B^{ed}-Tree: An All-Purpose Index Structure for String Similarity Search Based on Edit Distance

Zhenjie Zhang, Marios Hadjieleftheriou,
Beng Chin Ooi, Divesh Srivastava

School of Computing

at&t

Outline

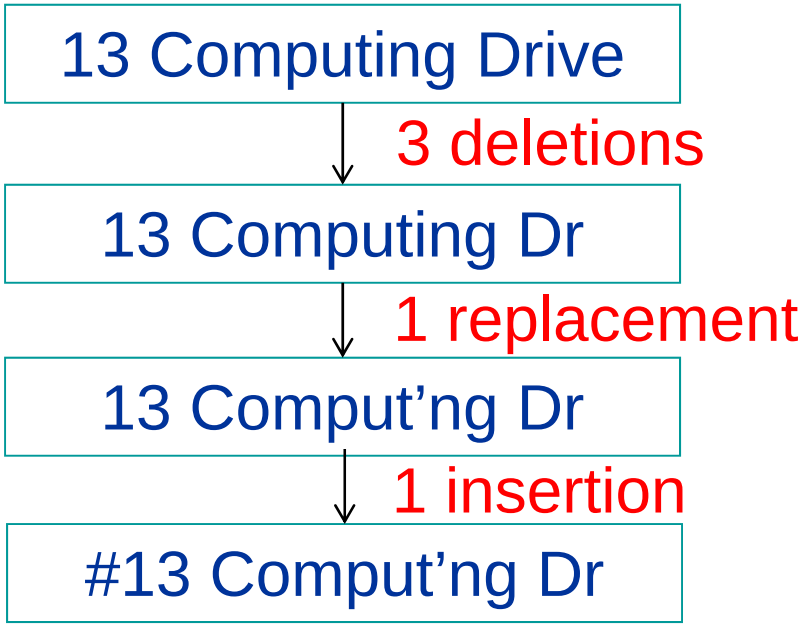
- Motivation and B^{ed}-Tree Framework
- String Orders
  - Dictionary order
  - Gram counting order
  - Gram location order
- Experiments
- Conclusion

Approximate String Search

- Information Retrieval
  - Web search query with string “Posgre SQL” instead of “Postgre SQL”
- Data Cleaning
  - “13 Computing Road” is the same as “#13 Comput’ng Rd”?
- Bioinformatics
  - Find out all protein sequences similar to “ACBCEEACCDECAAB”

