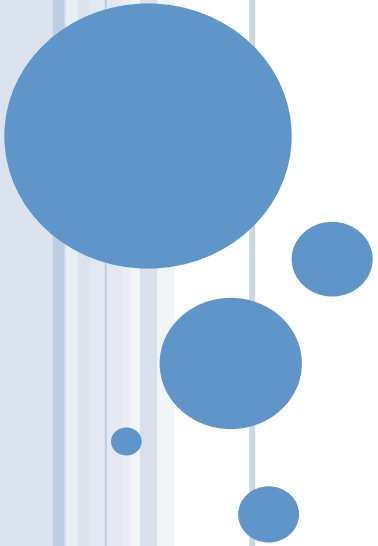


Truth Finding with Attribute Partitioning

M. Lamine BA (Télécom ParisTech), **Roxana HORINCAR** (Télécom ParisTech), Pierre SENELLART (Télécom ParisTech), Huayu WU (A*STAR, I2R)

WebDB Workshop, Melbourne, Australia May 31, 2015

Data Quality Problem

- Dirty data
  - Incomplete data
  - Duplicates
  - Inaccurate data
  - Inconsistent data
  - Stale data
  - Misformatted data
  - Undocumented data
  - Conflicting data

- How does data get dirty?
  - Data gathering
  - Storage
  - Transmission
  - Transformation
  - Integration
  - Deliberately falsified

OUTLINE

- Truth Finding
- Exploiting Structure
- Experimental Results
- Conclusions

OUTLINE

- Truth Finding
- Exploiting Structure
- Experimental Results
- Conclusions

TRUTH FINDING

- Context
  - Multiple sources: websites, blogs, forums, mailing lists, documents
  - Relations claimed by sources: objects, attributes
  - "One truth" setting: 1 true value, n false values

- Problem: determine which of the statements made by contradictory sources is correct
- Goals: source accuracy level, attribute value correctness level, correct attribute values discovery

- Real-world applications: query answering, source selection, crowdsourcing, data integration, Web data quality

TRUTH FINDING ALGORITHMS

- Majority Voting algorithm: true value provided by the largest number of sources

- Weighted Voting algorithms (e.g., AccuVote [Dong et al., 2009]): assigns a higher vote to a source with a higher accuracy, true value with the highest sum of votes
  - Source accuracy level
  - Value correctness level

$$A(S) = \frac{1}{|V(S)|} \sum_{v \in V(S)} P(v)$$

$$P(v) = \frac{\prod_{S \in S(v)} \frac{n \cdot A(S)}{1 - A(S)}}{normalizing\_factor}$$

- Improvements based on domain-specific characteristics
  - Attribute value similarity [Yin et al., 2008]
  - Copying relationships between sources [Dong et al., 2009]
  - Source correlations [Pochampally et al., 2014]
  - Fact hardness [Galland et al., 2010]

- But none of them look at the structure of the facts… *we do!*

OUTLINE

- Truth Finding
- Exploiting Structure
- Experimental Results
- Conclusions

EXAM EXAMPLE:
3 STUDENTS (SOURCES), 2 TESTS (OBJECTS), 2 QUESTIONS (ATTRIBUTES), 2 DOMAINS (MATH, GEOGRAPHY)

Test 1: **[Math]** Provide the set of prime numbers smaller than 10.

[Geography] What is the capital city of Australia?

Test 2: **[Math]** Give a natural number x satisfying x mod 4 = 0.

[Geography] What is the largest country in the European Union?

	Test	Math	Geography
Student 1	Test 1	**{2, 3, 5, 7}**	Melbourne
	Test 2	**24**	Spain
Student 2	Test 1	{2, 4, 6, 8}	**Canberra**
	Test 2	26	**France**
Student 3	Test 1	**{2, 3, 5, 7}**	Sydney
	Test 2	41	**France**

Correlated Attributes in Truth Finding

- Objects with inherent structure
- Different local source accuracies on different attribute subsets, rather than one global accuracy
- Unknown correlated attributes: challenge & opportunity

ATTRIBUTE PARTITIONING

- AccuPartition problem
  - Find an optimal partition of the attribute set such that running any base truth finding algorithm on each partition subset maximizes the overall precision of the truth finding process

- Optimal partition estimation
  - Partition weight
    - Estimates the precision of the truth finding process on a partition, the optimality level of a given partition
  - Subset score function
    - Estimates the precision of the truth finding process on a partition subset
    - Evaluated based on local source accuracy values: maxAccu, avgAccu, oracle

SOLVING ACCUPARTITION

- Start with any existing truth finding algorithm
- Explore the partition space and determine the optimal partition
- Use the truth finding algorithm separately on the subsets of the optimal partition

Finding the Optimal Partition

- Exact algorithm, optimal solution
  - GenAccuPartition
    - Exhaustive exploration
    - Exponential in the size of the attribute set
- Approximate algorithm, near-optimal solution
  - SamplingAccuPartition
    - Random uniform sampling approach
    - Restricts the optimal partition search to a limited number of candidates

