TELECOM ParisTech

Truth Finding with Attribute Partitioning

M. Lamine Ba Roxana Horincar
Pierre Senellart Huayu Wu

CCIPX final meeting, 19 April 2015

Outline

Truth Finding

- Context:
  - Set of sources stating facts about real-world objects
  - (Possible) functional dependencies between facts
  - Fully unsupervised setting: we do not assume any information on truth values of facts or inherent trust in sources
- Problem: determine which facts are true and which facts are false
- Real world applications: query answering, source selection, data quality assessment on the web, making good use of the wisdom of crowds

M. Lamine Ba Roxana Horincar Pierre Senellart Huayu Wu

Motivating Example

What are the capital cities of European countries?

	France	Italy	Poland	Romania	Hungary
Alice	Paris	Rome	Warsaw	Bucharest	Budapest
Bob	?	Rome	Warsaw	Bucharest	Budapest
Charlie	Paris	Rome	Katowice	Bucharest	Budapest
David	Paris	Rome	Bratislava	Budapest	Sofia
Eve	Paris	Florence	Warsaw	Budapest	Sofia
Fred	Rome	?	?	Budapest	Sofia
George	Rome	?	?	?	Sofia

Voting

Information: redundance

	France	Italy	Poland	Romania	Hungary
Alice	Paris	Rome	Warsaw	Bucharest	Budapest
Bob	?	Rome	Warsaw	Bucharest	Budapest
Charlie	Paris	Rome	Katowice	Bucharest	Budapest
David	Paris	Rome	Bratislava	Budapest	Sofia
Eve	Paris	Florence	Warsaw	Budapest	Sofia
Fred	Rome	?	?	Budapest	Sofia
George	Rome	?	?	?	Sofia
Frequence	**P.** 0.67	**R.** 0.80	**W.** 0.60	Buch. 0.50	Bud. 0.43
	R. 0.33	F. 0.20	K. 0.20	Bud. 0.50	**S.** 0.57
			B. 0.20		

Evaluating Trustworthiness of Sources

Information: redundance, trustworthiness of sources (= average frequence of predicted correctness)

	France	Italy	Poland	Romania	Hungary	Trust
Alice	Paris	Rome	Warsaw	Bucharest	Budapest	0.60
Bob	?	Rome	Warsaw	Bucharest	Budapest	0.58
Charlie	Paris	Rome	Katowice	Bucharest	Budapest	0.52
David	Paris	Rome	Bratislava	Budapest	Sofia	0.55
Eve	Paris	Florence	Warsaw	Budapest	Sofia	0.51
Fred	Rome	?	?	Budapest	Sofia	0.47
George	Rome	?	?	?	Sofia	0.45
Frequence	**P.** 0.70	**R.** 0.82	**W.** 0.61	**Buch.** 0.53	Bud. 0.46	
weighted	R. 0.30	F. 0.18	K. 0.19	Bud. 0.47	**S.** 0.54	
by trust			B 0.20			

Iterative Fixpoint Computation

Information: redundance, trustworthiness of sources with iterative fixpoint computation

	France	Italy	Poland	Romania	Hungary	Trust
Alice	Paris	Rome	Warsaw	Bucharest	Budapest	0.65
Bob	?	Rome	Warsaw	Bucharest	Budapest	0.63
Charlie	Paris	Rome	Katowice	Bucharest	Budapest	0.57
David	Paris	Rome	Bratislava	Budapest	Sofia	0.54
Eve	Paris	Florence	Warsaw	Budapest	Sofia	0.49
Fred	Rome	?	?	Budapest	Sofia	0.39
George	Rome	?	?	?	Sofia	0.37
Frequence	**P.** 0.75	**R.** 0.83	**W.** 0.62	**Buch.** 0.57	**Bud.** 0.51	
weighted	R. 0.25	F. 0.17	K. 0.20	Bud. 0.43	S. 0.49	
by trust			B 0.19			

Truth Finding Techniques

- Numerous truth finding techniques in the literature
- Model (depending on method):
  - trustworthiness of sources
  - hardness of facts
  - proximity of answers (edit distance, numerical difference)
  - copies across sources
  - which sources are useful to answer
- But none of them look at the structure of the facts

Outline

Example: Student Testing

- Test questions

  Test 1: 1. Provide the set of prime numbers smaller than 10.
  2. What is the capital city of Romania?

  Test 2: 1. Give a natural number x satisfying $x \bmod 4 = 0$.
  2. What is the largest country in the European Union?