Edit Distance

- Edit distance on strings



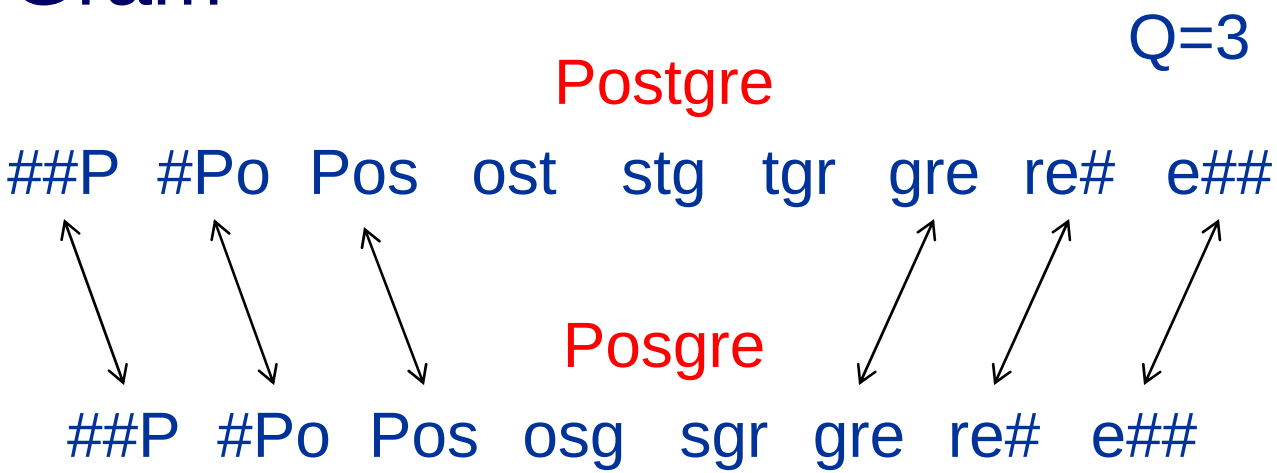
Edit distance: 5

- Normalized edit distance

$$\frac{ED(s_1,s_2)}{MaxLength(s_1,s_2)} \qquad \frac{5}{18}$$

Existing Solution

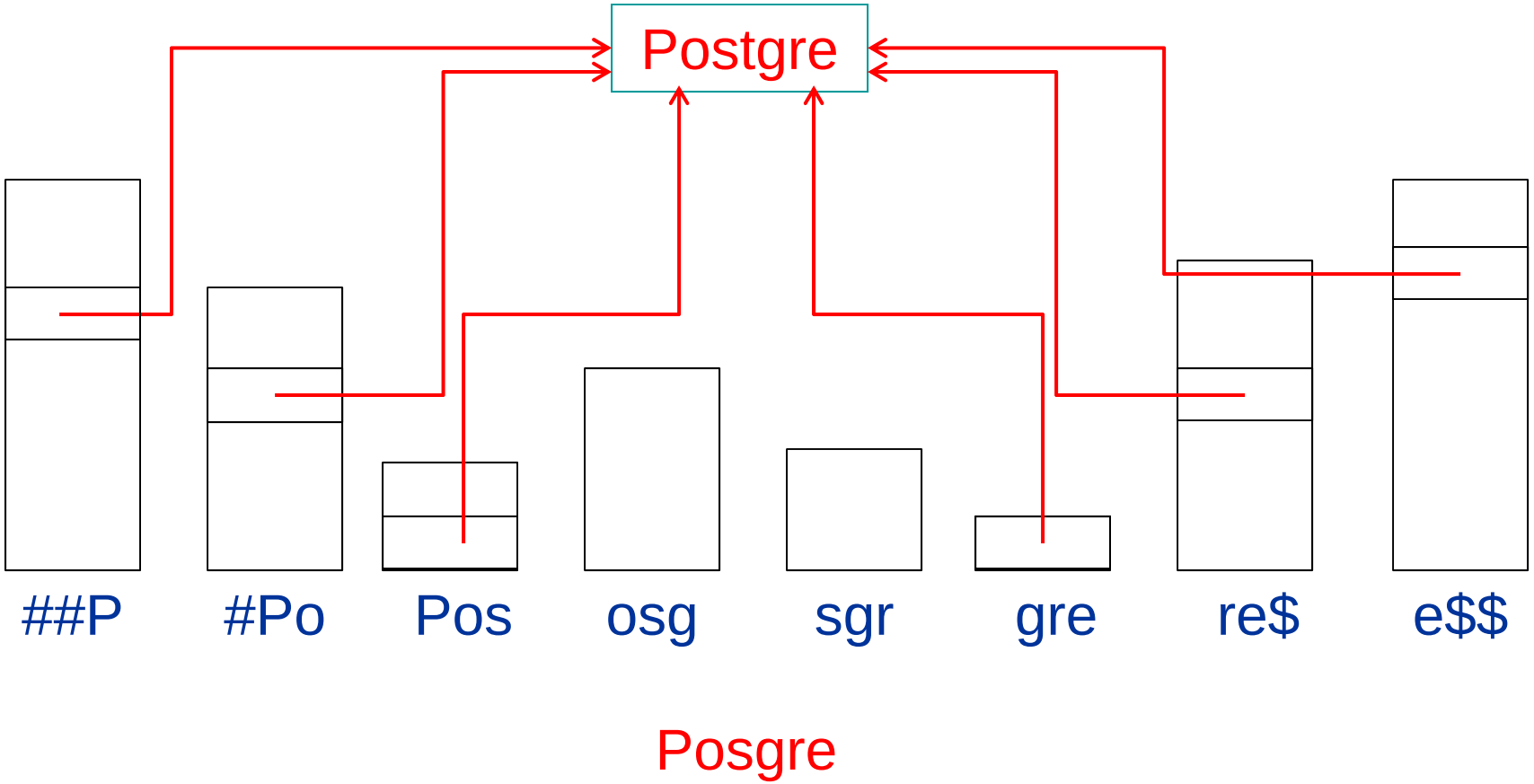
- Q-Gram



Observation: If $ED(s_1,s_2)=d$, they agree on at least $\min(|s_1|,|s_2|)+Q-1-d*(Q+1)$ grams

Existing Solution

- Inverted List



Limitations

- Inverted List Method
  - Limited queries supported

	Range Query	Join Query	Top-K Query	Top-K Join
Edit Distance	Y	Y	N	N
Normalized ED	N	N	N	N

  - Uncontrollable memory consumption
  - Concurrency protocol

Our Contributions

- B^{ed}-Tree
  - Wide support on different queries and distances

	Range Query	Join Query	Top-K Query	Top-K Join
Edit Distance	Y	Y	Y	Y
Normalized ED	Y	Y	Y	Y

