Dependency-Based Histogram Synopses for High-dimensional Data

Amol Deshpande, UC Berkeley
Minos Garofalakis, Bell Labs
Rajeev Rastogi, Bell Labs

Why Synopses ???

- Selectivity estimation for query optimization
- Approximate querying
  - Useful when not feasible to query the entire database
- Prevalent techniques :
  - Histograms, wavelets
    - Suffer from “Curse of dimensionality”
  - Random Sampling
    - Very few matches for selection in sparse high-dimensional data

Problem Statement

- Given a “counts table”, find an approximate answer to an aggregate range sum query
- The counts table can be thought of as a joint probability distribution
  - Evaluating an aggregate range sum query equivalent to finding a joint probability distribution

A	B	count
1	1	1
1	2	1
2	2	1

$$\begin{array}{cc|c|c|} & & & \\ & 2 & 1/3 & 1/3 \\ \cline{3-4} B & & & \\ & 1 & 1/3 & 0 \\ \cline{3-4} & & 1 & 2 \\ & & \multicolumn{2}{c}{A} \end{array} \times 3$$

Problem Statement

- Given a “counts table”, find an approximate answer to an aggregate range sum query
- The counts table can be thought of as a joint probability distribution
  - Evaluating an aggregate range sum query equivalent to finding a joint probability distribution

A	B	count
1	1	1
1	2	1
2	2	1

$$\begin{array}{cc|c|c|} & & \\ & 2 & 1/3 & 1/3 \\ B & & & \\ & 1 & 1/3 & 0 \\ \hline & & 1 & 2 \\ & & \multicolumn{2}{c}{A} \end{array} \times 3$$

Problem Statement

- Given a “counts table”, find an approximate answer to an aggregate range sum query
- The counts table can be thought of as a joint probability distribution
  - Evaluating an aggregate range sum query equivalent to finding a joint probability distribution

A	B	count
1	1	1
1	2	1
2	2	1

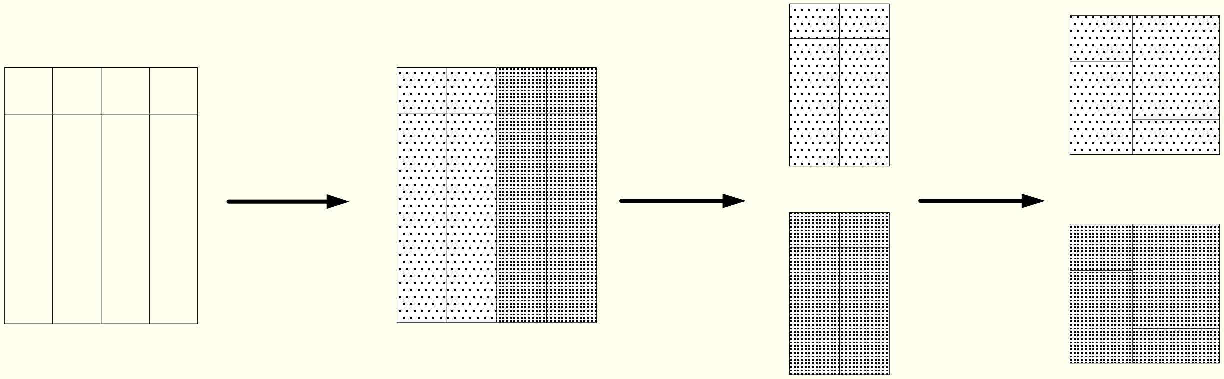
$$\begin{array}{cc|c|c|c} & 2 & 1/3 & 1/3 & \\ B & & & & \times 3 \\ & 1 & 1/3 & 0 & \\ & & 1 & 2 & \\ & & \multicolumn{2}{c}{A} & \end{array}$$

Histograms for High Dimensions

- Assume "attribute independence" and build per-attribute one-dimensional histograms
  - Simple to build and maintain
  - Highly inaccurate in presence of correlations
- "Multi-dimensional" histograms [PI'97, MVW'98, GKT'00]
  - Expensive to build and maintain
  - Large number of buckets required for reasonable accuracy in high dimensions
  - Not suitable for queries on lower-dimensional subsets of attributes
- Extremes in terms of the underlying correlations!!

