LSH Ensemble: Internet-Scale Domain Search

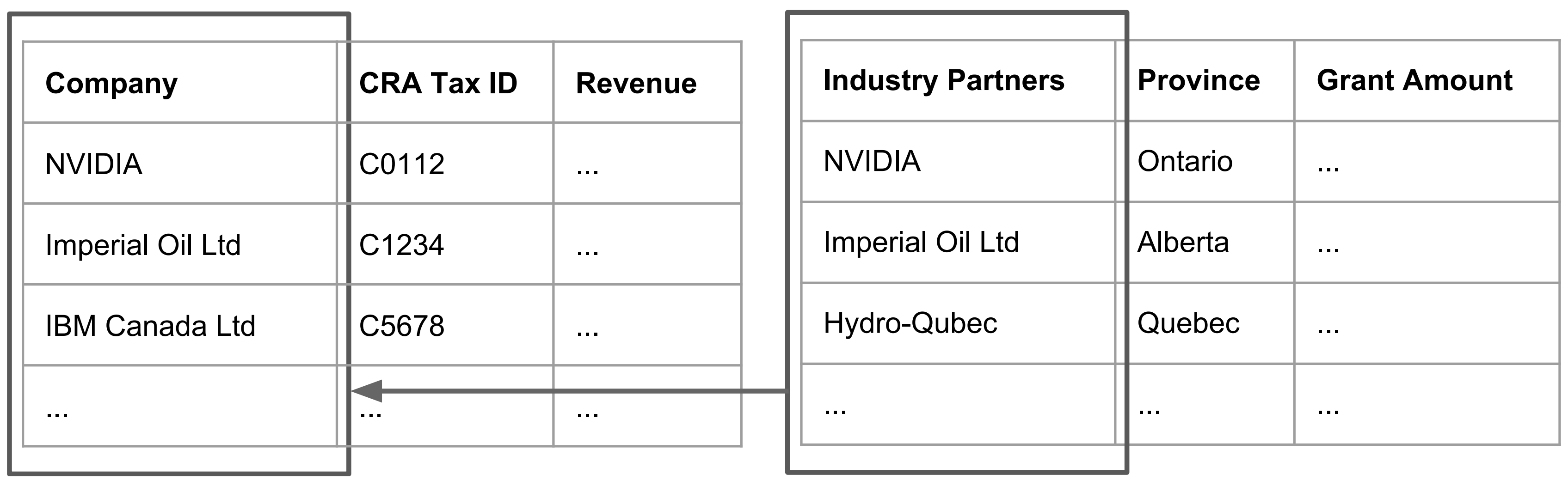
Erkang (Eric) Zhu[⋆] Fatemeh Nargesian[⋆] Ken Q. Pu[†] Renée J. Miller[⋆]

ekzhu@cs.toronto.edu fnargesian@cs.toronto.edu ken.pu@uoit.ca miller@cs.toronto.edu

[⋆]University of Toronto [†]University of Ontario Institute of Technology

1. Motivation

- **Domain search:** A domain is a set of values in a dataset (e.g. a column). We use containment score, $|Q \cap X|/|Q|$, as the measure of relevance of a domain X to a query domain Q.
- **Application:** Domain search is often essential to finding joinable datasets. As the landscape of Open Data and Web Data is growing fast, a domain search index can be useful to data scientists.



Company	CRA Tax ID	Revenue
NVIDIA	C0112	...
Imperial Oil Ltd	C1234	...
IBM Canada Ltd	C5678	...
...	...	...