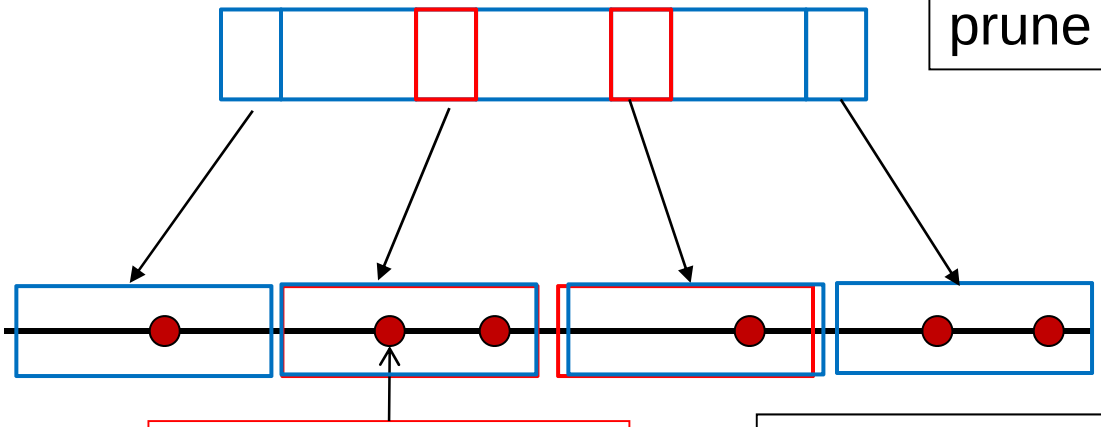
  - Adjustable buffer size and low I/O cost
  - Highly concurrent
  - Easy to implement
  - Competitive performance

Basic Index Framework

B^{ed}-Tree Framework

Index Construction follows standard B+ tree

Query: Posgre



Estimate the minimal distance to query and prune B+ tree nodes

Map all strings to a 1D domain

Result: Postgre

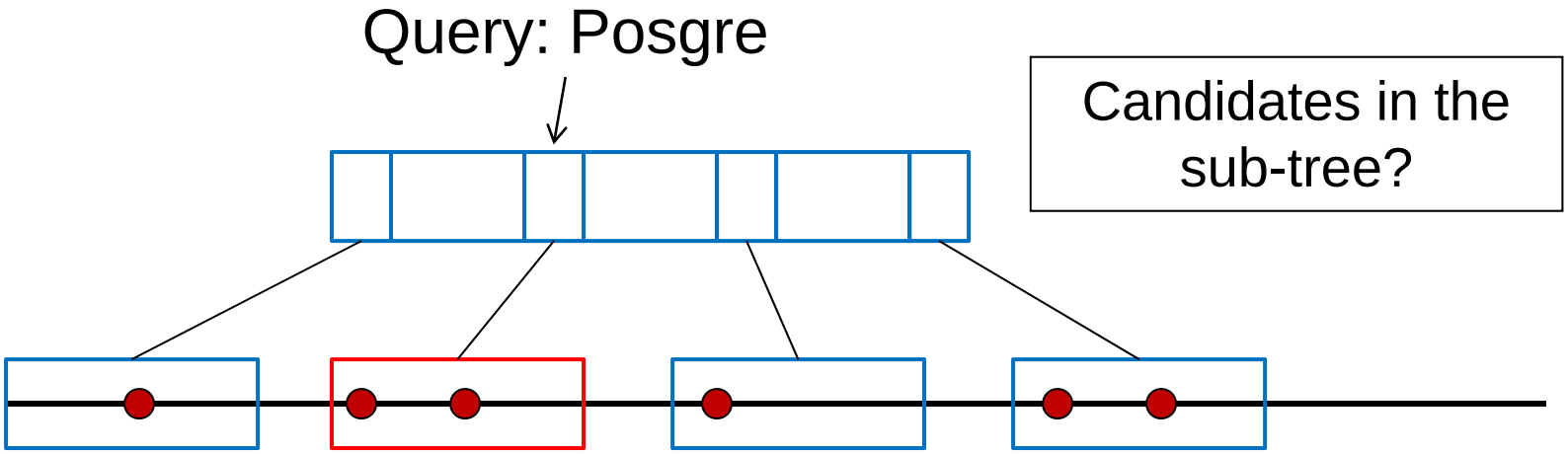
Refine the result by exact edit distance

Outline

- Motivation and B^{ed}-Tree Framework
- String Orders
  - Dictionary order
  - Gram counting order
  - Gram location order
- Experiments
- Conclusion

