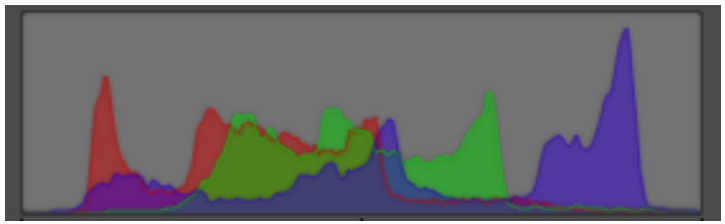
# Melody-Join: Efficient Earth Mover's Distance Similarity Joins Using MapReduce

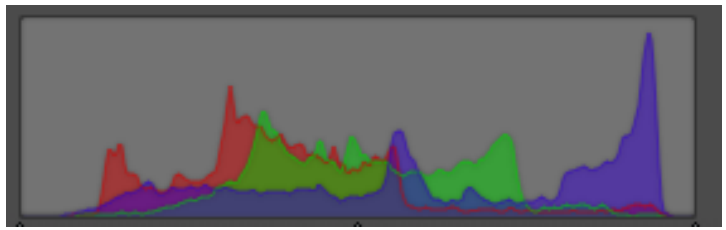
Jin Huang [†], Rui Zhang [†], Jian Chen [‡], Rajkumar Buyya [†]

[†] *Department of Computing and Information Systems*
*The University of Melbourne, Australia*

[‡] *School of Software Engineering*
*South China University of Technology, China*