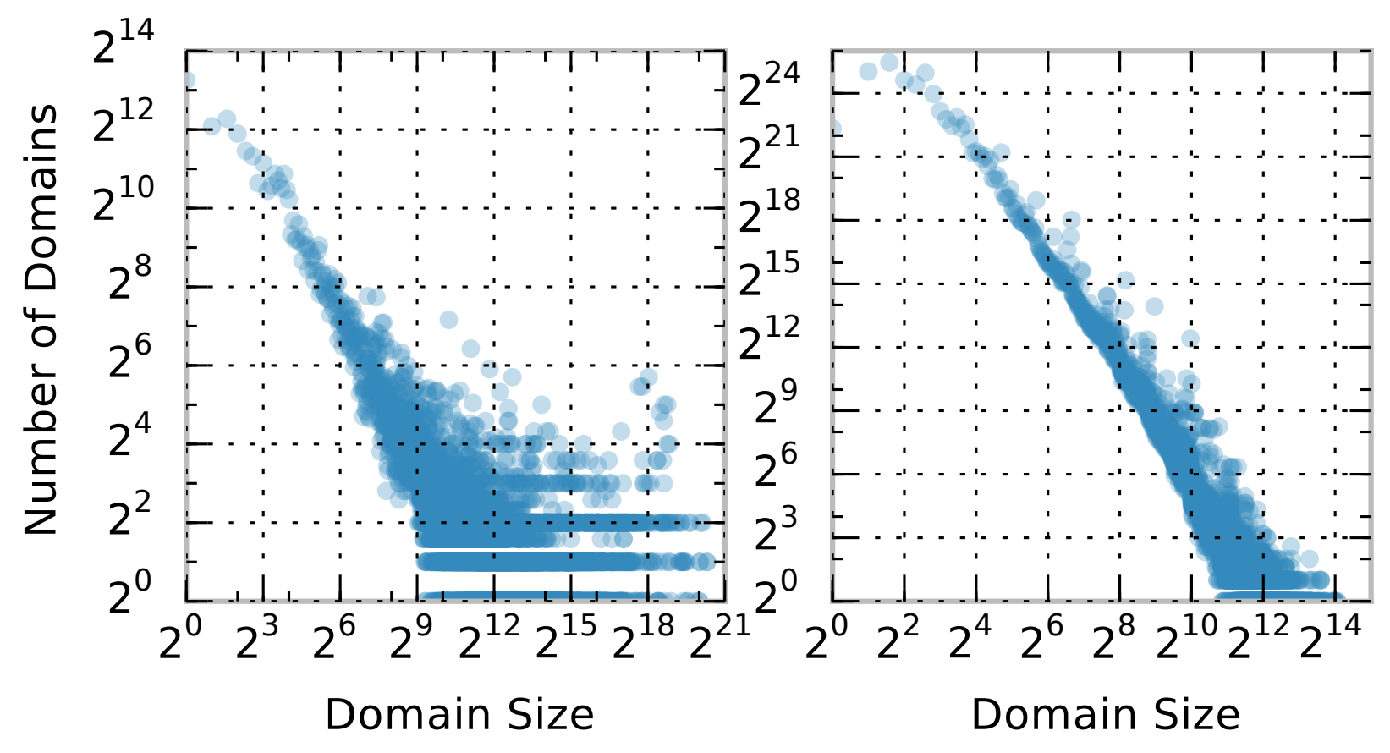
Industry Partners	Province	Grant Amount
NVIDIA	Ontario	...
Imperial Oil Ltd	Alberta	...
Hydro-Qubec	Quebec	...
...	...	...

2. Major Challenges

- **Massive scale:** For example, 263 million domains from attributes in 2015 English Relational WDC Web Table [1].
- **Small memory:** Index must be compact to handle hundreds of millions of domains, and search query must have small memory footprint as it needs to be exchanged over the Web.
- **Open world domains:** A fixed-vocabulary cannot be assumed – unseen values may occur in query/new domain.

MinHash LSH [3] can handle the above challenges, however, it does not support containment search natively.

- **Skewed distribution:** Existing approaches such as Asymmetric MinHash [2] require unbounded memory to maintain accuracy in the face of highly skewed domain size distribution (e.g., power-law), which often occurs in human-generated data.