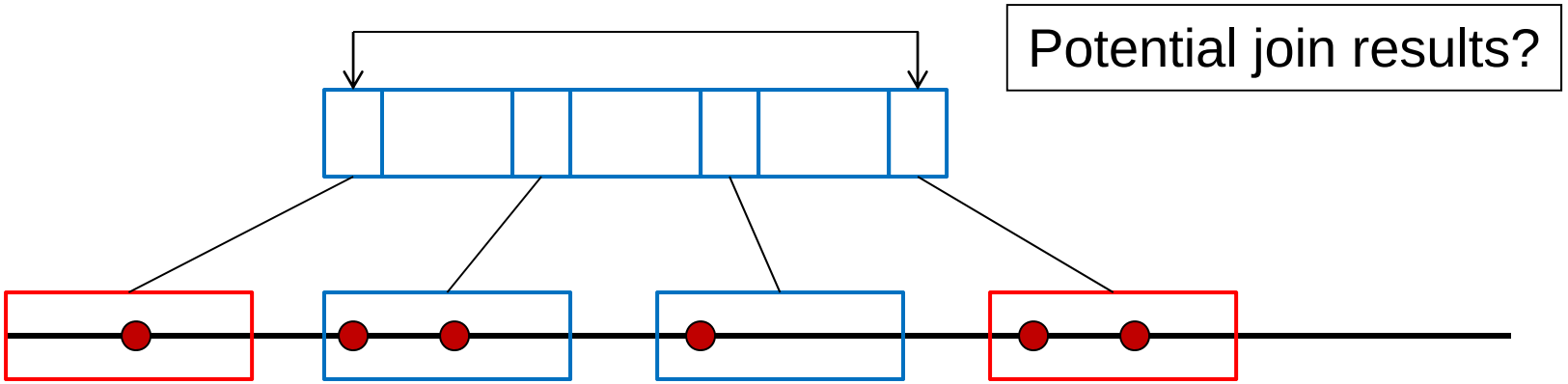
String Order Properties

- P1: Comparability
  - Given two string s1 and s2, we know the order of s1 and s2 under the specified string order
- P2: Lower Bounding
  - Given an interval [L,U] on the string order, we know a lower bound on edit distance to the query string



String Order Properties

- P3: Pairwise Lower Bounding
  - Given two intervals [L,U] and [L',U'], we know the lower bound of edit distance between s1 from [L,U] and s2 from [L',U']
- P4: Length Bounding
  - Given an interval [L,U] on the string order, we know the minimal length of the strings in the interval



String Order Properties

- Properties v.s. supported queries and distances

	Range Query	Join Query	Top-K Query	Top-K Join
Edit Distance	P1, P2	P1, P3	P1, P2	P1, P3
Normalized ED	P1, P2, P4	P1, P3, P4	P1, P2, P4	P1, P3, P4

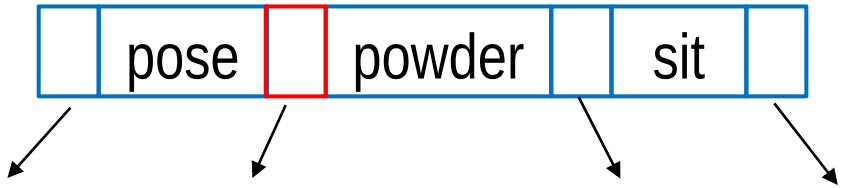
	Description
P1	Comparability
P2	Lower Bounding
P3	Pair-wise Lower Bounding
P4	Length Bounding