Our Approach : Dependency Based (DB) Histograms

- Build a statistical model on the attributes of data
- Based on model, build a set of low dimensional histograms
- Use this collection of histograms to provide approximate answers



Outline

- Motivation
- Decomposable Models
- Building a collection of histograms
- Query evaluation
- Experimental evaluation

Decomposable Models

- Specify correlations between attributes
- Examples :
  - Partial Independence :

    $p(\text{salary} = s, \text{height} = h, \text{weight} = w) =$

    $p(\text{salary} = s)p(\text{height} = h, \text{weight} = w)$
  - Conditional Independence :

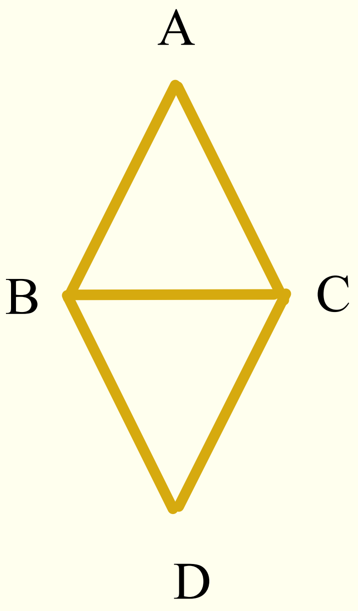
    $p(\text{salary} = s, \text{age} = a \mid \text{YPE} = y) =$

    $p(\text{salary} = s \mid \text{YPE} = y)\ p(\text{age} = a \mid \text{YPE} = y)$
- Advantages of Decomposable Models:
  - Closed form estimates for the joint probability exist
  - Interpretation in terms of partial and conditional independence statements
  - Can be represented as a graph

Decomposable Models - Example

- Interpretation
  - Attributes A and D are conditionally independent given attributes B and C, i.e.,

    p(AD|BC) = p(A|BC)p(D|BC)
- Graphical Representation :
  - Markov Property : If T separates U and V, then p(UV|T) = p(U|T)p(V|T)

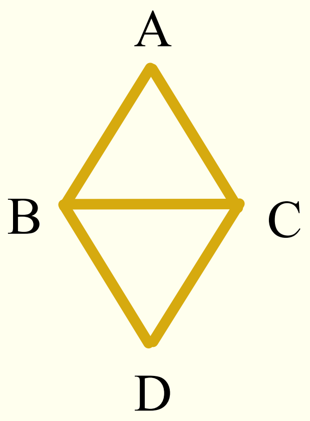


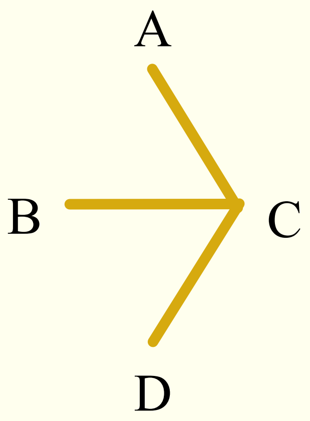
- Joint Probability Distribution

  p(ABCD) = p(ABC)p(DBC)/p(BC)

What to do with the Model ?

- Build clique histograms on marginals corresponding to the maximal cliques of the model





$p(ABCD) = p(ABC)p(BCD)/p(BC)$

Histograms on : ABC, BCD

$p(ABCD) = p(AC)p(BC)p(DC)/p(C)^2$

Histograms on : AC, BC, DC

Searching for the Best Model

- NP-hard to find the best model
- Heuristic forward selection :
  - Start with full independence assumption and grow the model greedily
  - Growing a model :
    - Need to stay in the space of decomposable models
    - Naïve approach : Try every possible extension of the current model
      - Works for small number of attributes
    - Developed a more sophisticated algorithm [DGJ01]