Canadian Open Data (Left) and WDC Web Table (Right)

3. LSH Ensemble - Containment Search

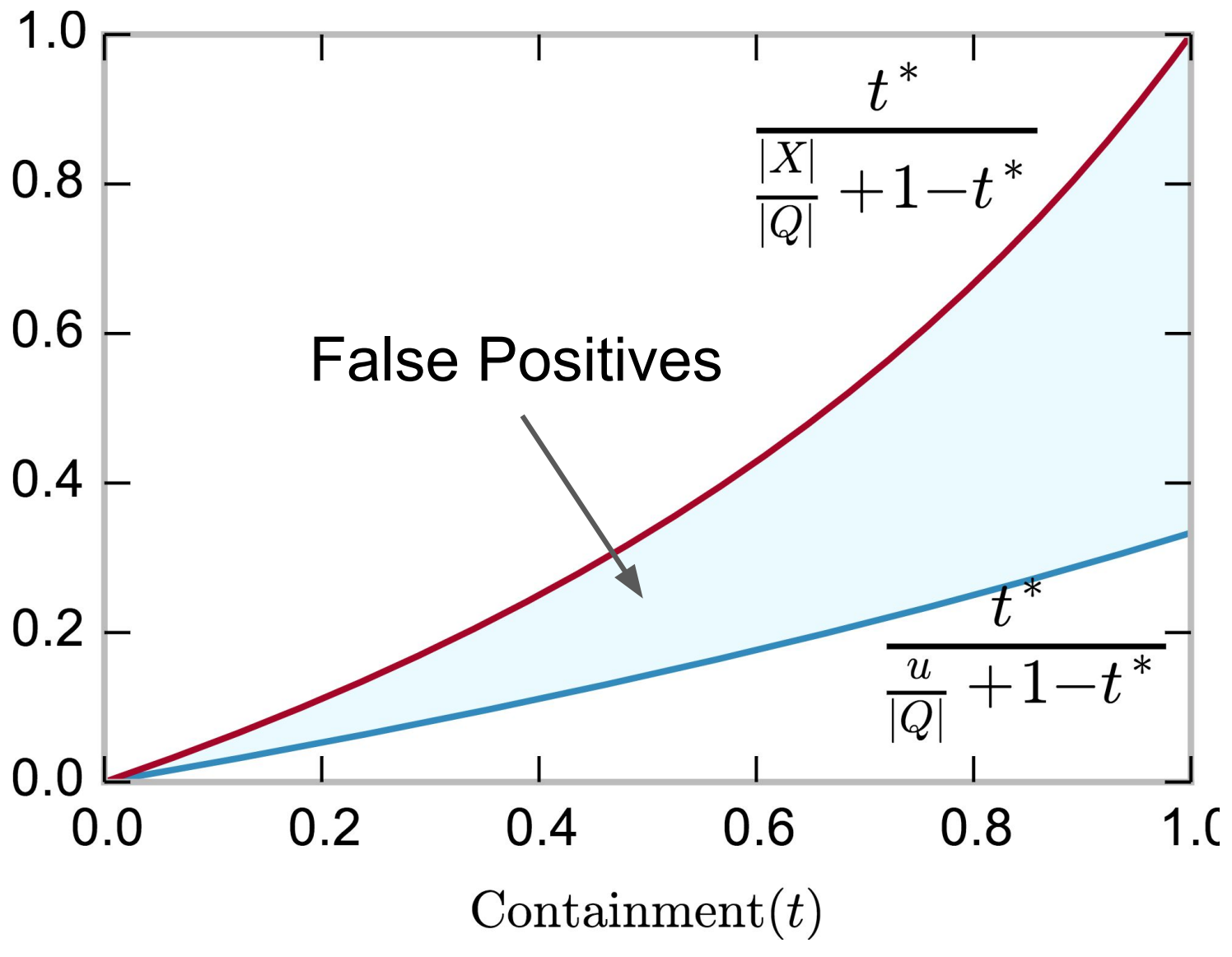
LSH Ensemble solves the domain search problem with a **containment threshold**: Given a collection of domains $\mathcal{D}$, a query domain Q, and a threshold $t^* \in [0,1]$ on the containment score, find a set of relevant domains from $\mathcal{D}$ defined as

$$\{X : \text{Containment}(Q,X) \geq t^*, X \in \mathcal{D}\} \qquad (1)$$

Containment can be converted to Jaccard using the following equation:

$$\text{Jaccard}(Q,X) = \frac{\text{Containment}(Q,X)}{\frac{|X|}{|Q|} + 1 - \text{Containment}(Q,X)} \qquad (2)$$

Given domain sizes in the range $[l,u]$, we use the new Jaccard threshold $s^* = \frac{t^*}{\frac{u}{|Q|}+1-t^*}$ on a MinHash LSH index [3] to approximate containment search, creating false positive domains. Removing the false positives incurrs additional cost.



