Entity Profiling with Varying Source Reliabilities

Furong Li, Mong Li Lee, Wynne Hsu

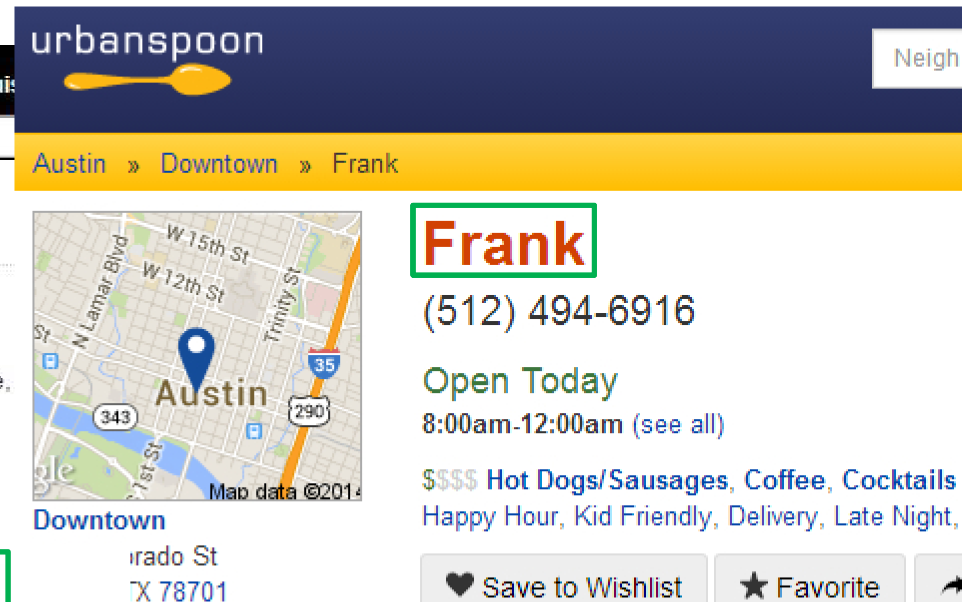
{furongli, leeml, whsu} @ comp.nus.edu.sg

National University of Singapore

Outline

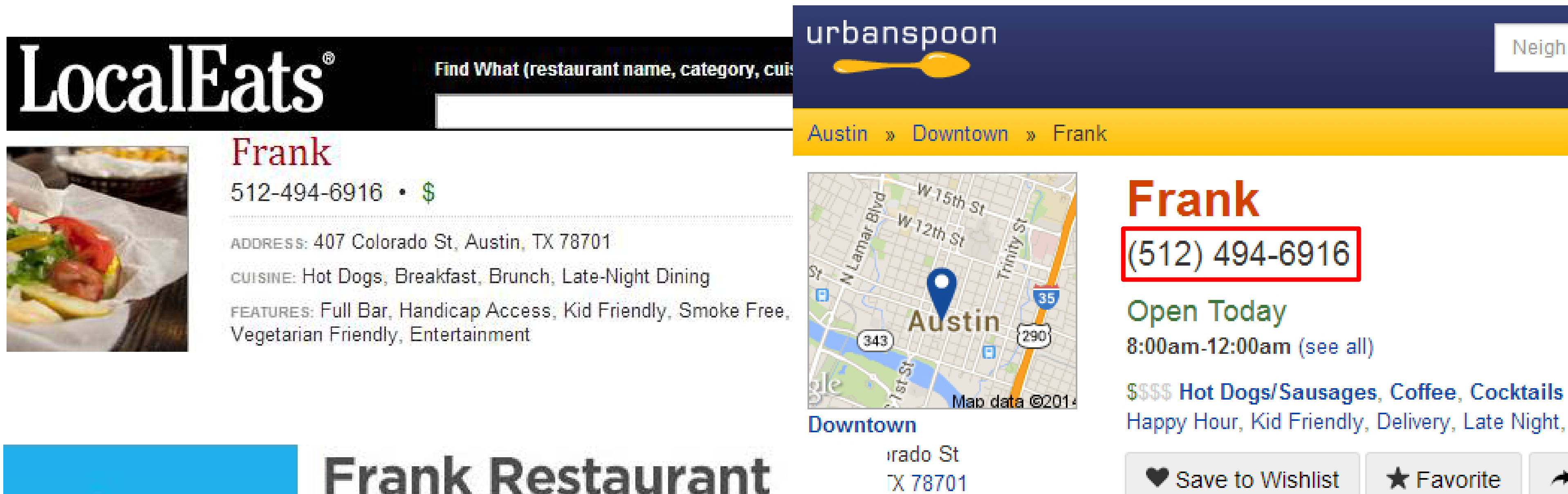
- Motivation
- Proposed Method
- Performance Study
- Conclusion