Choosing among Models

- Kullback-Leibler Information Divergence
  - A measure of "distance" between two probability distributions

$$KL(p \| p') = \sum_x p(x) \log \frac{p(x)}{p'(x)}$$

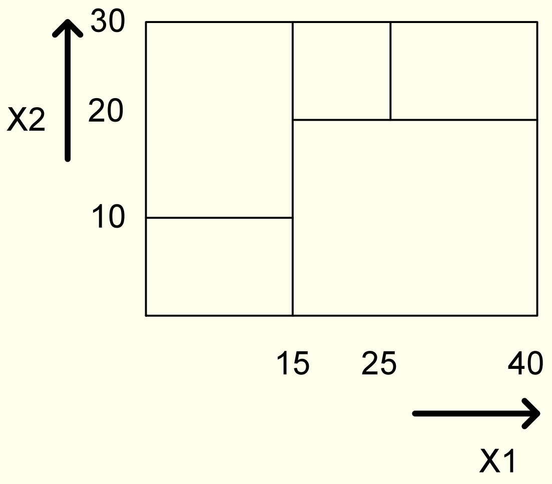
- Choosing among possible extensions
  - Maximize the increase in approximation accuracy due to increase in complexity
  - Maximize ratio of increase approximation accuracy and the increase in total state space
- When to stop ?
  - Limit the maximum dimensionality of a histogram

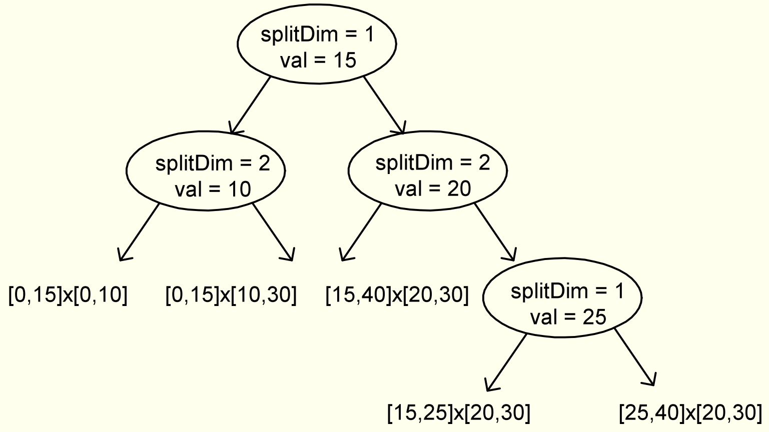
Outline

- Motivation
- Decomposable Models
- Building a collection of histograms
- Query evaluation
- Experimental evaluation

Building Clique Histograms

- MHIST approach [PI97] :
  - Partition the space to be covered through recursive splits.
- Split Tree representation of MHISTs





- MHIST projection and multiplication :
  - Performed directly on the Split Tree representation

Storage Space Allocation

- Minimize total error for given storage space
- Can be solved in time $O(CB^2)$:
  - C is # histograms and B is total space available
- More efficient heuristic :
  - Greedily allocate additional buckets to the histogram that maximizes the decrease in error per unit space
  - Optimal if the individual histogram error functions follow the law of diminishing returns

Outline

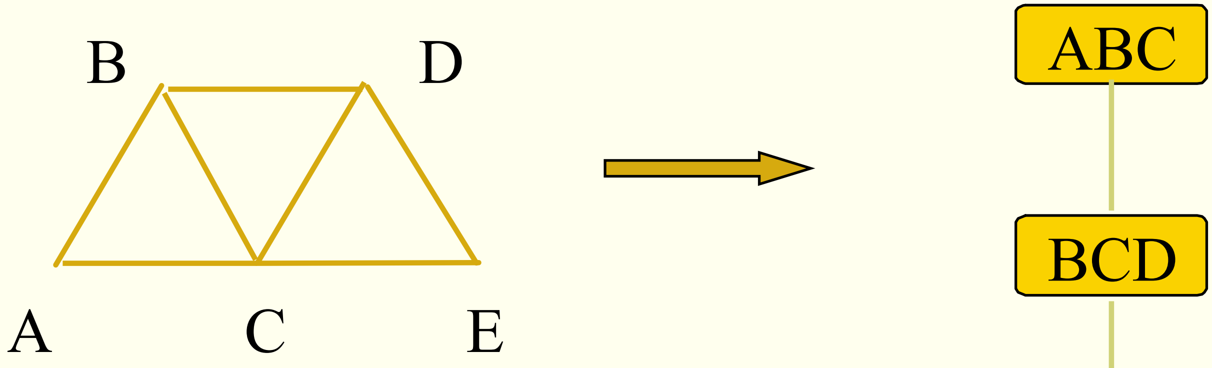
- Motivation
- Decomposable Models
- Building a collection of histograms
- Query evaluation
- Experimental evaluation