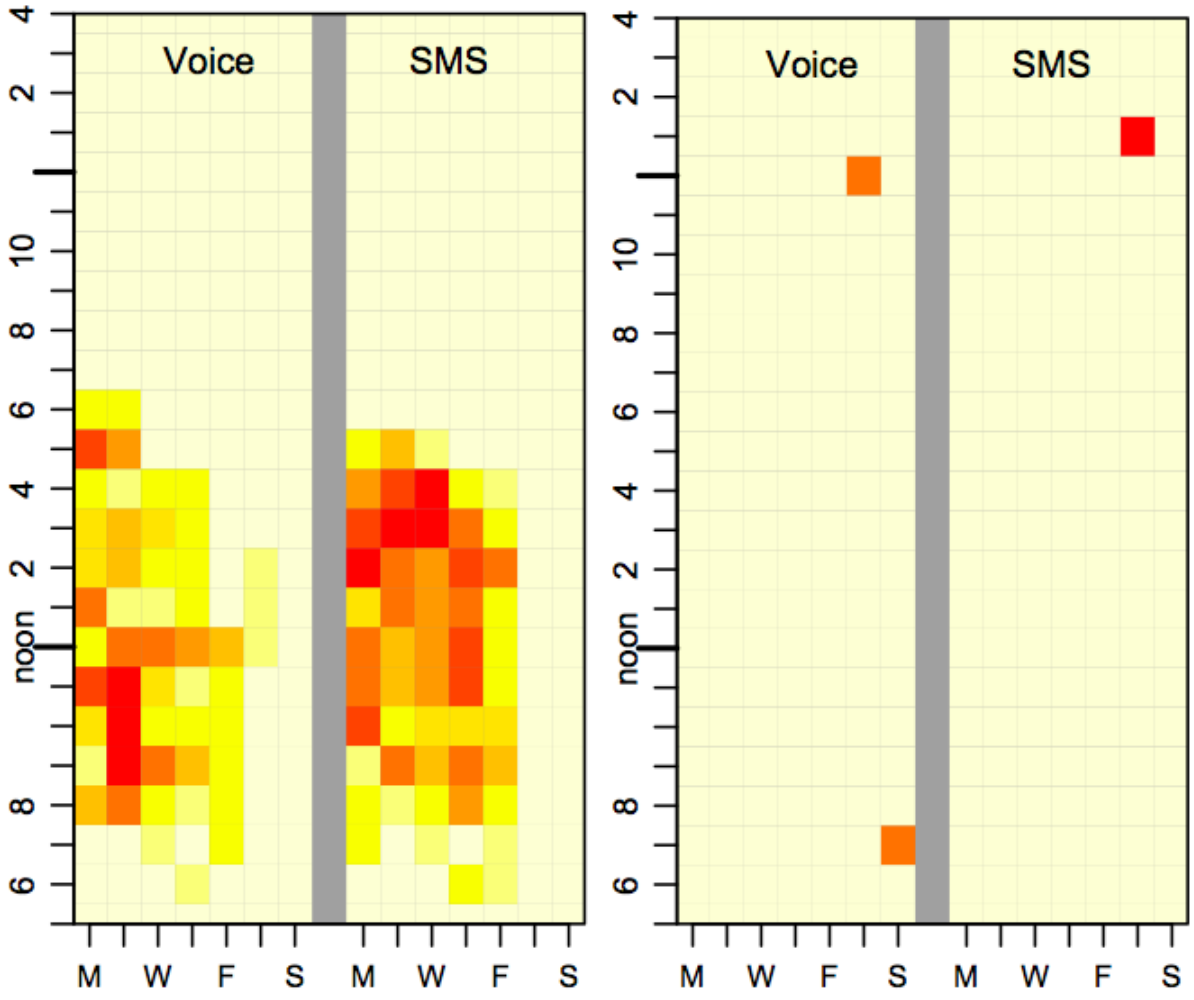
# Motivation I: Similar Image Detection





# Motivation II: Stock Distribution Analysis


# Motivation III: Usage Pattern Analysis

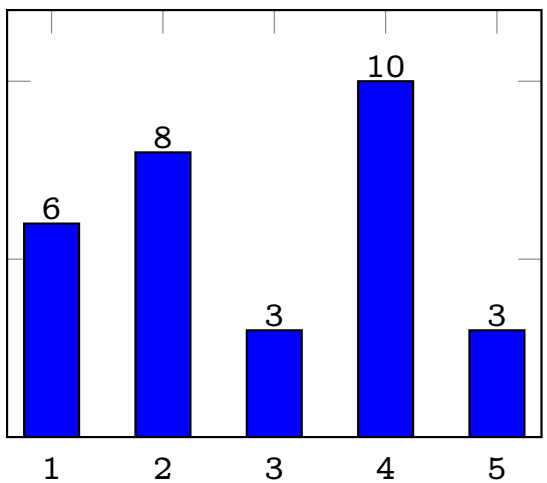



D. Applegate, T. Dasu, S. Krishnan, and S. Urbanek. Unsupervised Clustering of Multidimensional Distributions using Earth Mover Distance. KDD'11

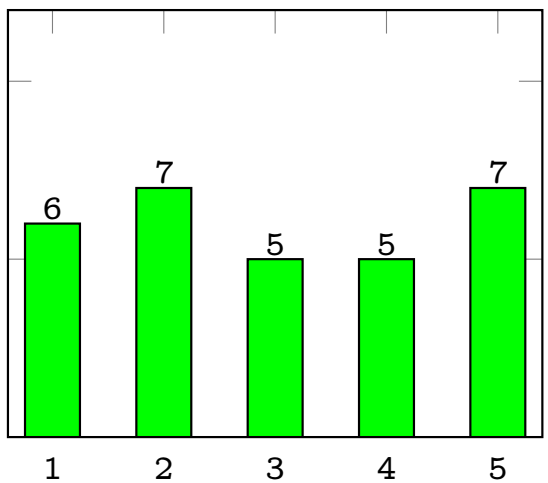
# Preliminaries: Earth Mover's Distance [Rubner et al. 2000]

Given two histograms with equal weights, the **minimum** cost of transforming one to the other

- cost $= \sum$ amount transformed $\times$ distance transformed
- $\text{EMD}(\text{A}, \text{B}) = 1 \times dist(2,3) + 1 \times dist(3,4) + 4 \times dist(4,5)$




