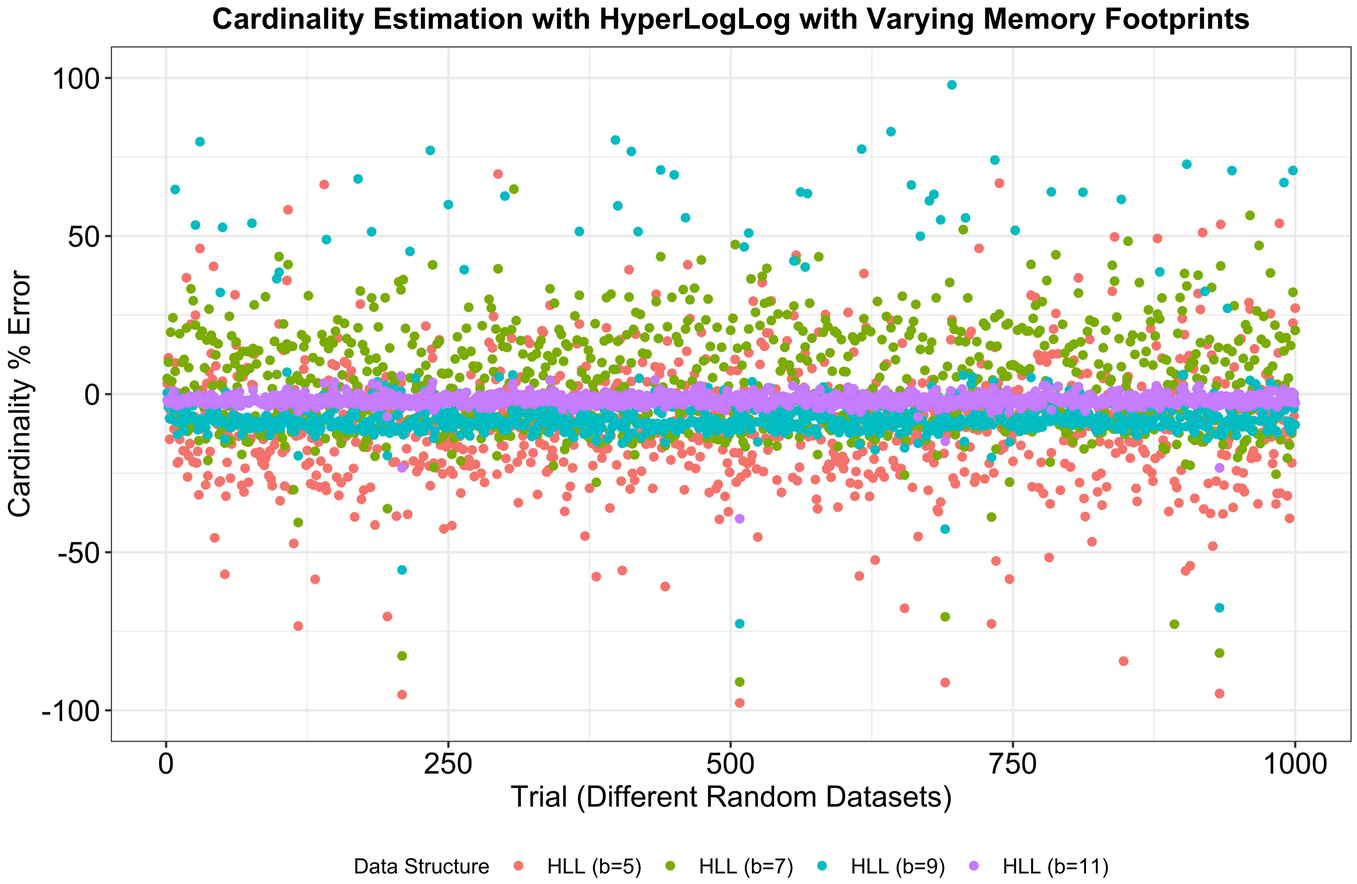
- Question: Does the implemented MinHash estimate set cardinality correctly?
- Answer: It seems to be working. We see the expected pattern as k gets larger …