Query Evaluation

- Compute joint probability distribution and project
- More efficient evaluation algorithm :
  - Use "Junction Tree" of the model graph
    - Nodes : Maximal cliques of the model
    - Edges : An edge between two nodes A and B only if S = A ∩ B separates A - S and B - S.
  - Minimize the number of operations for computing any marginal probability distribution

- Operation ordering is related to join order optimization

Junction Trees



Computing *p*(AD)

$$p(AD) = \sum_{B,C} p(ABC)\,p(BCD) / p(BC)$$

Computing *p*(AE)

$$p(ACD) = \sum_{B} p(ABC)\,p(BCD) / p(BC)$$

$$p(AE) = \sum_{C,D} p(ACD)\,p(CDE) / p(CD)$$

Outline

- Motivation
- Decomposable Models
- Building the collection of histograms
- Query evaluation
- Experimental evaluation

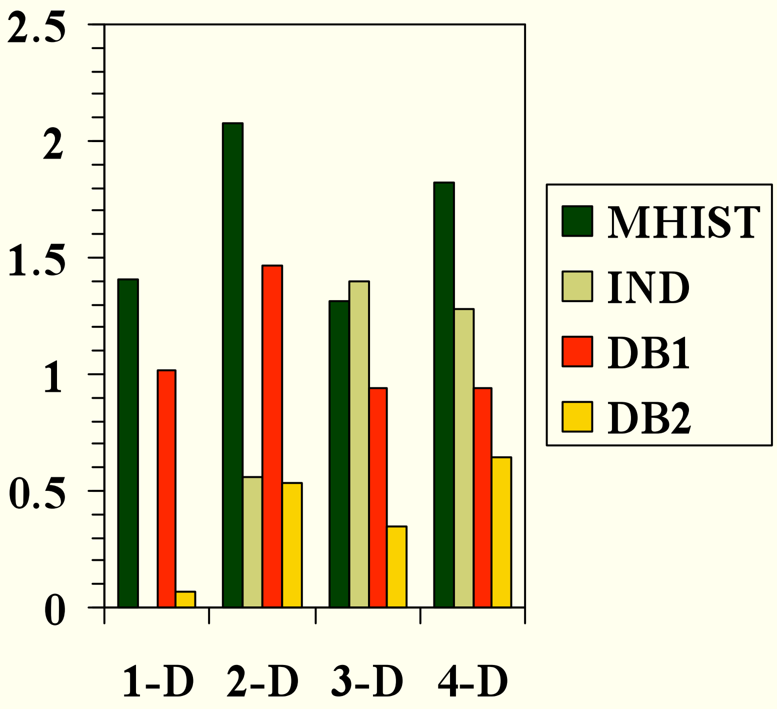
Experimental Evaluation

- Census Data :
  - Census-6 :
    - citizenship, native country of father, native country of mother, native country of the sample person, occupation Code, age
  - Census-12 :
    - industry code, hours worked, education, state, county, race
- Error Metrics :
  - Absolute Relative Error :
    - |correct – approx|/correct
  - Multiplicative Error :
    - max{correct, approx}/min{correct,approx}