(a) $\text{A} = \{6, 8, 3, 10, 3\}$




(b) $\text{B} = \{6, 7, 5, 5, 7\}$

- we consider $\ell_2$ for $dist(bin, bin)$

# Preliminaries: Problem Definition and Observation

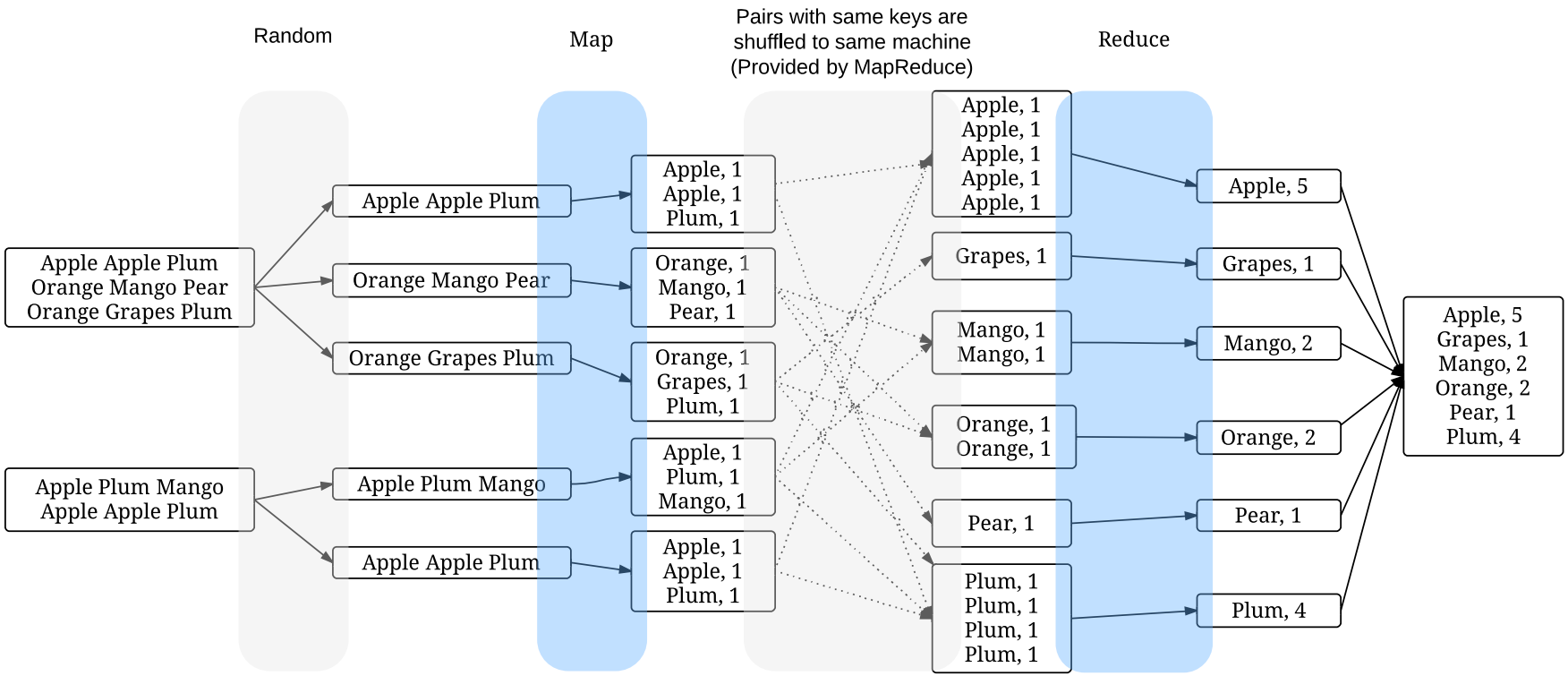
- **EMD based Similarity Join** Given two histogram datasets $H_R$ and $H_S$ and a EMD threshold $\epsilon$, the join returns $\{(h_R, h_S) | EMD(h_R, h_S) \leq \epsilon, h_R \in H_R, h_S \in H_S\}$

- Highly computation-intensive
  - Linear programming optimization in each distance
  - Simplex method: $O(n^3 \log n)$
  - EMD (n = 32) vs. $\ell_2$ : 50 ms vs. 0.002 ms
  - Join $O(H_R \times H_S)$

- We propose to use MapReduce to join large amount of data

# Preliminaries: MapReduce (MR)

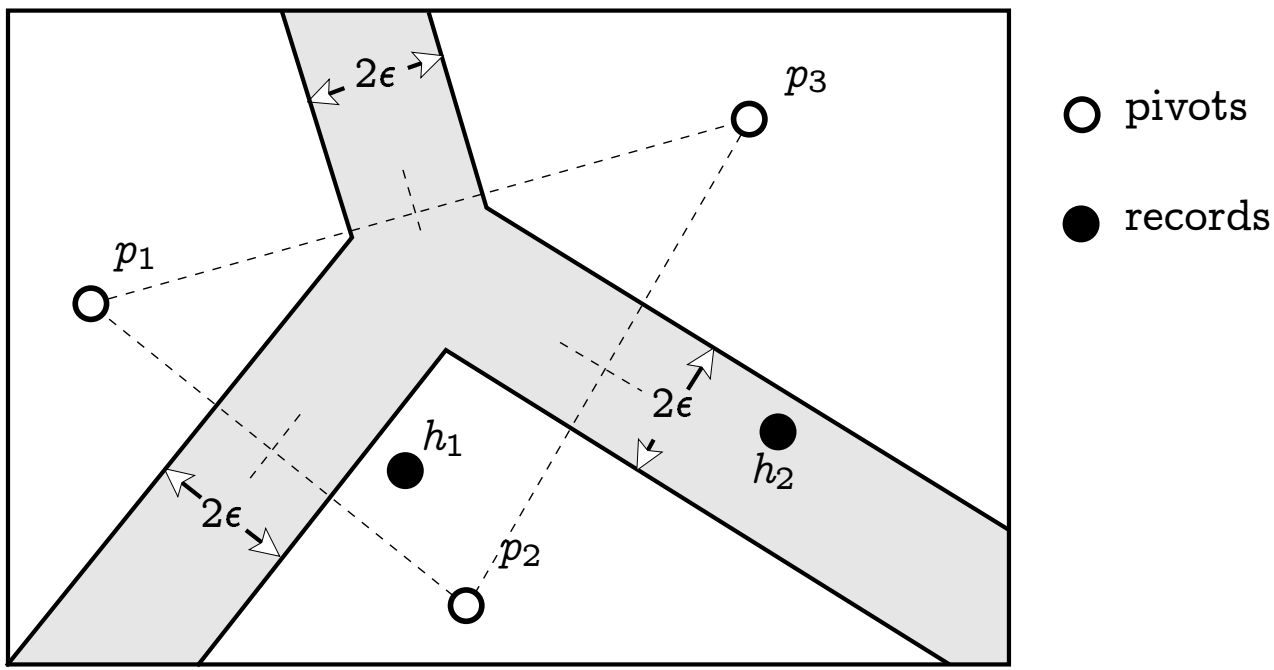
## Example: Counting word frequency in text files