Dictionary Order

- All strings are ordered alphabetically, satisfying P1, P2 and P3

Search: Posgre with ED=1

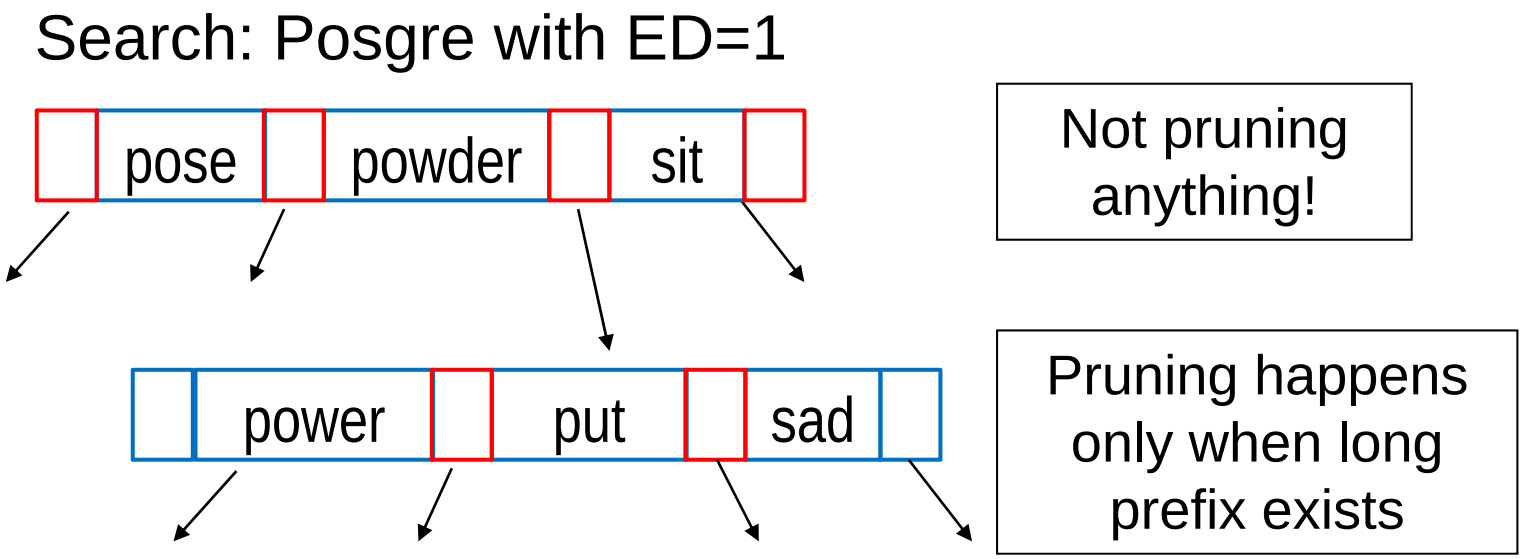
Insertion: Postgre



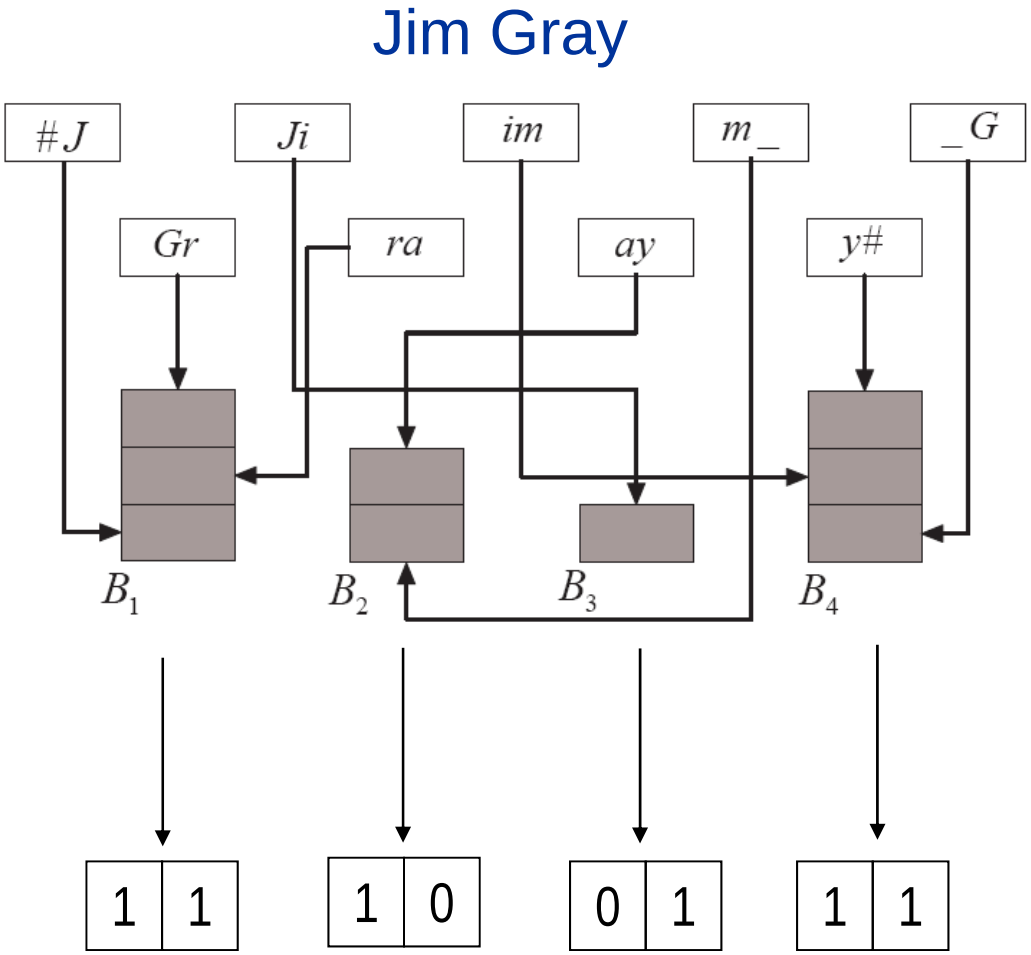
It's between "pose" and "powder"