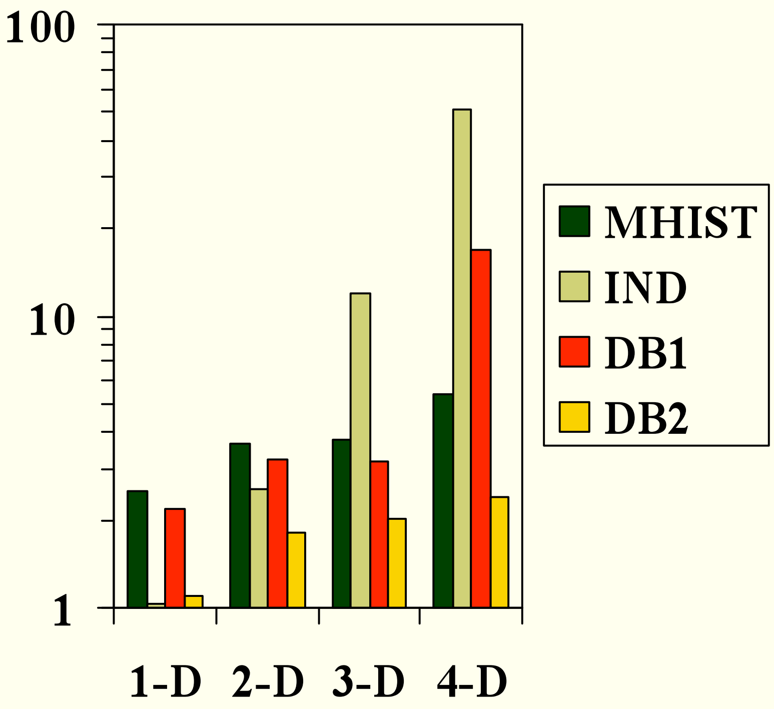
Methods Compared

- MHIST :
  - Multi-dimensional histogram on all the attributes
- IND :
  - Per attribute one-dimensional histograms
- Dependency-based histograms :
  - DB1 : Model selected based on statistical significance
  - DB2 : Model selected with the goal of minimizing the ratio of approximation accuracy and total state space