# Preliminaries: State-of-the-Art MRSimJoin [Silva et al. 2012]

MRSimJoin:

- Intensive distance computation in data partitioning
- Vulnerability towards skewed datasets

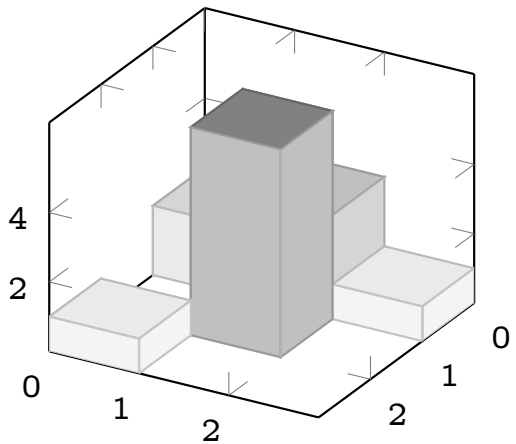


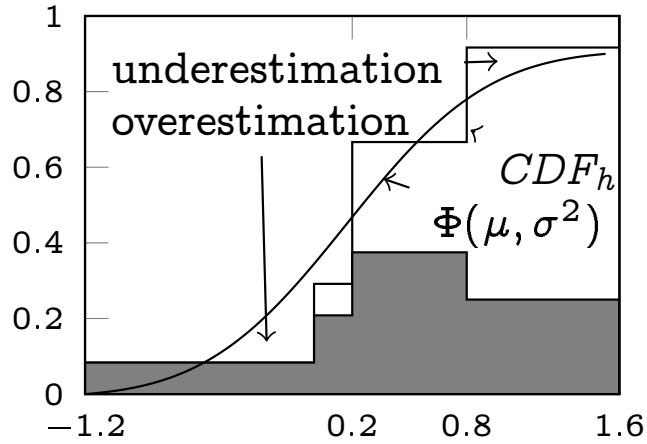
# Preliminaries: Normal Lower Bounds [Ruttenberg and Singh 2012]

1. Project high-D histograms to 1D histograms
2. Approx. 1D Cumulative Distribution Function (CDF) with normal CDF

$$EMD(A,B) \geq \left| \int CDF_A - \int CDF_B \right|$$

$$\geq \left| \int \Phi(\mu,\sigma^2)_A - \int \Phi(\mu,\sigma^2)_B + error_A - error_B \right|$$




3. Hough transform normal CDF for record-group LB

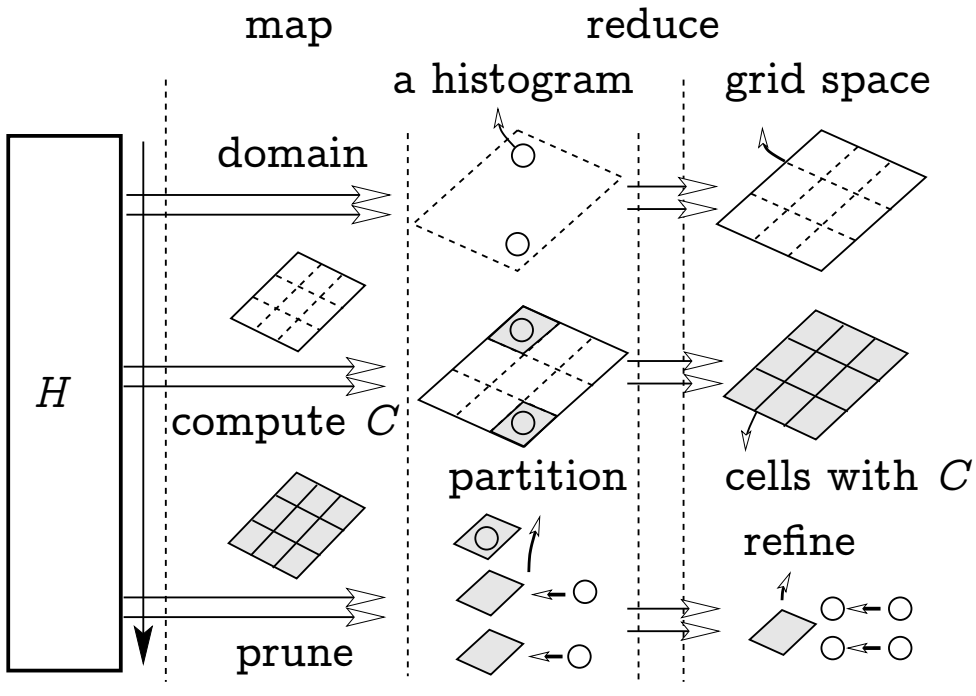
# Melody-Join: *Mapreduce Earth mover's distance Lower bOund baseD similaritY Join*

- Avoid EMD: use lower bounds to prune and partition data
- Three MR jobs to implement the idea with grids