Dictionary Order

- All strings are ordered alphabetically, satisfying P1, P2 and P3



Gram Counting Order

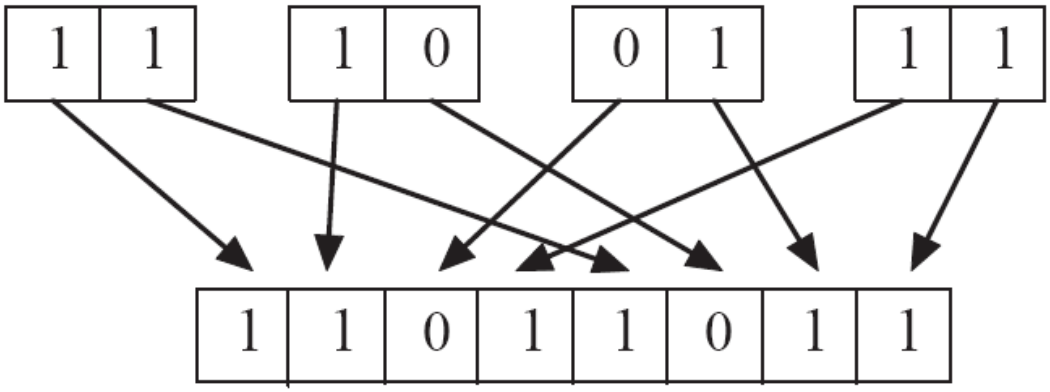


Hash all grams to 4 buckets

Count the grams in binary

Gram Counting Order

- Transform the count vector to a bit string with z-order

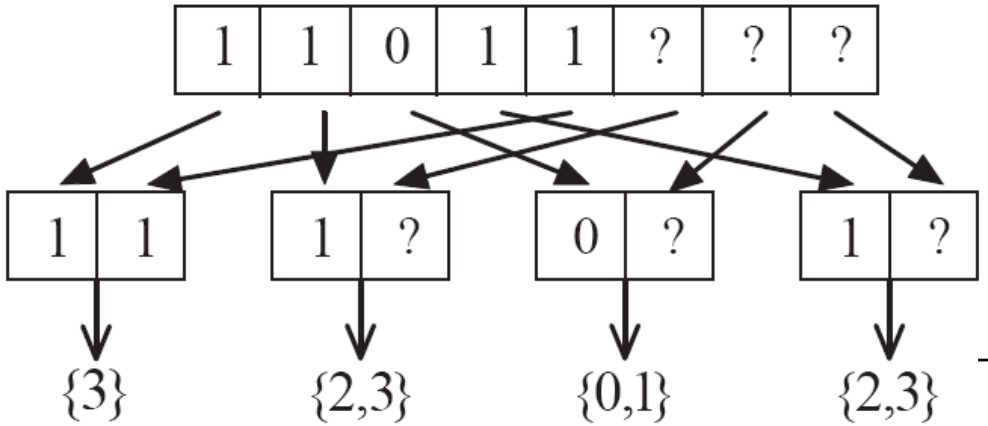


Gram Counting Order

- Lower Bounding

"11011011" to "11011101"

Prefix: "11011???"



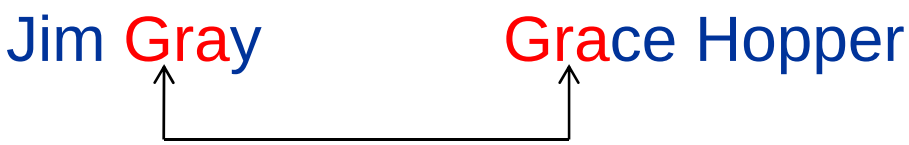
Query: Jim Gary

signature: (4,1,2,2)

Minimal edit distance: 1

Gram Location Order

- Extension of Gram Counting Order
  - Include positional information of the grams



  - Allow better estimation of mismatch grams
  - Harder to encode

Outline

- Motivation and Framework
- String Orders
  - Expected properties
  - Dictionary order
  - Gram counting order
  - Gram location order
- Experiments
- Conclusion

Experiment Settings

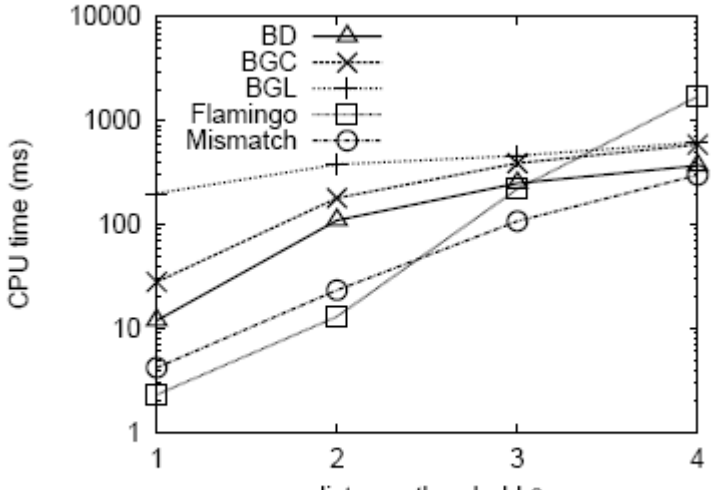
- Data

Dataset	# of Strings	Max. Length	Avg. Length
Author	2,948,929	56	22
Title	1,158,648	675	74
Actor	1,213,391	80	23
Movie	1,568,893	247	26
Protein	508,038	1,999	347