OUTLINE

- Truth Finding
- Exploiting Structure
- Experimental Results
- Conclusions

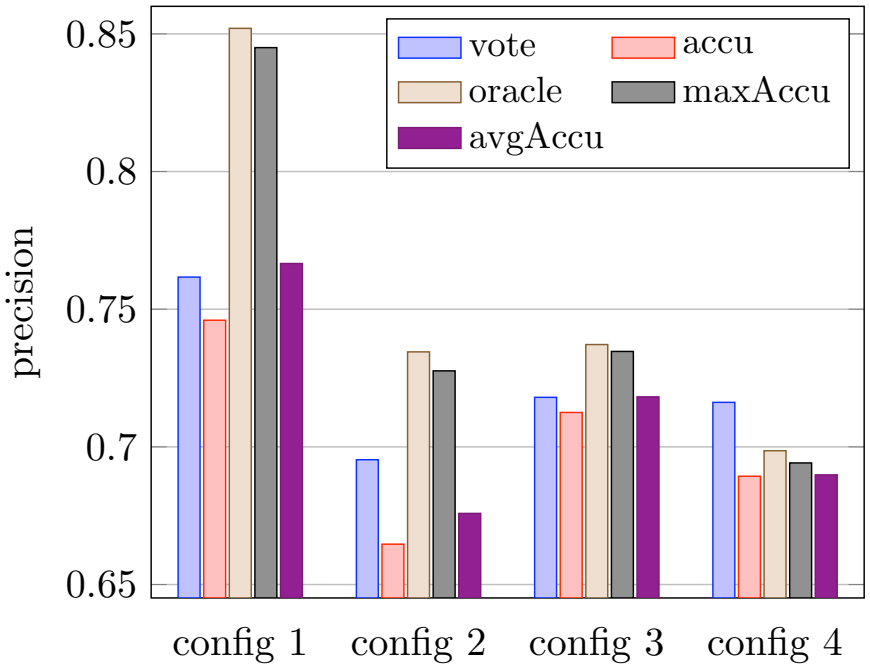
Experimental Evaluation

- Precision results: GenAccuPartition, SamplingAccuPartition
- Baseline: AccuVote; also Vote

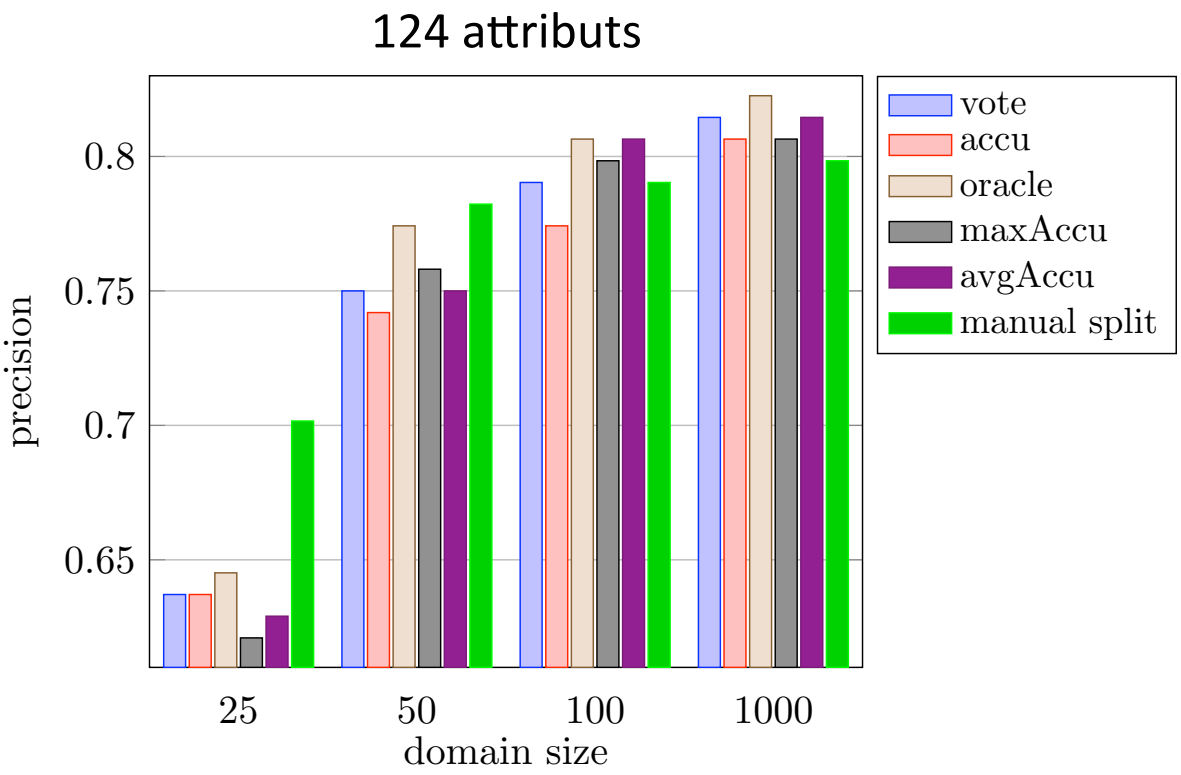
- Experimental setup
  - Synthetic data
    - 60,000 data items: 10 sources, 1000 objects, 6 attributes
    - Generator: uniform distribution functions m1, m2
  - Real data: Exam dataset
    - 247 students (sources), 1 test (object), 124 questions (attributes)
    - 9 domains: Math 1A, Physics, Chemistry 1, Math 1B, Electrical Engineering, Computer Science, Chemistry 2, Life Sciences, Math 2
    - False value generator, domain size: 25, 50, 100, 1000