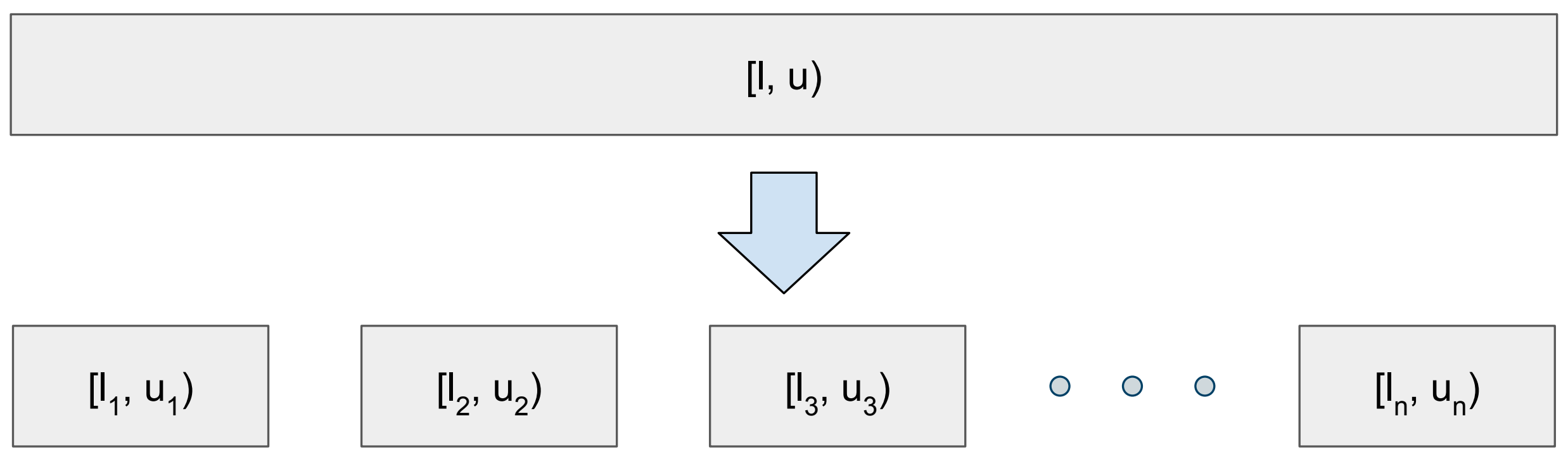
The query cost of containment search can be characterized as:

$$T_{\text{Containment}} = T_{\text{Jaccard}} + \Theta(N^C_{l,u}) + \Theta(N^{FP}_{l,u})$$

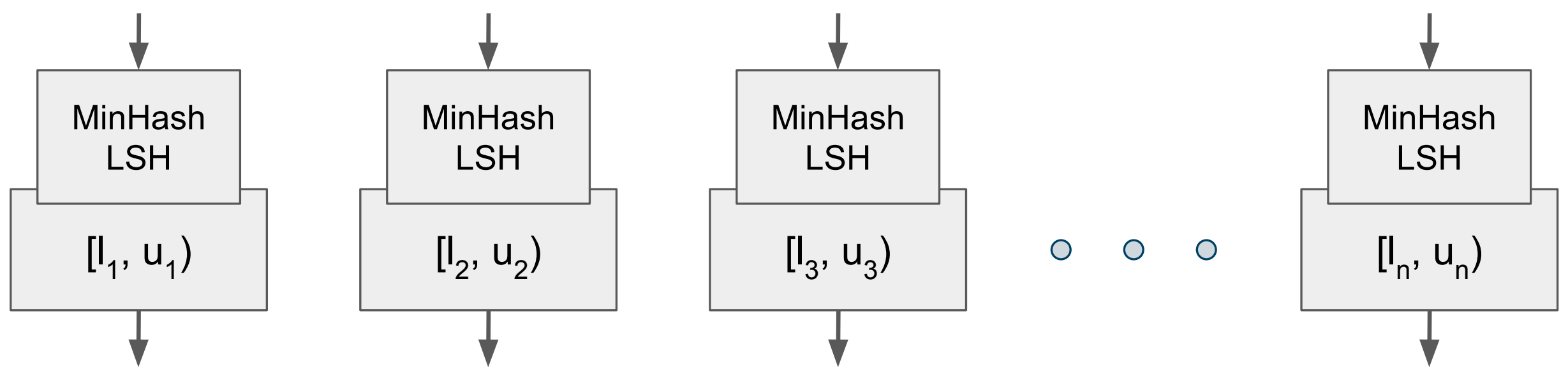
where $N^C_{l,u}$ is the number of correct domains and $N^{FP}_{l,u}$ is the number of false positive domains.