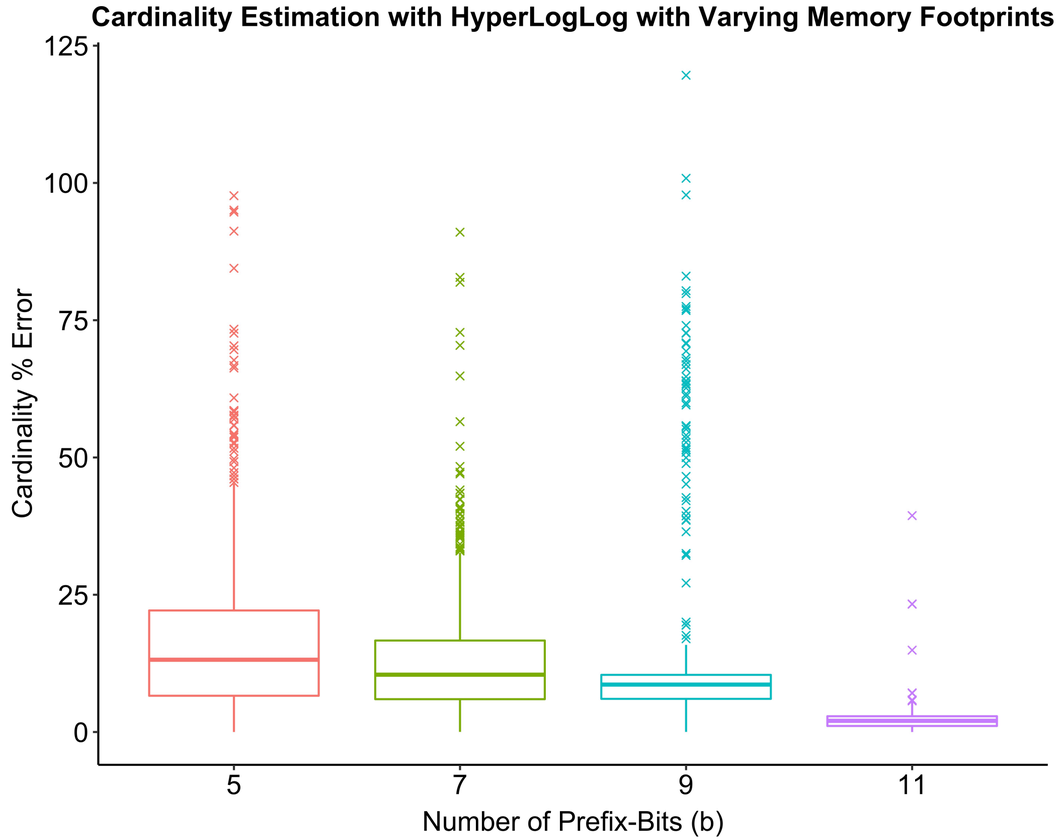




Set Cardinality Estimation with HyperLogLog

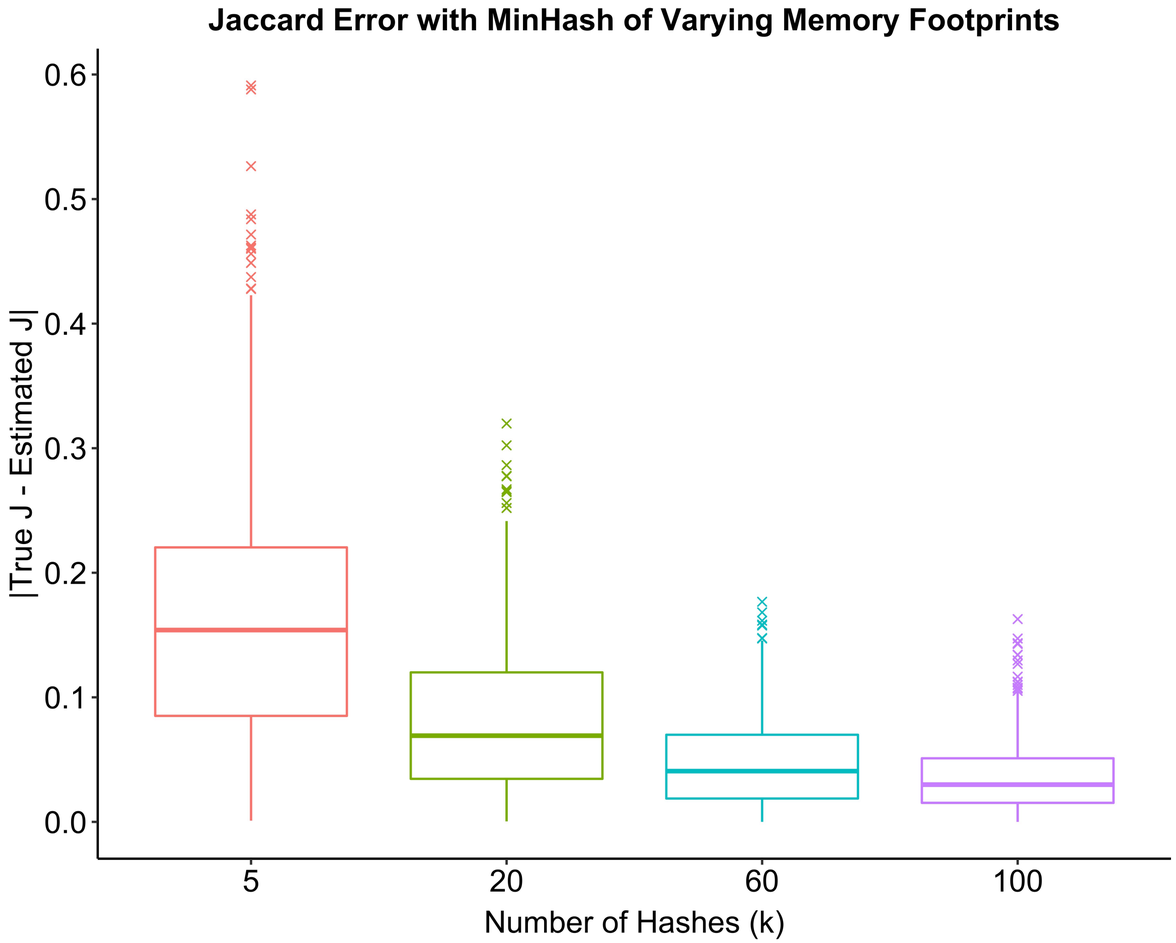
- Question: Does the implemented HyperLogLog estimate set cardinality correctly?
- Answer: It works! We see the expected pattern as b gets larger …