Entities in Multiple Sources



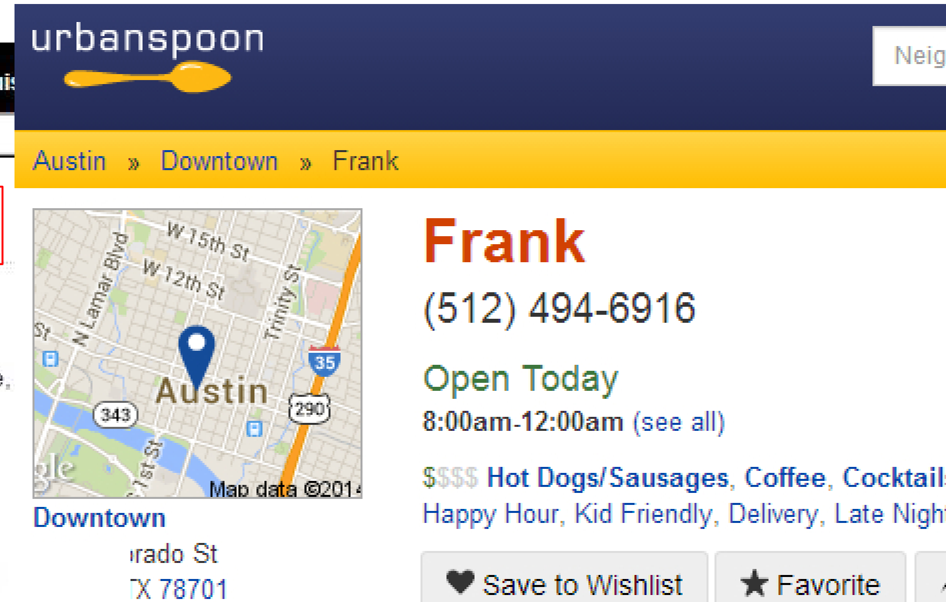
- Various name representations

Entities in Multiple Sources



- Various name representations
- Erroneous attribute values

Entities in Multiple Sources



- Various name representations
- Erroneous attribute values
- Incomplete information

Entities in Multiple Sources

- Various name representations
- Erroneous attribute values
- Incomplete information
- Ambiguous references

What We Want

Name	Address	Phone	Cuisine	Recommend	Price	Weekday Hours	Weekend Hours	Rating	Source
Frank	407 Colorado St	512-494-6894		Hot dog		Normal[1]	Extend[2]	8.7	Urbanspoon[4]
Frank	407 Colorado St	512-494-6916			$	Normal	Normal	6.0	FindMeGF[5]
Frank Restaurant	btw 4th & 5th	512-494-6894			$	Normal	Extend	9.4	Foursquare[6]
Frank	407 C								LocalEats[7]
Frank	407 C					al	Extend	8.2	Yelp[8]
Frank	407 C							8.2	TripAdvisor[9]
Frank&Angie's Pizzeria	508					al	Night[3]	8.8	Urbanspoon
Frank & Angie's	508					al	Night	8.6	Foursquare
Frank&Angie's Pizzeria	508					al	Night		LocalEats
Frank & Angie's	508					al	Night	7.0	Yelp
Frank&Angie's Pizzeria	508 West Ave	512-472-3534	Italian	Pizza	$$			8.0	Tripadvisor

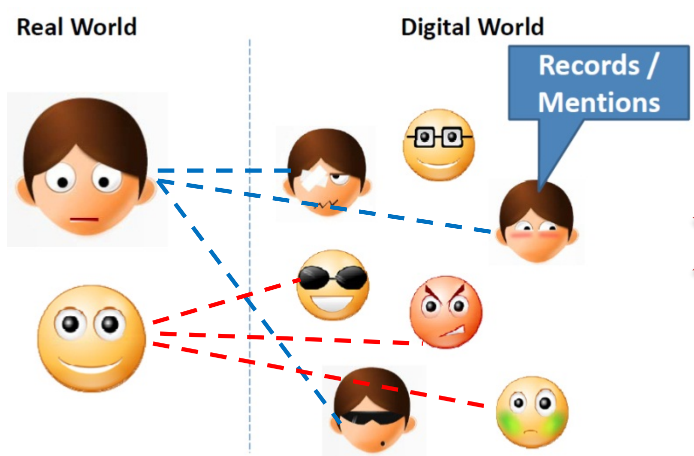
- Various name representations
- Erroneous attribute values
- Incomplete information
- Ambiguous references