Results on Census-6



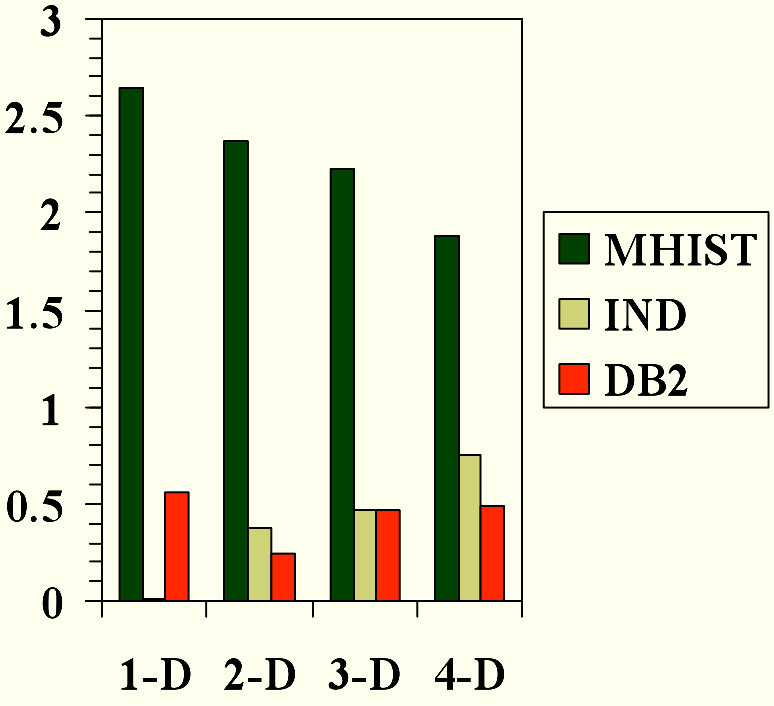
Absolute Relative Error



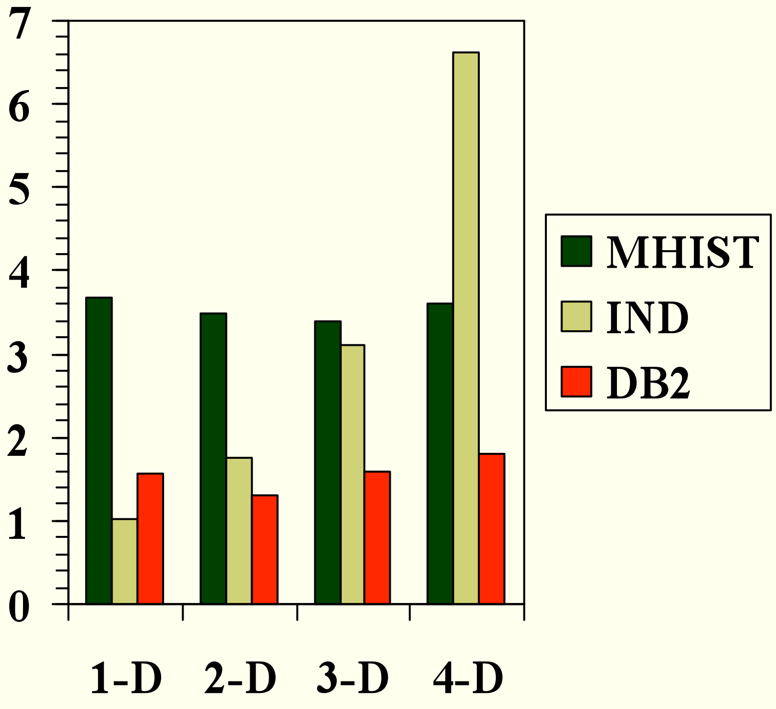
Multiplicative Error

Results on Census-12

Absolute Relative Error



Multiplicative Error



Summary of Results

- Decomposable Models are Effective
  - Good approximations with models of small complexity
- Better Approximate Answer Quality
  - As much as 5 times lower errors
- Storage Efficient
  - Fairly accurate answers with less than 1% space

Conclusions

- Proposed an approach to building synopses by explicitly identifying and using correlations present in the data
- Developed an efficient forward selection procedure
- Developed efficient algorithms for building and using collections of histograms
- General methodology presented applicable to other synopsis methods as well

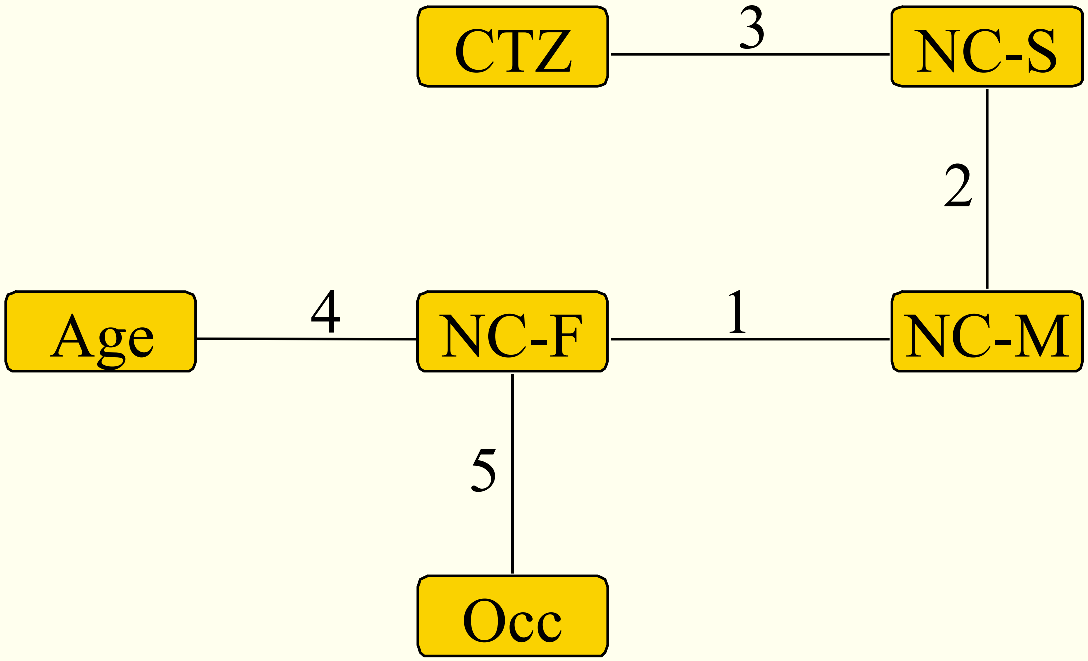
Future Work

- Maintenance
  - Additional problem of maintaining the underlying model
- Error Guarantees
- Applicability to other synopsis techniques
- Exploiting more general class of models for storing clique marginals