Configs	m1	m2
config 1	1.0	0.0
config 2	0.8	0.0
config 3	0.8	0.2
config 4	0.6	0.4

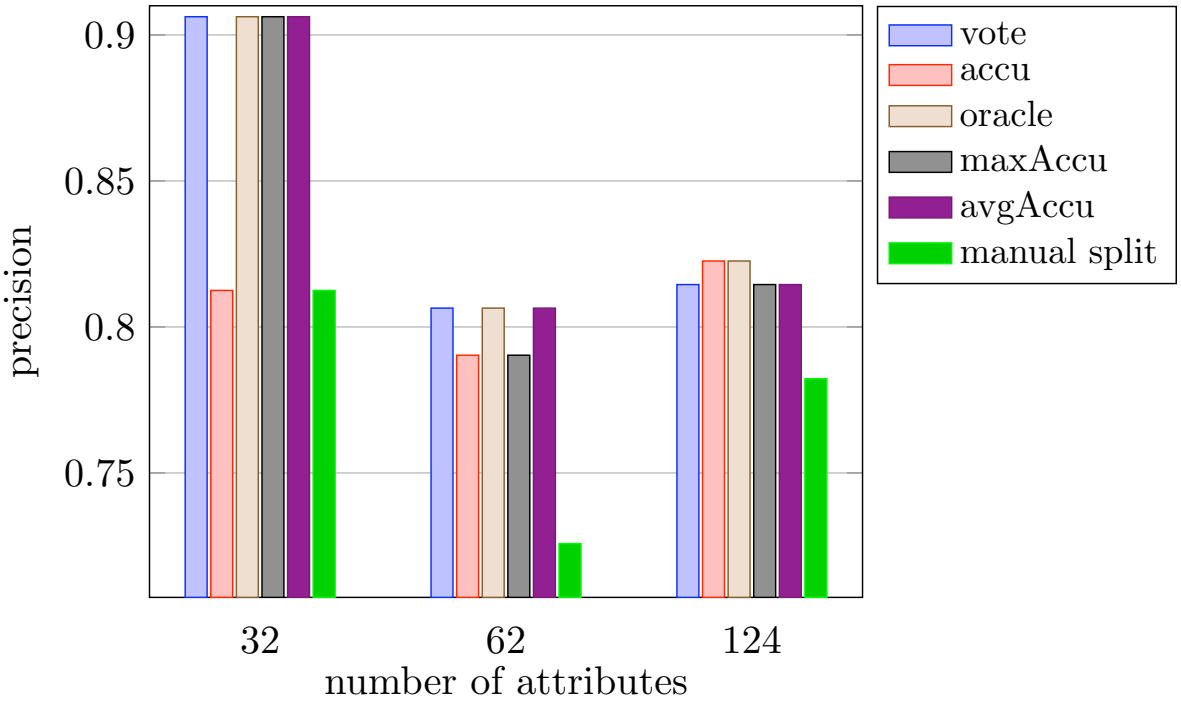
SYNTHETIC DATA



Real-World Data with Artificial Coverage



REAL-WORLD DATA



OUTLINE

- Truth Finding
- Exploiting Structure
- Experimental Results
- Conclusions

Conclusions & Perspectives

- Conclusions
  - Attributes with inherent structure
  - Possible to use structure to improve quality of truth finding
  - AccuPartition: can be used on top of any truth finding method

- Perspectives
  - New subset scores and partition weighting functions
  - Automatic partition generation
    - Greedy approach
    - Functional dependencies discovery
  - Combine attribute partitioning with source selection methods

REFERENCES

- X. L. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global detection of complex copying relationships between sources. VLDB Endow., 3(1), 2010.
- X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. PVLDB, 2(1), 2009.
- X. L. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. PVLDB, 2(1), 2009.
- A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In WSDM, 2010. sources for data integration. PVLDB, 5(6), 2012.
- X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? PVLDB, 6(2), 2012.
- R. Pochampally, A. Das Sarma, X. L. Dong, A. Meliou, and D. Srivastava. Fusing data with correlations. In SIGMOD, 2014.
- X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. IEEE TKDE, 2008.

Thank you!