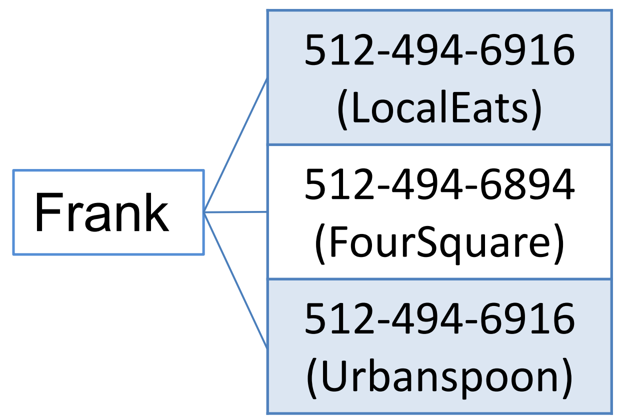
Entity Profiling

Name	Address	Phone	Cuisine	Recommend	Price	Weekday Hours	Weekend Hours	Rating
Frank Restaurant	407 Colorado St	512-494-6894	American	Hot dog	$	Normal	Extend	8.2
Frank & Angie's Pizzeria	508 West Ave	512-472-3524	Italian	Pizza	$$	Normal	Night	8.0

The Problem Involves Two Tasks

- Record linkage
  - [Getoor et al, VLDB’12], [Negahban et al, CIKM’12]
- Truth discovery
  - [Li et al, VLDB’13], [Yin et al, KDD’07]





✓ The erroneous values may prevent correct linkages
✓ Incomplete picture of the data limits the effectiveness of truth discovery

State-of-the-art Method

- [Guo et al, VLDB'10]
- Assume a set of (soft) uniqueness constraints
- Transform records into attribute value pairs

- Limitations
  - Make wrong associations when the percentage of erroneous values increases
  - Computationally expensive
  - The uniqueness constraint limits its generality

Outline

- Motivation
- Proposed Method
- Performance Study
- Conclusion

A Motivating Example

Table 1: Reference Records

	Name	Affiliation
q_1	Rakesh Agrawal	MS
q_2	Charu Aggarwal	IBM
q_3	Alon Y. Halevy	Google

matching -> profiles

p_1 =<*Rakesh Agrawal, MS, DM, Wisconsin*>
p_2 =<*Charu Aggarwal, IBM, DM, MIT*>
p_3 =<*Alon Y. Halevy, Google, DB, Stanford*>

Table 2: Input Records from Various Data Sources

	Name	Affiliation	Field	Education	Source
r_1	Rakesh Agrawal	*Bell*	DM	Wisconsin	src_1
r_2	Alon Halevy	Google	DB	Stanford	
r_3	Rakesh Agrawal	MS	DM		src_2
r_4	A. Halevy	Google	DB		
r_5	Agrawal	MS		Wisconsin	src_3
r_6	Charu Aggarwal	IBM		MIT	
r_7	Agrawal	*IBM* ?		Wisconsin √	src_4
r_8	Halevy	*UW* x	DB	Stanford √	
r_9	Charu Aggarwal	*UIC* x	DM	MIT √	
r_{10}	Agrawal	IBM	DM	*Wisconsin*	src_5

True matchings: $\{q_1, r_1, r_3, r_5, r_7\}, \{q_2, r_6, r_9, r_{10}\}, \{q_3, r_2, r_4, r_8\}$

A Motivating Example

Table 1: Reference Records

	Name	Affiliation
q_1	Rakesh Agrawal	MS
q_2	Charu Aggarwal	IBM
q_3	Alon Y. Halevy	Google

matching -> profiles

p_1 =<*Rakesh Agrawal, MS, DM, Wisconsin*>
p_2 =<*Charu Aggarwal, IBM, DM, MIT*>
p_3 =<*Alon Y. Halevy, Google, DB, Stanford*>