Job 1 Obtain Hough space domain and grid

Job 2 Compute the approximation errors for cells

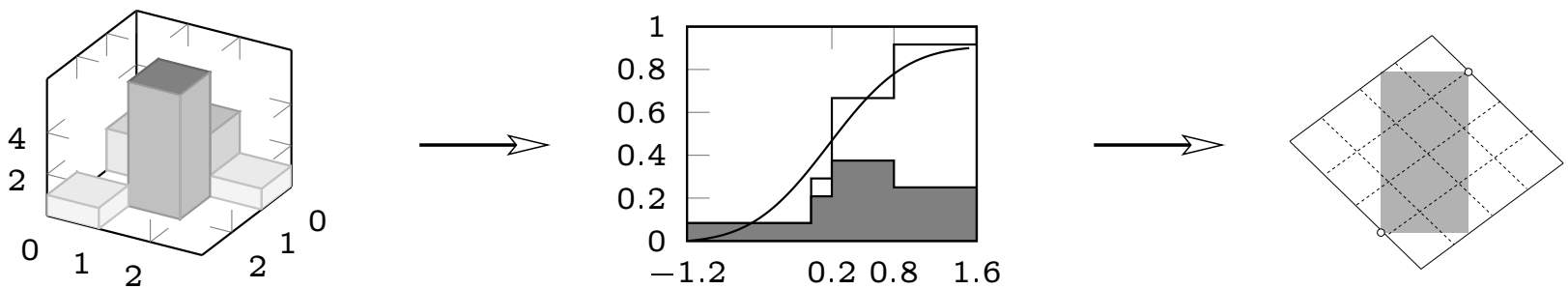
Job 3 Prune, partition, and refine records



# Melody-Join: Job 1

## Obtain space and grid for $LB_{normal}$

- Map:
  - Project high-D histogram to 1D histogram
  - Approximate 1D CDF with normal CDF
  - Transform normal CDF $\Phi(\mu, \sigma^2)$ to record $(\frac{1}{\sigma}, \frac{-\mu}{\sigma})$ in Hough space
- Reduce: Obtain domain of the Hough space and divide space into (diamond-shape) grid



# Melody-Join: Job 2

Aggregate approximation errors of cells for $LB_{normal}$

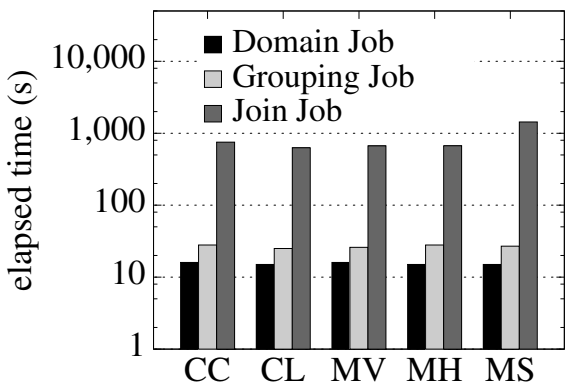
- Map:
  - Compute containing cell for each histogram
  - Distribute histograms to their containing cells
- Reduce:
  - Aggregate the approximation errors for each cell
  - Count the histograms in each cell

- Similar to counting word frequency
  - Each line: each histogram
  - Word frequency: errors and counts of cell