4. LSH Ensemble - Partitioning

Query cost can be reduced by reducing the number of false positives, which is bounded by $N^{FP}_{l,u} \leq N_{l,u} \cdot \frac{u-l+1}{2u}$. We achive this by **domain partitioning**.



By executing query in parallel, query cost is determined by the partition with the most false positive domains.



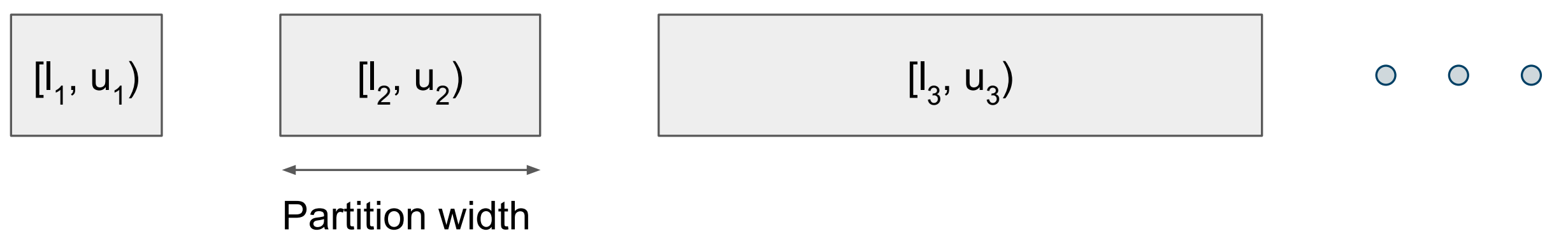
This led us to formulate an optimization problem for partitioning using the upper bound of $N^{FP}_{l,u}$ on each partition.

$$\Pi^* = \arg\min_{\Pi} \max_{1\leq i \leq n} M_i, \; M_i = N_{l_i,u_i} \cdot \frac{u-l+1}{2u}$$

This is equivalent to finding a partitioning such that all partitions have the same M_i.

$$\exists \Pi^* \; s.t. \; M_i = M_j, \; \forall i,j$$

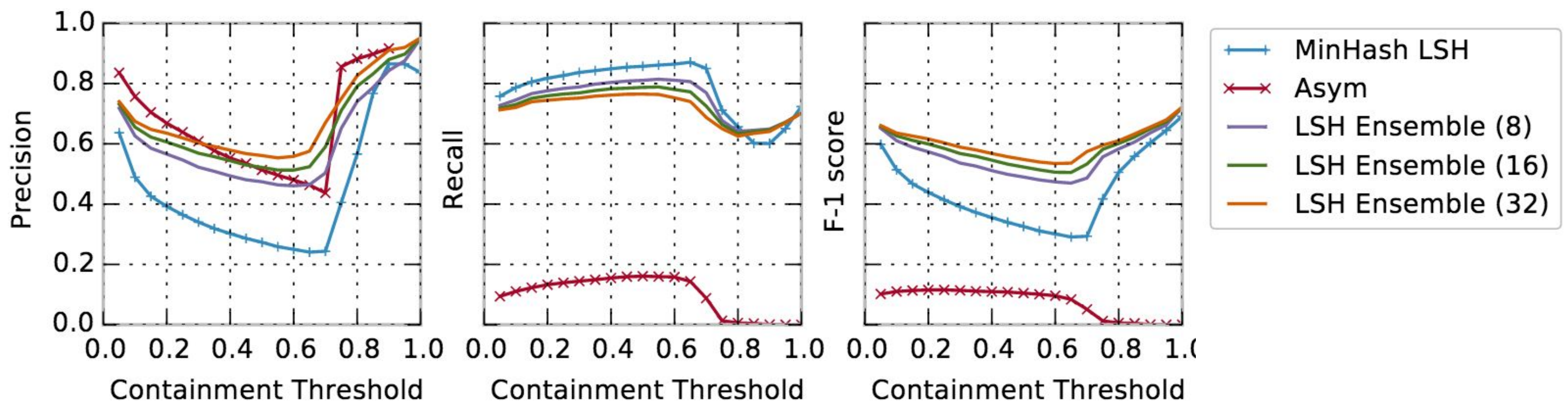
We proved that **equi-depth** partitioning is optimal for domains following **power-law** distribution.