Table 2: Input Records from Various Data Sources

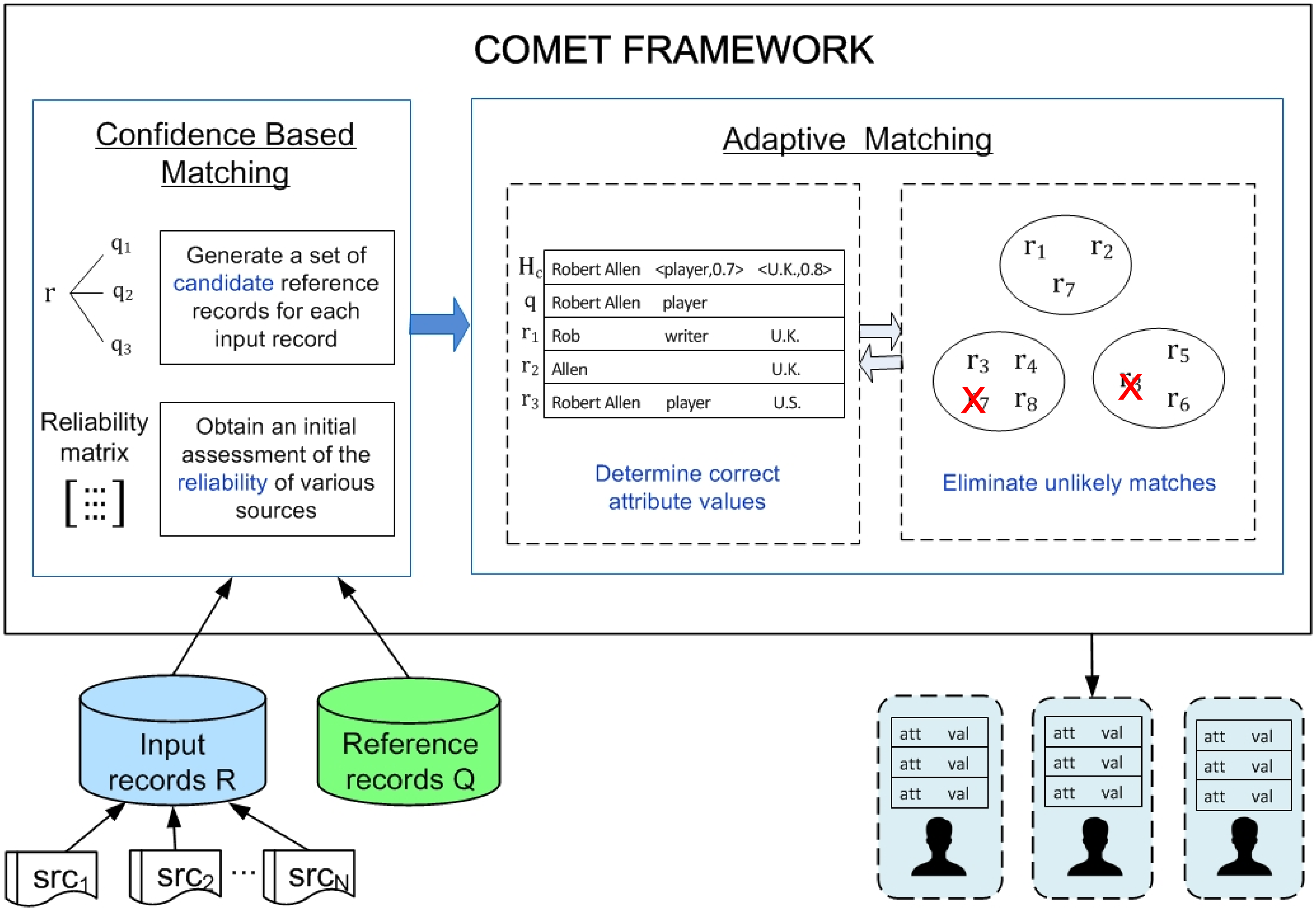
	Name	Affiliation	Field	Education	Source
r_1	Rakesh Agrawal	*Bell*	DM	Wisconsin	src_1
r_2	Alon Halevy	Google	DB	Stanford	src_1
r_3	Rakesh Agrawal	MS	DM		src_2
r_4	A. Halevy	Google	DB		src_2
r_5	Agrawal	MS		Wisconsin	src_3
r_6	Charu Aggarwal	IBM		MIT	src_3
r_7	Agrawal	*IBM* ?		Wisconsin √	src_4
r_8	Halevy	*UW* X	DB	Stanford √	src_4
r_9	Charu Aggarwal	*UIC* X	DM	MIT √	src_4
r_{10}	Agrawal	IBM	DM	*Wisconsin*	src_5

True matchings: $\{q_1, r_1, r_3, r_5, r_7\}$, $\{q_2, r_6, r_9, r_{10}\}$, $\{q_3, r_2, r_4, r_8\}$