- Five Index Schemes
  - B^{ed}-Tree: BD, BGC, BGL
  - Inverted List: Flamingo, Mismatch
- Default Setting
  - Q=2, Bucket=4, Page Size=4KB

Empirical Observations

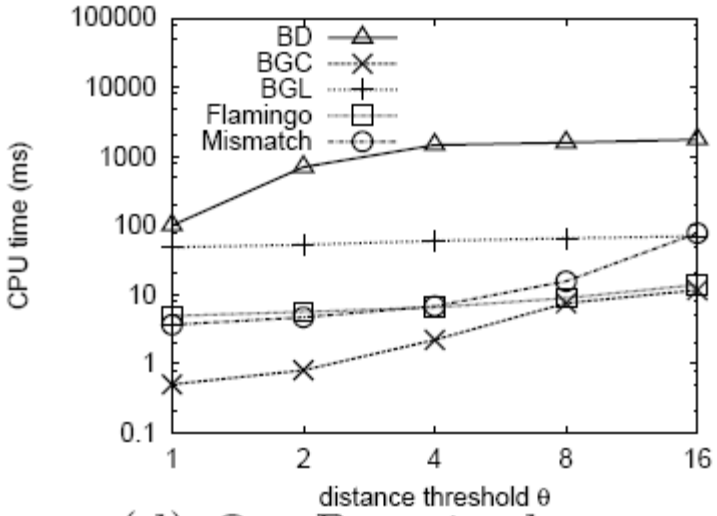
- How good is B^{ed}-Tree?
  - With small threshold, Inverted Lists are better
  - When threshold increases, B^{ed}-Tree is not worse



(a) On *Author* data

Empirical Observations

- Which string order is better?
  - Gram counting order is generally better
  - Gram Location order: tradeoff between gram content information and position information



(d) On *Protein* data

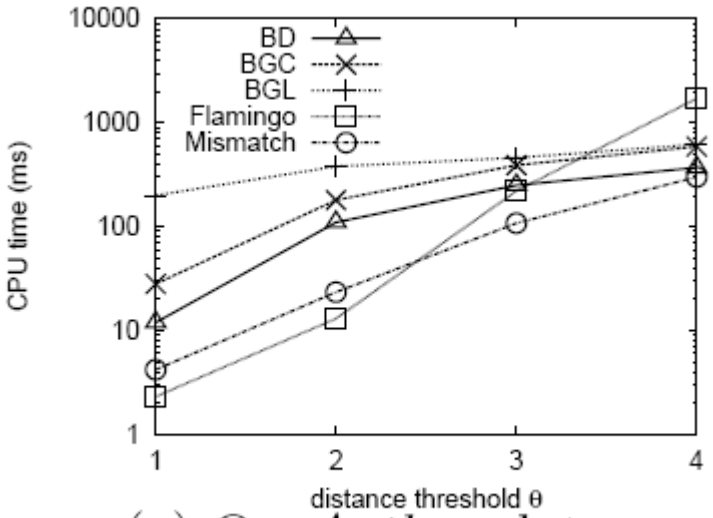
Conclusion

- A new B+ tree index scheme
  - All similarity queries supported
  - Both edit distance and normalized distance
  - General transaction and concurrency protocol
  - competitive efficiencies