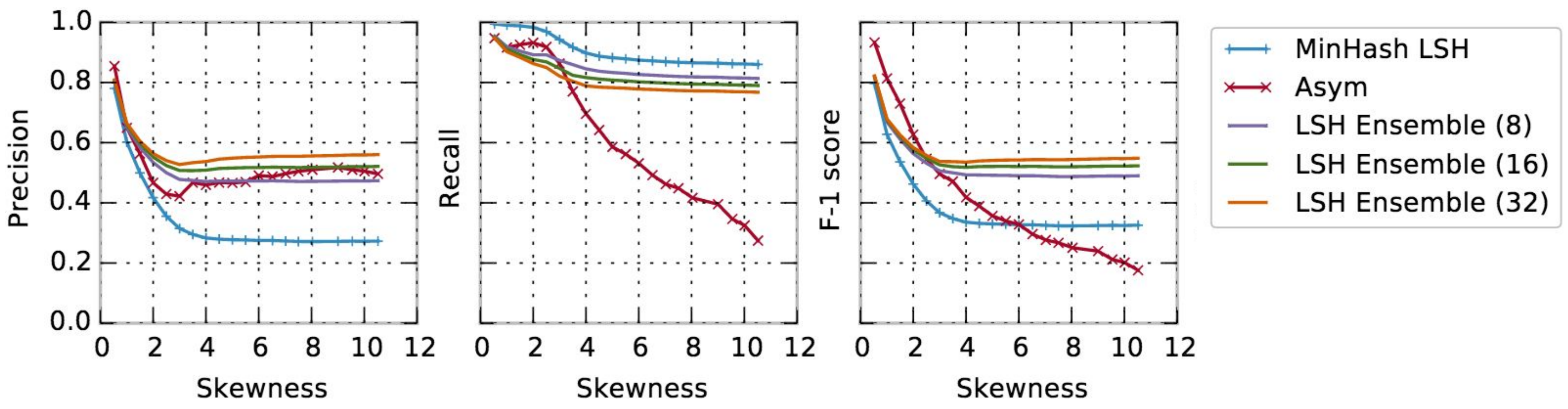


We also proposed a query-time tuning strategy for MinHash LSH given any containment threshold. See the paper for details.

5. Evaluation

Compared against Asym MinHash [2] and MinHash LSH [3] on accuracy, using Canadian Open Data (65,533 domains).





Scalability experiment used the complete 2015 English Relational WDC Web Table [1] (263 million domains), and $t^* = 0.5$.

	Mean Query (sec)	Precision Before Pruning
MinHash LSH	45.13	0.27
LSH Ensemble (8)	7.55	0.48
LSH Ensemble (16)	4.26	0.53
LSH Ensemble (32)	3.12	0.58

References

[1] O. Lehmberg, D. Ritze, R. Meusel, and C. Bizer. A large public corpus of web tables containing time and context metadata. In WWW, 2016.

[2] A. Shrivastava and P. Li. Asymmetric minwise hashing for indexing binary inner products and set containment. In WWW, pages 981-991, 2015.

[3] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In STOC, pages 604-613, 1998.