The Example Tells Us

- The data sources are not equally reliable among different attributes
  - Introduce a reliability matrix $M[s, a]$
  - Lower the impact of erroneous values on matching decisions
- Rectifying errors in attribute values provides additional evidence for linking records
  - Interleave the processes of record linkage and error correction

The Proposed Two-phase Method



Outline

- Motivation
- Proposed Method
- Performance Study
- Conclusion

Comparative Methods

- PIPELINE
  - Record linkage [1] + Truth discovery [2]
- MATCH [3]
  - State-of-the-art method
- COMET
  - The proposed method

1. Negahban et al. Scaling multiple-source entity resolution using statistically efficient transfer learning. In CIKM, 2012
2. Yin et al. TruthFinder. In KDD, 2007
3. Guo et al. Record linkage with uniqueness constraints and erroneous values. VLDB, 2010

Results on Restaurant Dataset

Table 5: Record Linkage on Restaurant Dataset

	Precision	Recall
COMET	**96.6**	**96.6**
MATCH	93.0	88.1
PIPELINE	89.1	83.5

Table 6: Truth Discovery on Restaurant Dataset

	Accuracy	Coverage
COMET	**86.4**	**83.2**
MATCH	75.3	76.8
PIPELINE	82.3	71.2

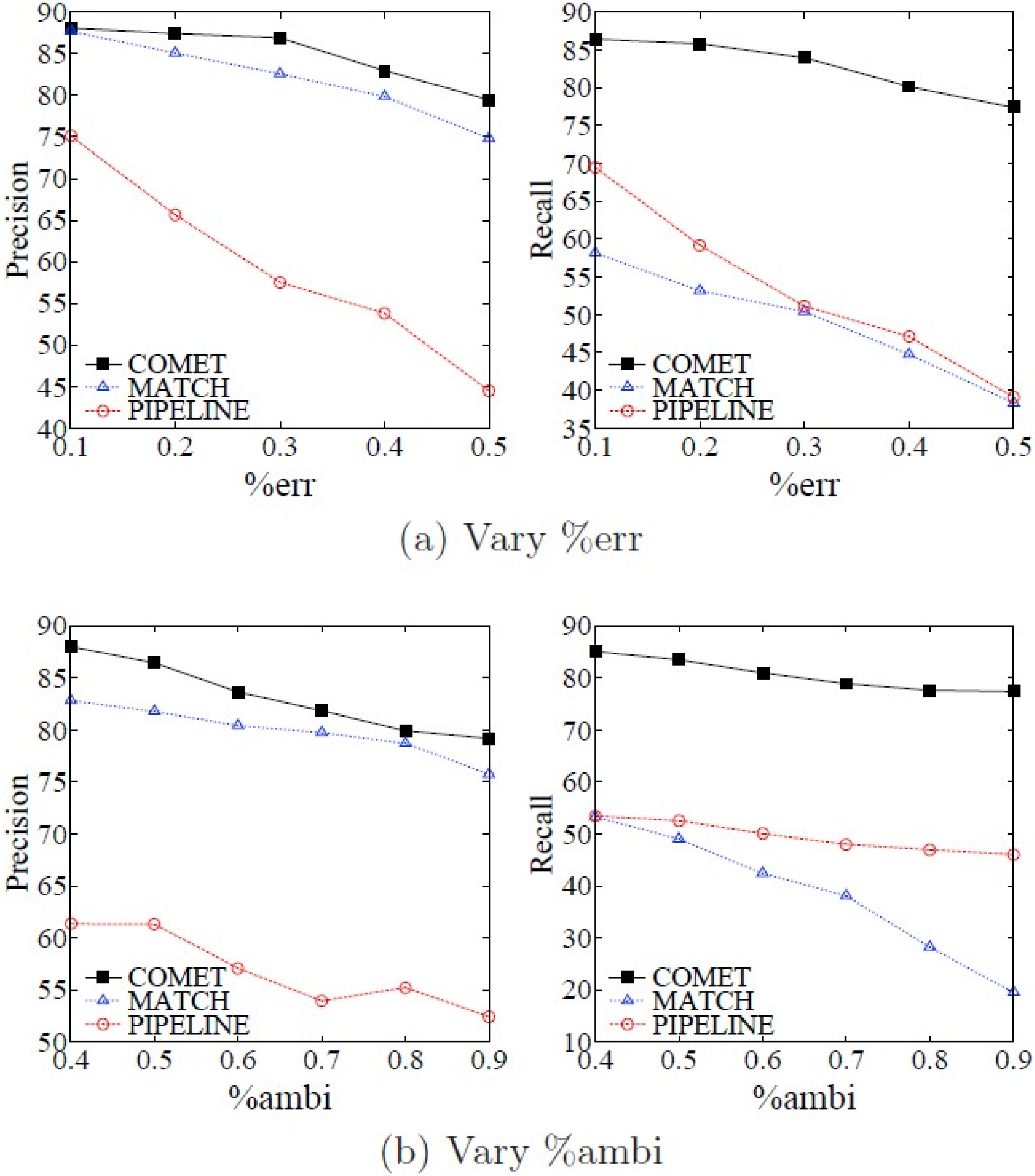


Figure 2: Record linkage on Football dataset

- %err: percentage of erroneous values
- %ambi: percentage of records with abbreviated names