# Melody-Join: Job 3

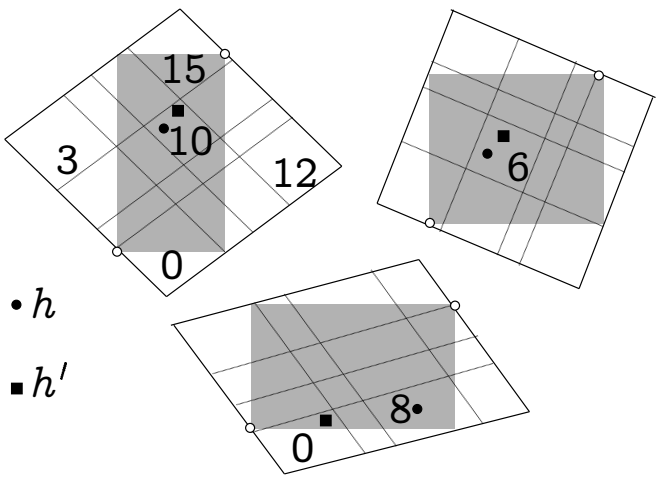
## Prune, partition, and refine pairs

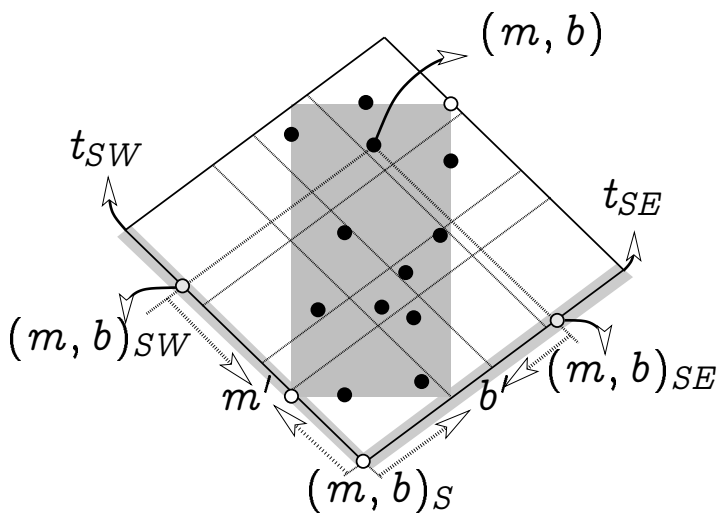
- Map:
  - Prune cells from a histogram $h$ if $LB_{normal}(h, cell) > \epsilon$
  - Distribute histogram to containing cell and unpruned cells
- Reduce: refine pairs of histograms by computing EMD
- This job dominates the running time



# Enhancing: Separating Histograms

- Employ multiple projection vectors for multiple spaces
  - Compose cell keys: $G_h = \{10, 6, 8\}$, i.e., $h$ is in the cell 10 in space 1, cell 6 in space 2, and cell 8 in space 3
  - Prune the cell from a histogram if any $LB_{normal} > \epsilon$
- A quantile based grid to handle skewed data
  - Compute quantile values in Job 1 in addition to the domain

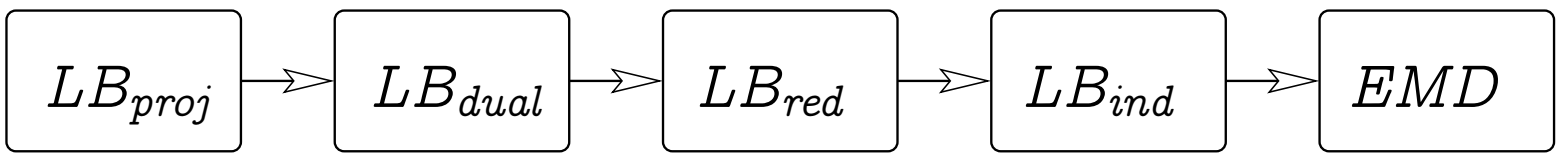


# Enhancing: Multiple Types LB

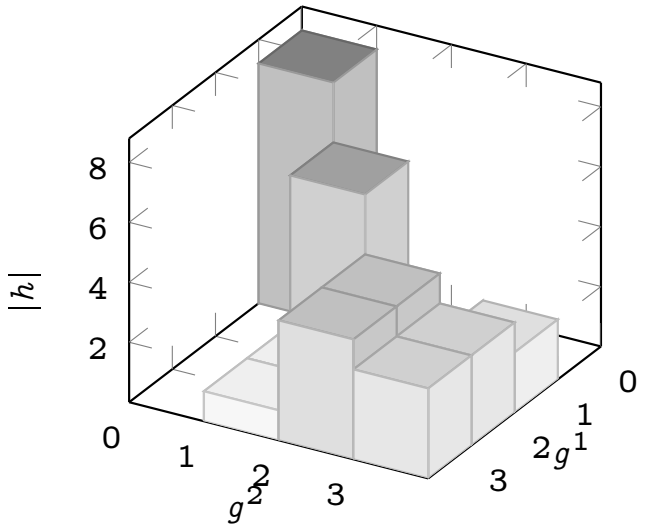
- Plug in pruning: LBs that support record-group computation, e.g., the Dual Lower Bound [Xu et al. 2012]
  - Compute dual keys for histograms in Job 1 and Job 2
  - Compute $LB_{dual}$ in Job 3 Map in addition to $LB_{normal}$, i.e., $LB_{dual}(h, cell) \geq \epsilon$ or $LB_{normal}(h, cell) \geq \epsilon$ lead pruning

- Plug in refining: Chain Projection LB, Dual LB, Reduction LB, Independent Min LB



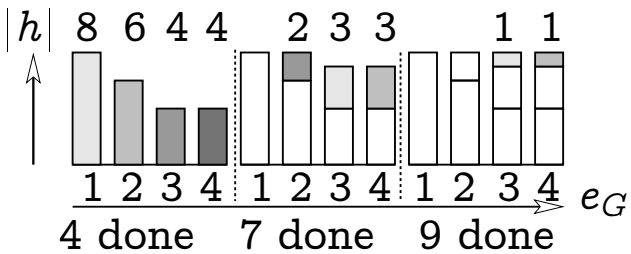
# Enhancing: Cardinality Based Load Balancing