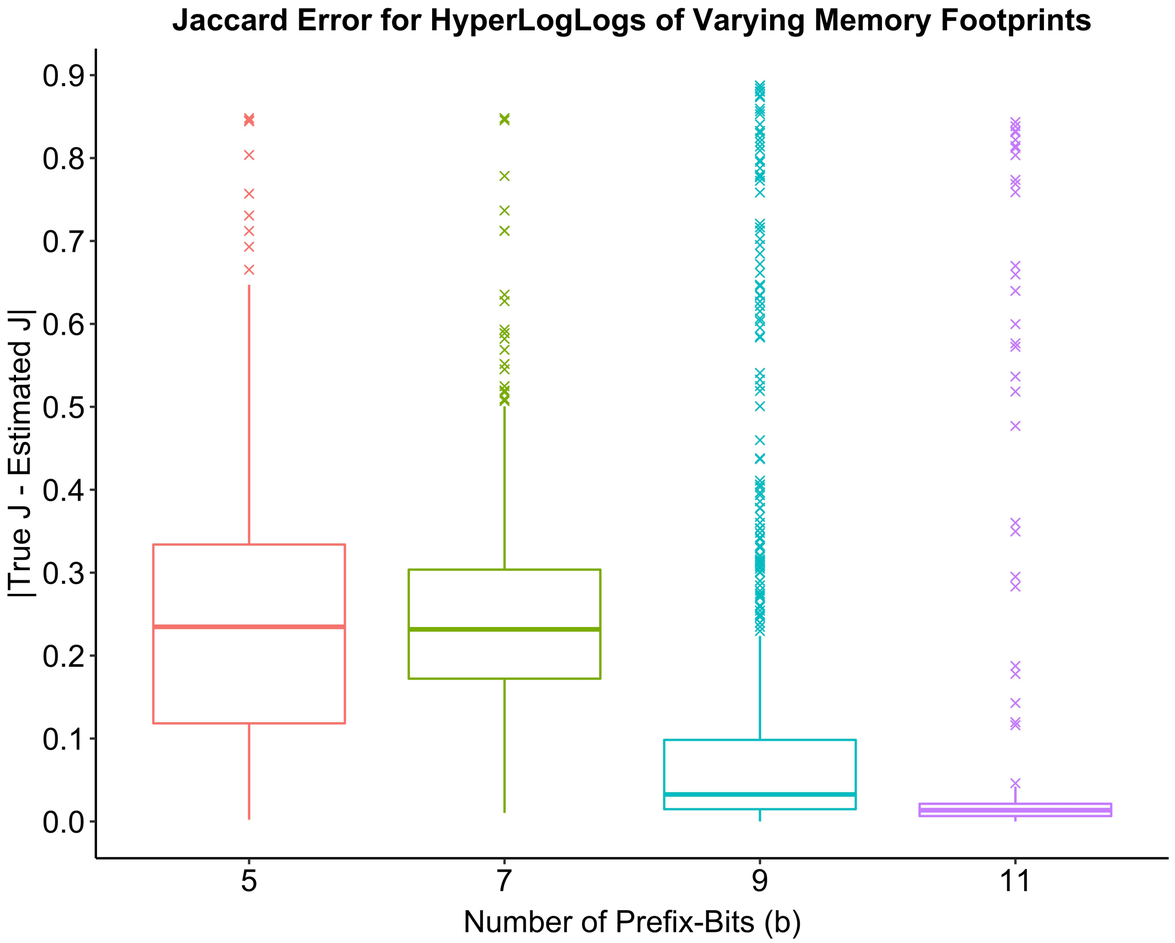




Jaccard Estimation with MinHash & HyperLogLog

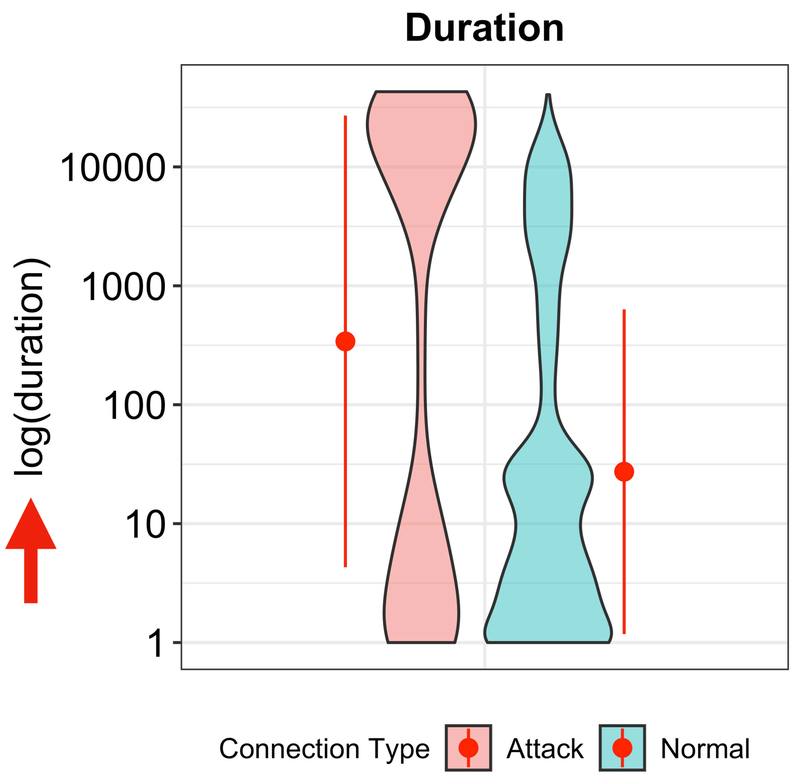
- Question: Can the two data-structures estimate the jaccard of two sets accurately?
- Answer: Yes, it seems like both can get within 0.1 of true jaccard with reasonable parameters.