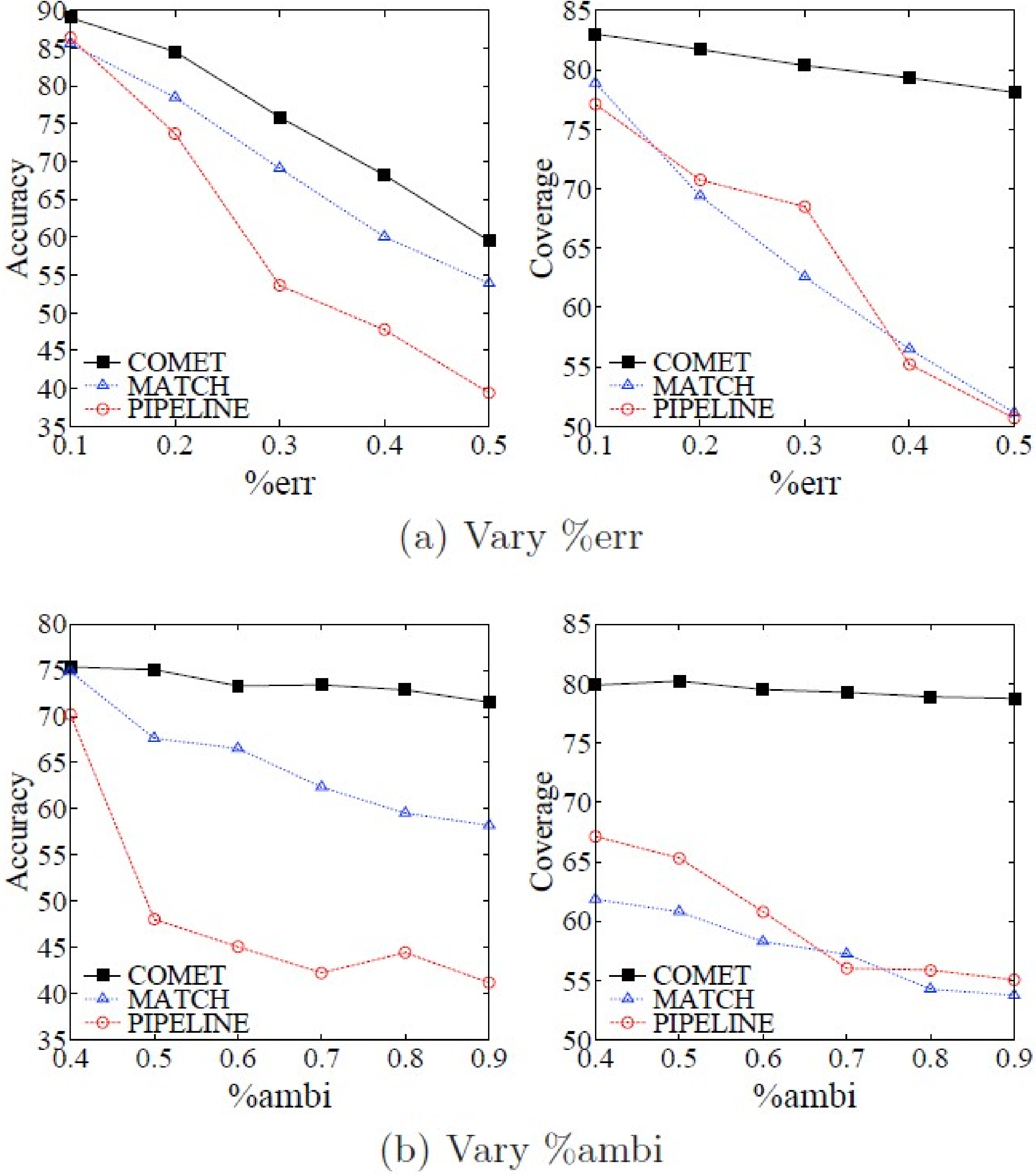
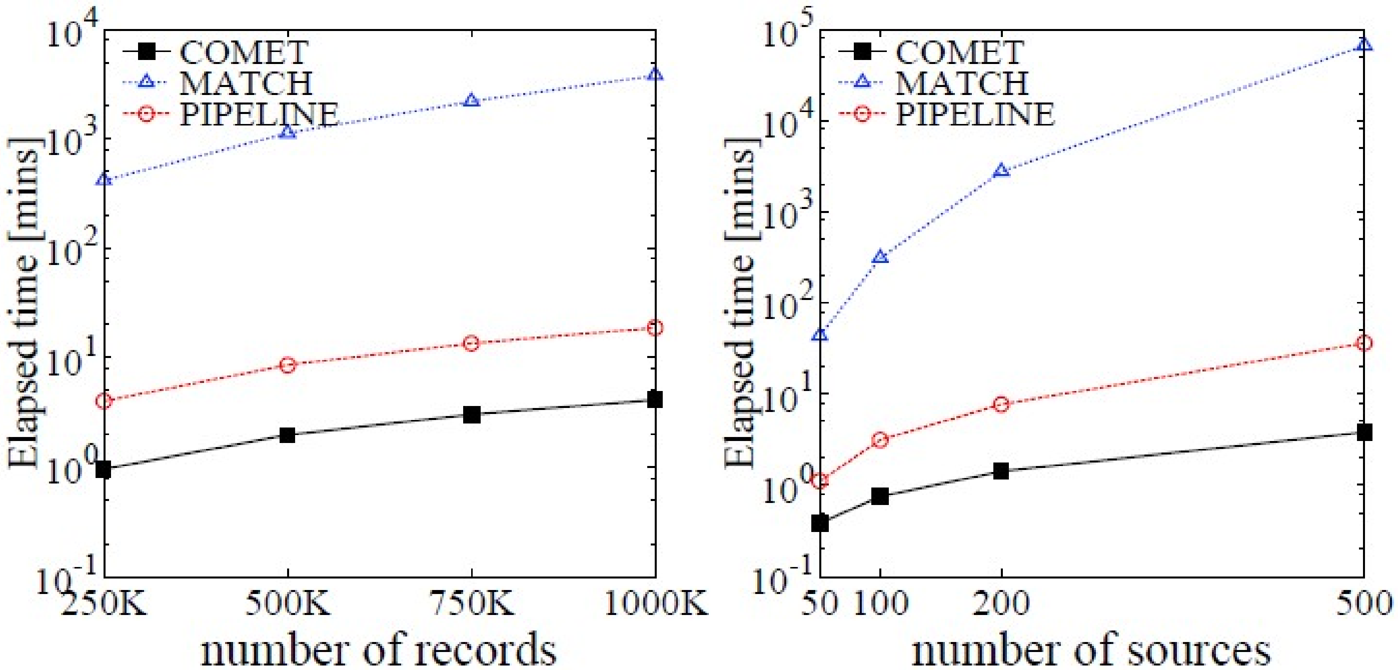