- Composite keys are skewed on number of histograms




| $g$ | $\lvert h\rvert$ | $\{g^1, g^2\}$ | $\lvert h\rvert$ |
|---|---|---|---|
| $g^1 = 0$ | 8 | $\{0, 0\}$ | 8 |
| $g^1 = 1$ | 8 | $\{1, 1\}$ | 6 |
| $g^1 = 2$ | 8 | $\{2, 2\}$ | 4 |
| $g^1 = 3$ | 8 | $\{3, 2\}$ | 4 |
| $g^2 = 0$ | 8 | $\{2, 3\}$ | 3 |
| $g^2 = 1$ | 8 | $\{3, 3\}$ | 3 |
| $g^2 = 2$ | 8 | $\{1, 3\}$ | 2 |
| $g^2 = 3$ | 8 | $\{2, 1\}$ | 1 |
| | | $\{3, 1\}$ | 1 |
| | | Others | 0 |

- Group composite cells to achieve balanced loads

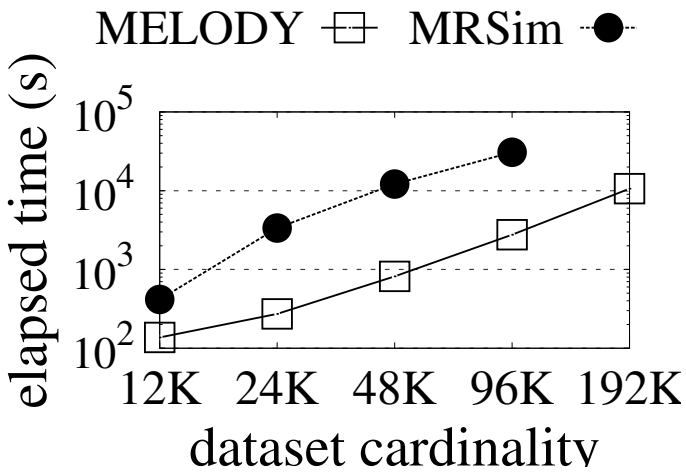



| $e_G$ | $G$ | $\lvert h\rvert$ |
|---|---|---|
| 1 | $\{0, 0\}$ | 8 |
| 2 | $\{1, 1\}$, $\{1, 3\}$ | 8 |
| 3 | $\{2, 2\}$, $\{2, 3\}$, $\{2, 1\}$ | 8 |
| 4 | $\{3, 2\}$, $\{3, 3\}$, $\{3, 1\}$ | 8 |

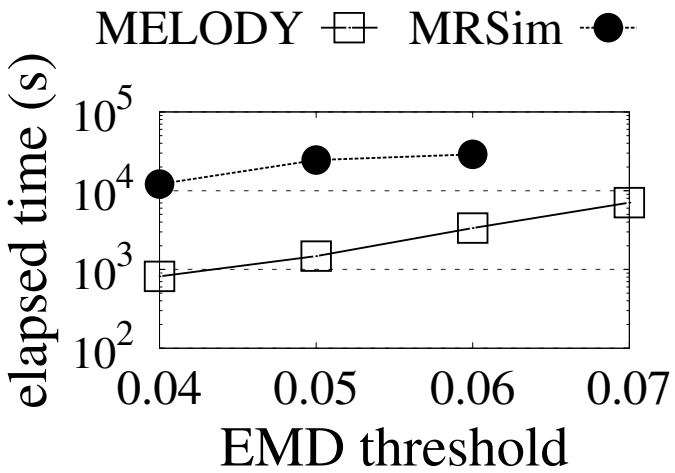
# Experiments: Settings

- Image collections:
  - COREL: 68040 images
  - MIRFLICKR: 1 millions images
- Image Feature Representations (for evaluation only):
  - COREL: MPEG-7 Dominant Color Histogram (CC), MPEG-7 Color Layout Histogram (CL)
  - MIRFLICKR: MPEG-8 Edge Histogram Vertical (MV), Horizontal (MH), and Slash (MS)
- Default to 3 projections, $4 \times 4$ grid, and on a 48-node Hadoop instance

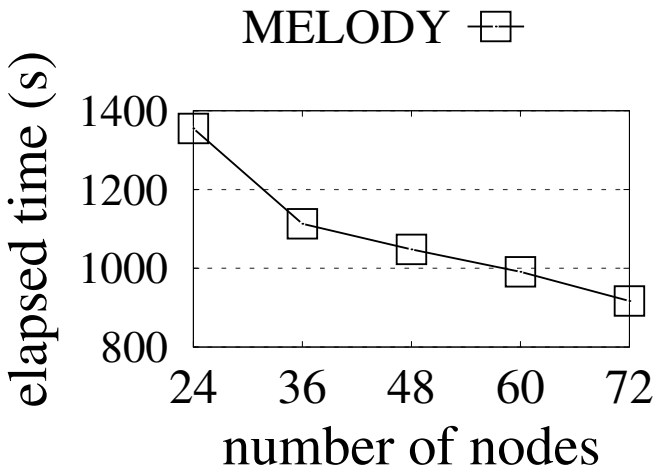
# Experiments: Results




(l) Varying Cardinarlity




(m) Varying Threshold




(n) Scaling Out

# Conclusion

- First study on EMD similarity join
- A novel framework Melody-Join to leverage lower bounds for pruning
- Enhancepruning power by multiple lower bounds; balance load by quantile based grid and cardinality based grouping.
- Extensive experiments confirming orders of magnitude improvement on the state-of-the-art technique.