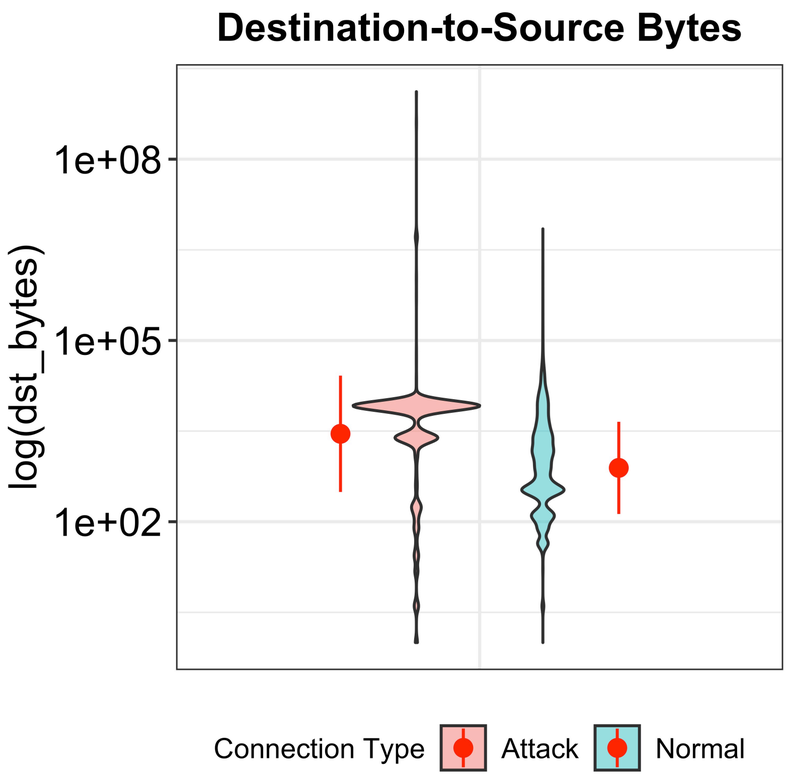


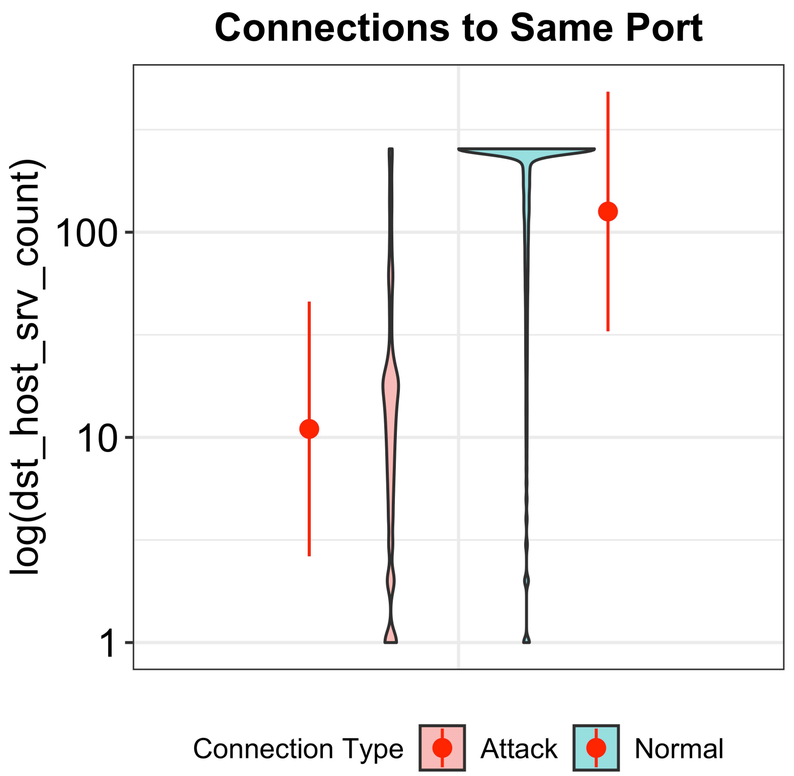


Analysis of NSL-KDD Features

- Question: Can we visually show the difference in features between attack and normal records?
- Answer: Looking at the distribution of different features, we can see that they differ …