Figure 3: Truth discovery on Football dataset

- %err: percentage of erroneous values
- %ambi: percentage of records with abbreviated names

Scalability Experiments



(a) Vary number of entities (b) Vary number of sources

Figure 7: Scalability results

Outline

- Motivation
- Proposed Method
- Performance Study
- Conclusion

Conclusion

- Address the problem of building entity profiles by collating data records from multiple sources in the presence of erroneous values
- Interleave record linkage with truth discovery
- Varying source reliabilities
- Reduce the impact of erroneous values on matching decisions

Thanks!

Q & A

furongli@comp.nus.edu.sg

Confidence Based Matching

- Bootstrap the framework with a small set of confident matches
- Initialize reliability matrix based on confident matches

q_1	Rakesh Agrawal	MS			
r_1	Rakesh Agrawal	*Bell*	DM	Wisconsin	src_1
r_3	Rakesh Agrawal	MS	DM		src_2
q_2	Charu Aggarwal	IBM			
r_6	Charu Aggarwal	IBM		MIT	src_3
r_9	Charu Aggarwal	*UIC*	DM	MIT	src_4
q_3	Alon Y. Halevy	Google			
r_2	Alon Halevy	Google	DB	Stanford	src_1

$$
\begin{aligned}
M[src_1, \text{Affiliation}] &= 0.5, \\
M[src_2, \text{Affiliation}] &= 1.0, \\
M[src_3, \text{Affiliation}] &= 1.0, \\
M[src_4, \text{Affiliation}] &= 0.2, \\
M[src_5, \text{Affiliation}] &= \epsilon.
\end{aligned}
$$