- Future work: top-$k$ similarity join, other frameworks


Jin Huang: jin.huang@unimelb.edu.au

# Appendix A: Discussion on MRSimJoin

MRSimJoin does not support LB pruning:

- Most LB are not metric, e.g., without triangular inequality and transitivity
- Even metric LB may not preserve the *locality* of EMD, e.g., the nearest pivot of a record in terms of LB may not be the nearest pivot in terms of EMD

LB can be integrated into MRSimJoin in refining

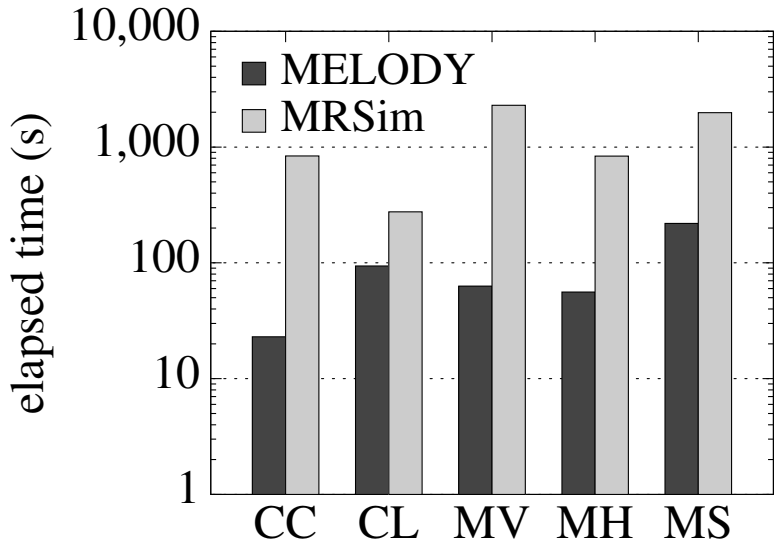
- Experimental comparison is conducted on implementation with the identical chain of LB from Melody-Join

# of pivots in MRSimJoin is selected to produce the number of reduce tasks that fits into the capacity of the cluster

# Appendix B: Discussion on Load Balance

Effects of our load balancing efforts

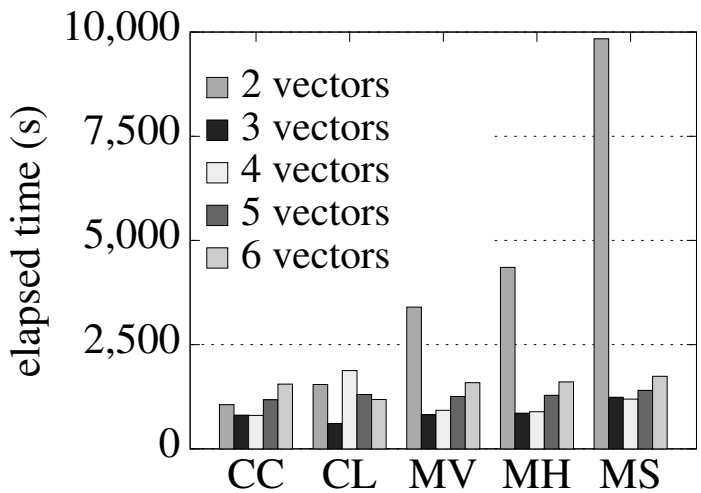
- Use the standard deviation on the completion time of reducers

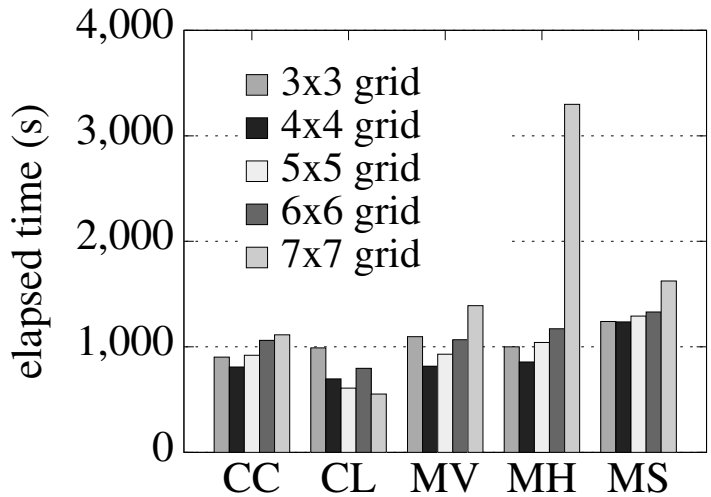



- The standard deviation of Melody-Join is orders of magnitude smaller than that of MRSimJoin

# Appendix C: # of Projections and Grid Granularity

Effects of parameters on the performance of MELODY-JOIN




3 projection vectors and $4 \times 4$ grid achieve the best performance in most datasets