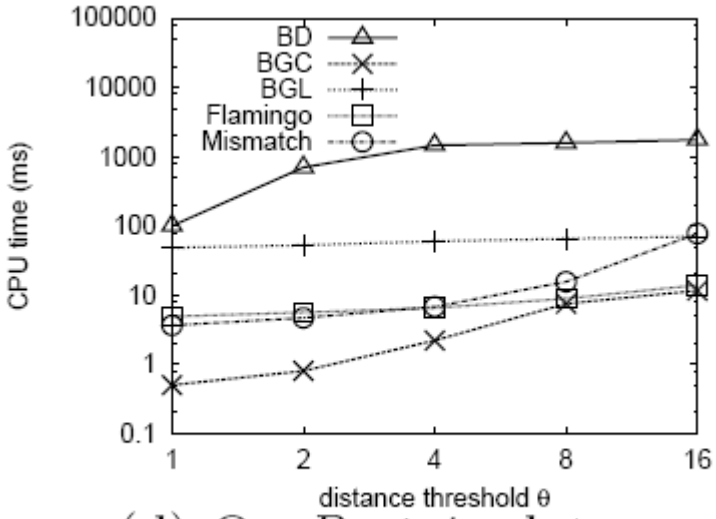
Q&A

Results

Range Query



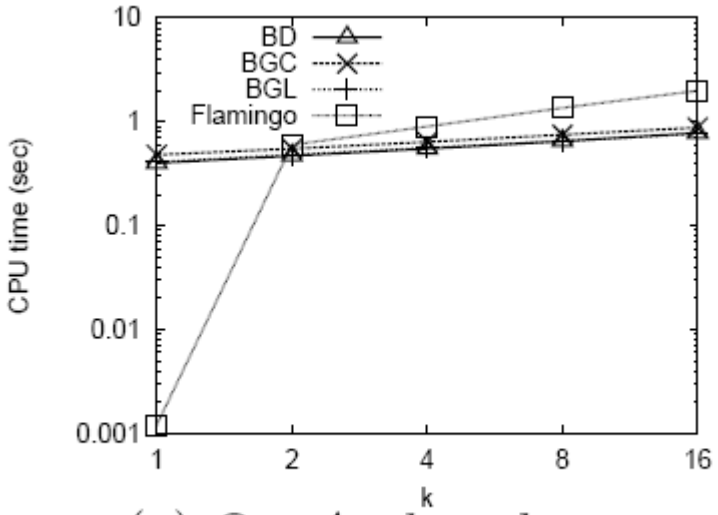
(a) On *Author* data



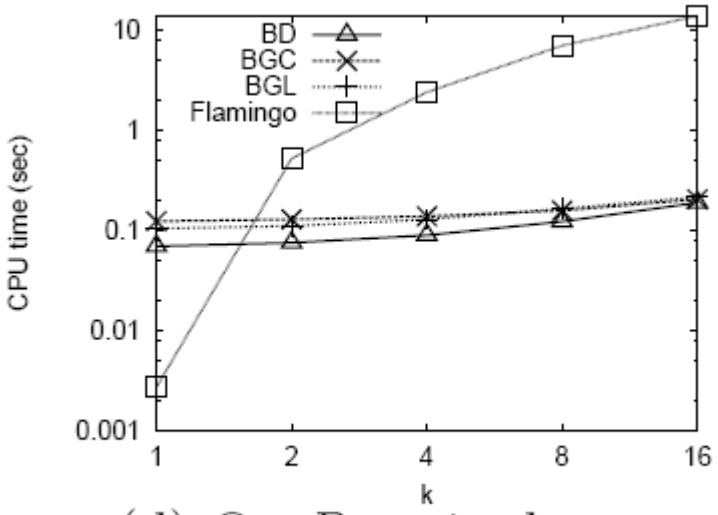
(d) On *Protein* data

Results

- Top-K Query



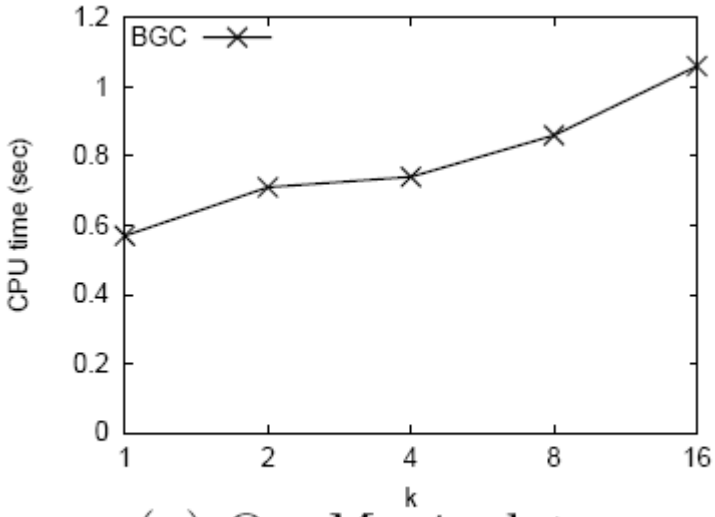
(a) On *Author* data



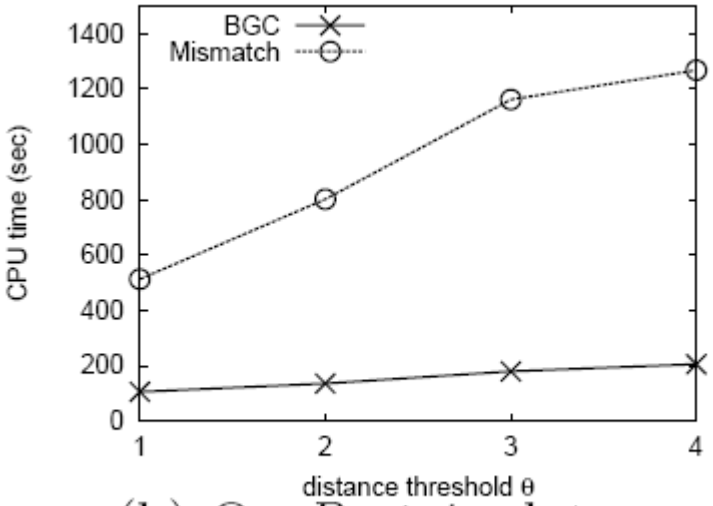
(d) On *Protein* data

Results

- Normalized Edit Distance & Join Query



(a) On *Movie* data



(b) On *Protein* data