- Student's answers

	Test	Math	Geography
student 1	Test 1	$\{2, 3, 5, 7\}$	Budapest
	Test 2	24	Spain
student 2	Test 1	{2, 4, 6, 8}	Bucharest
	Test 2	26	France
student 3	Test 1	$\{2, 3, 5, 7\}$	Belgrade
	Test 2	41	France

Attribute Partitioning

- Sources have different accuracy levels rather than a global accuracy
- In the previous examples, better to give a trustworthiness score per student per topic than per student
- What if we do not know in advance the optimal level at which to estimate accuracy?

AccuPartition problem

Find an optimal partitioning of the input attribute set which maximizes the precision of the truth finding process

Estimation of optimal partition

- Computation of a weight value for each partition
  - Estimates the precision of the truth finding process on each partition
  - For example, the average value of the (estimated) precision scores of its blocks
- Scoring function for blocks of correlated data attributes
  - Estimates the precision of the process on each block
  - Determines scores based on source local accuracy values
- Various possible scoring strategies
  - Average accuracy value
  - Maximum accuracy value
  - Oracle accuracy value

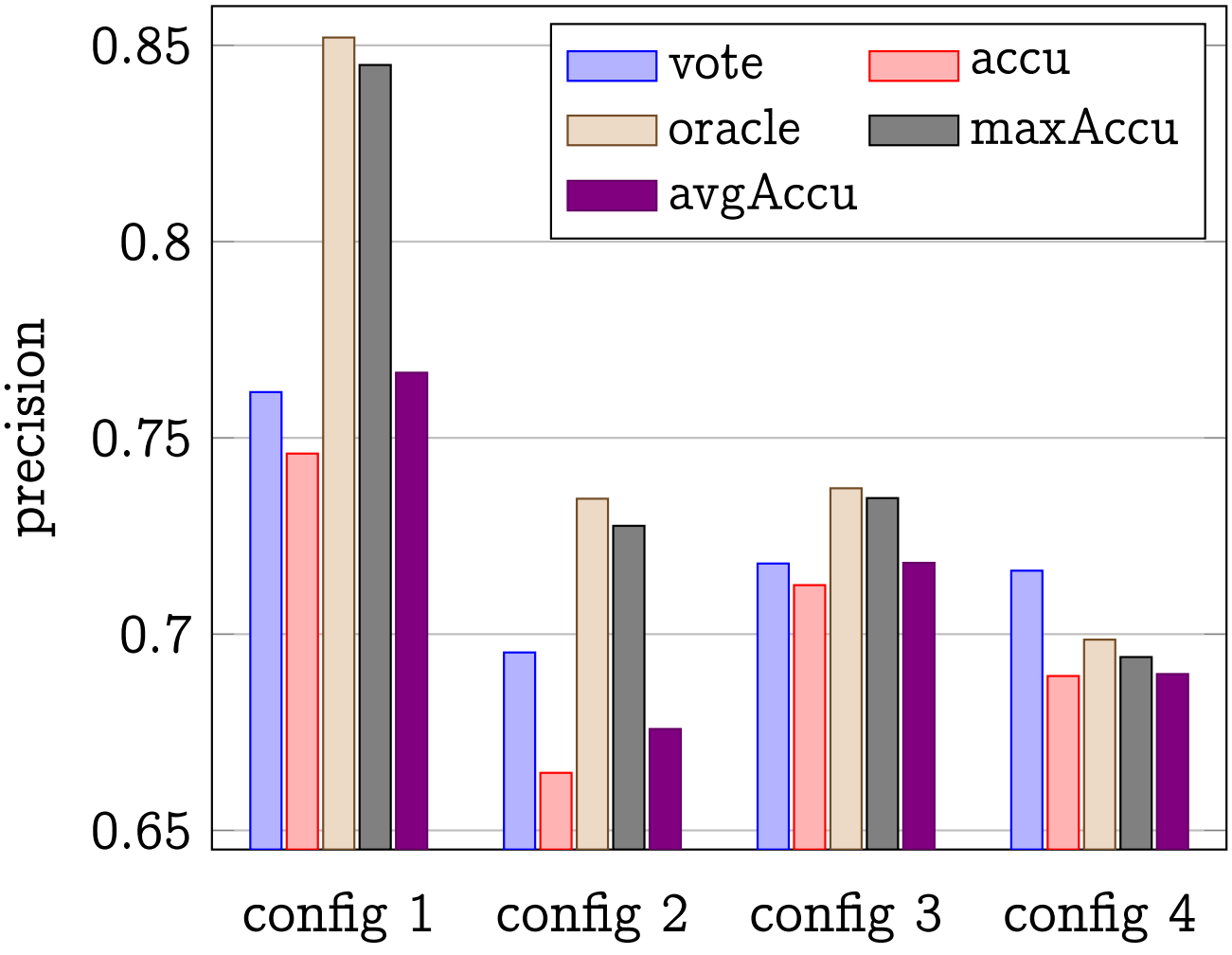
Solving AccuPartition

- Start with any existing truth finding algorithm
- Explore the space of all partitions and determine the best one:
  - Exhaustive exploration
  - Bottom-up greedy partition construction
  - Use sampling to explore only a subset of the partitions
- Use the truth finding algorithm on the optimal partition