References

- [MD88] Muralikrishna and Dewitt; Equi-depth histograms for estimating selectivity factors for multi-dimensional queries; SIGMOD'88
- [PI97] Poosala and Ioannidis; Selectivity Estimation Without the Attribute Value Independence Assumption; VLDB'97
- [BFH75] Bishop, Fienberg and Holland; Discrete Multivariate Analysis; MIT Press, 1975
- [MVW98] Matias, Vitter and Wang; Wavelet-Based Histograms for Selectivity Estimation; SIGMOD'98
- [DGJ01] Deshpande, Garofalakis and Jordan; Efficient Stepwise Selection in Decomposable Models; UAI'01

Census-6 : Model Found



Previous Work

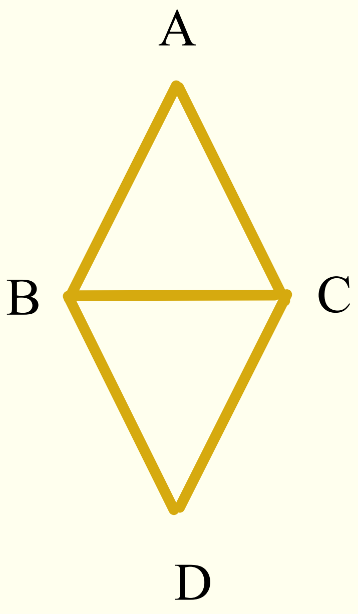
- Approximate Querying
  - Maintaining a collection of histograms to answer queries
    - [PK00] Based on the query workload. Use "Iterative Proportional Fitting" to answer queries.
    - [PG99] Driven by a pre-specified error bound.
- Issues not addressed :
  - Selection of Histograms
  - Answering higher-dimensional queries using lower-dimensional histogram.

References

- [PG99] Poosala and Ganti; Fast approximate answers to aggregate queries on a data cube; SSDBM'99
- [PK00] Palpanas and Koudos; Entropy based approximate querying and exploration of datacubes; TR, Univ of Toronto, 2000

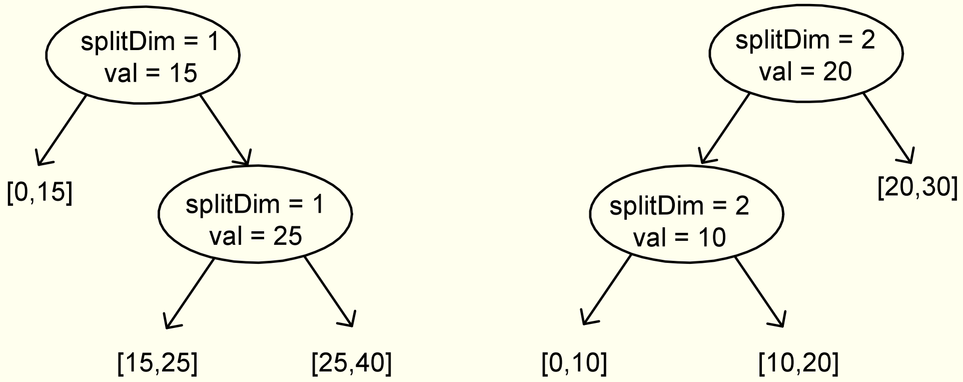
What to do with the Model ?

- Build clique histograms on the probability marginals corresponding to the maximal cliques of the model
- Example :
  - Build histograms on the attribute sets ABC and BCD
- Full probability distribution can be recovered from clique marginals
  - p(ABCD) = p(ABC)p(BCD)/p(BC)



Operations on Mhists

- Our algorithms perform required operations directly on the Split Tree representation
- Operations :
  - Projection



  - Multiplication