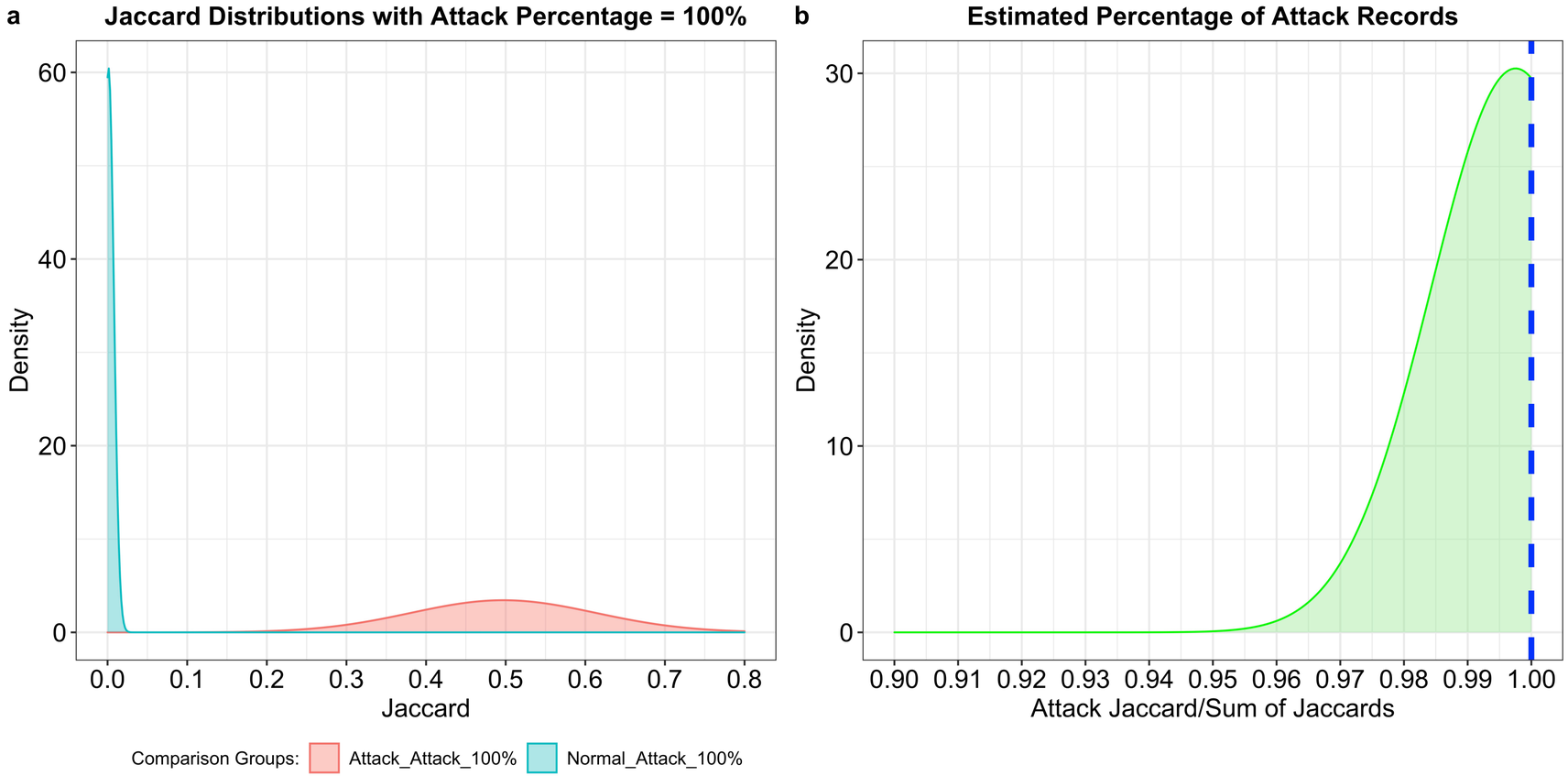






Jaccard Distributions of NSL-KDD

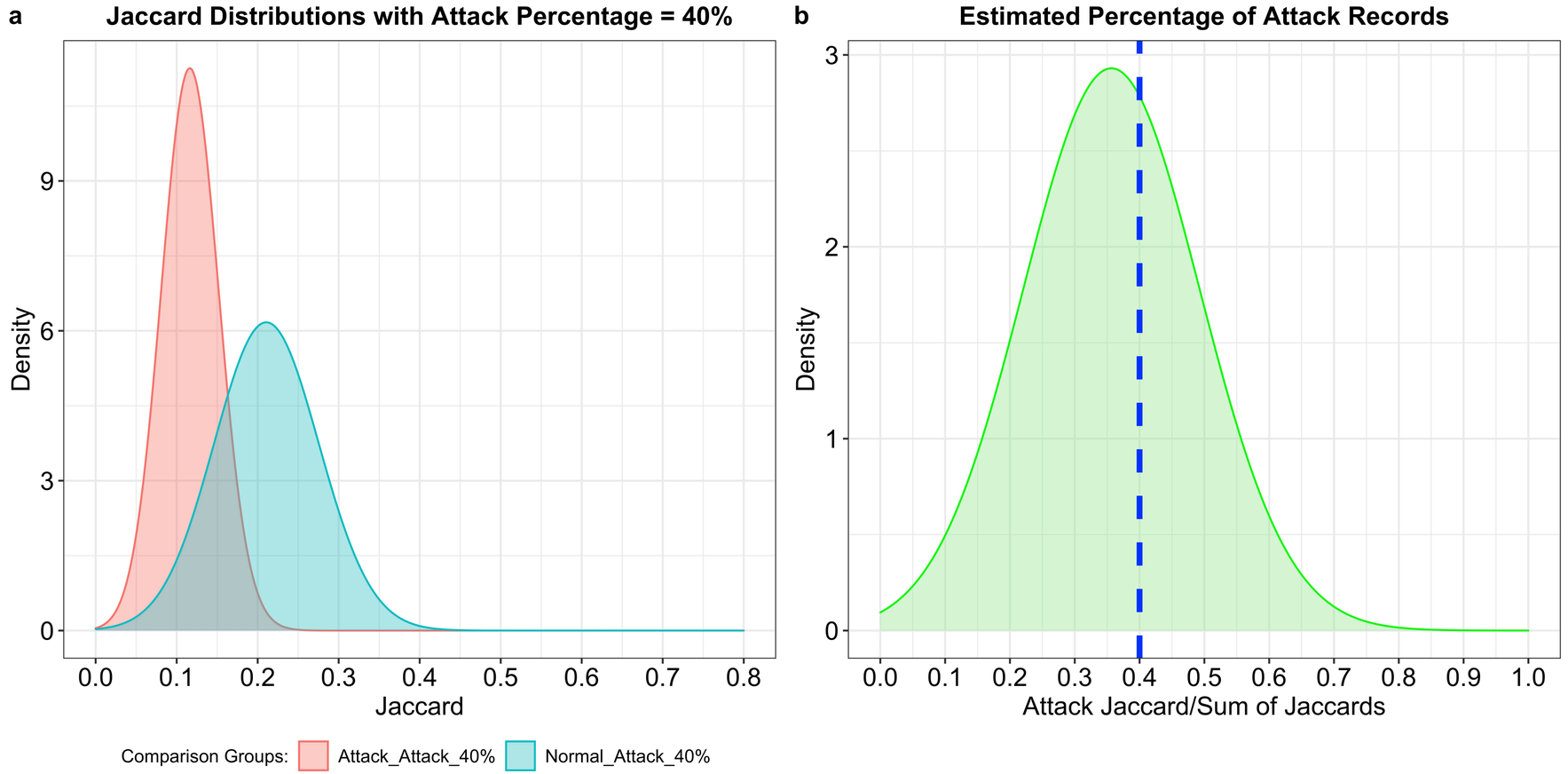
- Question: Will we see a difference when we compare windows of data to reference sketches?
  - Hypothesis: Time windows with a majority of attack records will be more similar to attack sketch
  - Each test window had 100% attack records to **make it easy at first** …
- Answer: Yes at 100%, it is easy to see most records more closely resemble attack reference.



$$percent = \frac{J_a}{J_a + J_n}$$

Jaccard Distributions of NSL-KDD

- Question: Will we see a difference when we compare windows of data to reference sketches?
  - Hypothesis: Time windows with a majority of attack records will be more similar to attack sketch
  - **Now lets try a more difficult percentage where 40% of records are security anomalies …**
- Answer: At 40%, a majority of windows are more similar to normal sketch than attack one.



$$percent = \frac{J_a}{J_a + J_n}$$

Key Takeaways & Future Work

- My project is an approach to **detecting security anomalies** found in US Air Force LAN connection data.
  - Used cardinality estimation data-structures like MinHash and HyperLogLog
  - Built sketches over windows of data, and compared them to reference sketches
  - Based on initial results, it seems like our approach can potentially do a good job of binary classifying windows as anomalous or not, and estimating anomalous %

- There were a couple of areas I wanted to finish before the end of semester …
  - Compute classification metrics like accuracy to see just how well we can do
  - Compare results to baseline approach of sampling to see if it improves upon it

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] 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.

[4] 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.

[5] Tune, P., & Veitch, D. (2011, April). Sampling vs sketching: An information theoretic comparison. In *2011 Proceedings IEEE INFOCOM* (pp. 2105-2113). IEEE.

[6] NSL-KDD dataset. Available at https://www.unb.ca/cic/datasets/nsl.html