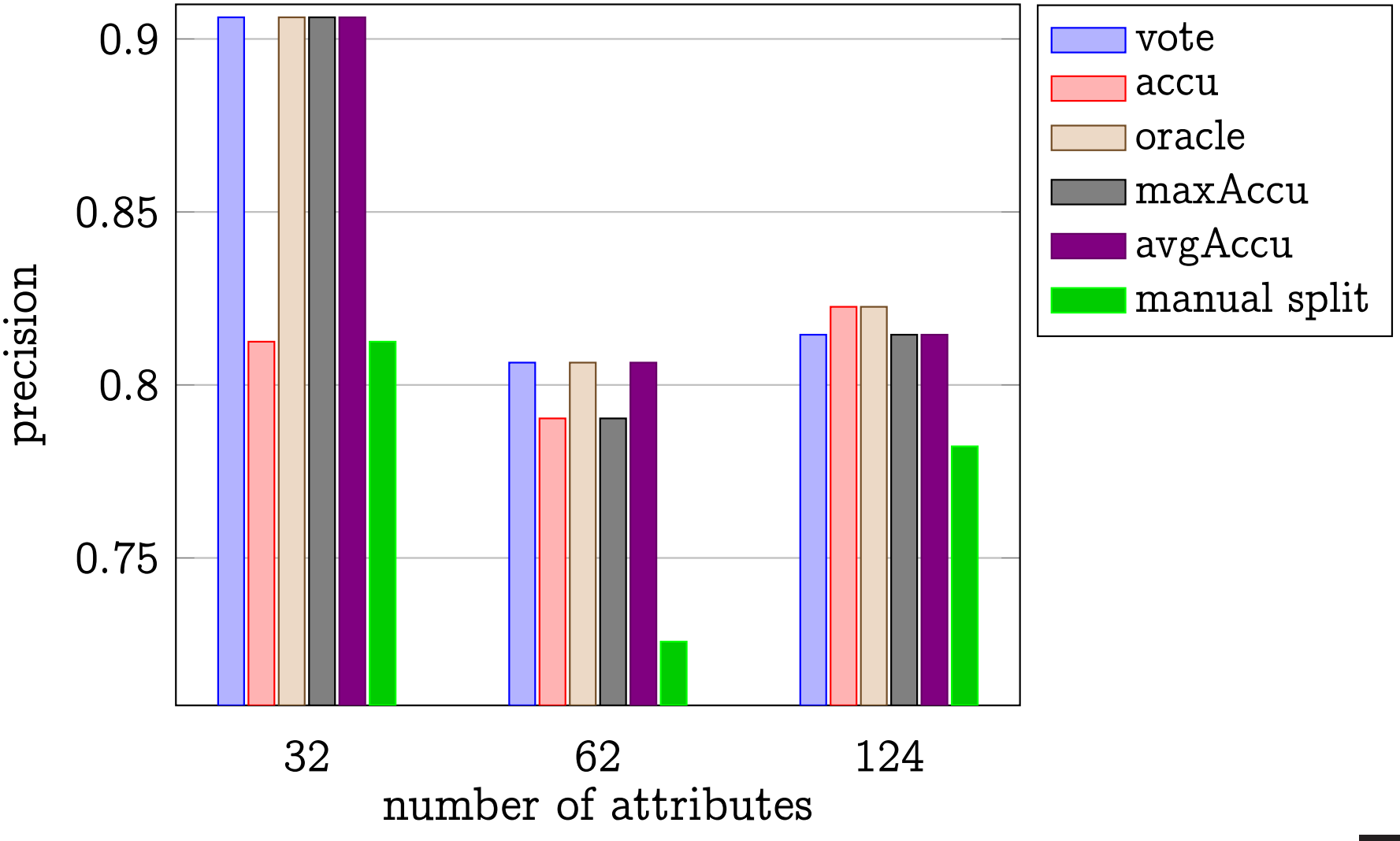
M. Lamine Ba Roxana Horincar Pierre Senellart Huayu Wu

Outline

Results on Synthetic Data



Results on Real-World Data



Outline

Conclusions

- Possible in certain cases to use structure to improve quality of truth finding
- Can be used on top of any truth finding method