Confidence Based Matching

- Distinguish sources
  - Reliable: $\{src_1, src_2, src_3\}$
  - $\{r_5, r_4\}$
  - Unreliable: $\{src_4, src_5\}$
  - $\{r_7, r_8, r_{10}\}$

q_1	Rakesh Agrawal	MS			
r_1	Rakesh Agrawal	*Bell*	DM	Wisconsin	src_1
r_3	Rakesh Agrawal	MS	DM		src_2

q_2	Charu Aggarwal	IBM			
r_6	Charu Aggarwal	IBM		MIT	src_3
r_9	Charu Aggarwal	*UIC*	DM	MIT	src_4

q_3	Alon Y. Halevy	Google			
r_2	Alon Halevy	Google	DB	Stanford	src_1

Confidence Based Matching

- Distinguish sources
  - Reliable: $\{src_1, src_2, src_3\}$
  - $\{r_5, r_4\}$
  - Unreliable: $\{src_4, src_5\}$
  - $\{r_7, r_8, r_{10}\}$

q_1	Rakesh Agrawal	MS			
r_1	Rakesh Agrawal	*Bell*	DM	Wisconsin	src_1
r_3	Rakesh Agrawal	MS	DM		src_2
r_5	Agrawal	MS		Wisconsin	src_3

q_2	Charu Aggarwal	IBM			
r_6	Charu Aggarwal	IBM		MIT	src_3
r_9	Charu Aggarwal	*UIC*	DM	MIT	src_4
r_5	Agrawal	MS		Wisconsin	src_3

q_3	Alon Y. Halevy	Google			
r_2	Alon Halevy	Google	DB	Stanford	src_1
r_4	A. Halevy	Google	DB		src_2

Confidence Based Matching

- Distinguish sources
  - Reliable: $\{src_1, src_2, src_3\}$
  - $\{r_5, r_4\}$
  - Unreliable: $\{src_4, src_5\}$
  - $\{r_7, r_8, r_{10}\}$

q_1	Rakesh Agrawal	MS			
r_1	Rakesh Agrawal	*Bell*	DM	Wisconsin	src_1
r_3	Rakesh Agrawal	MS	DM		src_2
r_5	Agrawal	MS		Wisconsin	src_3
r_7	Agrawal	*IBM*		Wisconsin	src_4
r_{10}	Agrawal	IBM	DM	*Wisconsin*	src_5

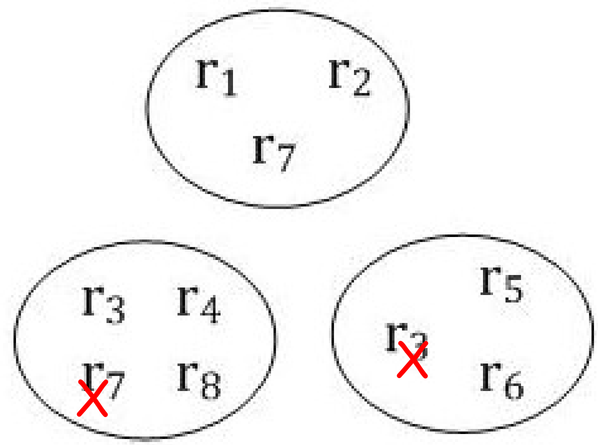
q_2	Charu Aggarwal	IBM			
r_6	Charu Aggarwal	IBM		MIT	src_3
r_9	Charu Aggarwal	*UIC*	DM	MIT	src_4
r_5	Agrawal	MS		Wisconsin	src_3
r_7	Agrawal	*IBM*		Wisconsin	src_4
r_{10}	Agrawal	IBM	DM	*Wisconsin*	src_5

q_3	Alon Y. Halevy	Google			
r_2	Alon Halevy	Google	DB	Stanford	src_1
r_4	A. Halevy	Google	DB		src_2
r_8	Halevy	*UW*	DB	Stanford	src_4

Adaptive Matching

H_c	Robert Allen	<player,0.7>	<U.K.,0.8>
q	Robert Allen	player	
r_1	Rob	writer	U.K.
r_2	Allen		U.K.
r_3	Robert Allen	player	U.S.

Cluster signatures <- current truth
Update reliability matrix



Refine clusters by pruning the most dissimilar records

$$\text{match}(r, c) = \frac{\sum_{a \in \mathcal{A}} \boxed{M[s_r, a]} \cdot \text{sim}(r.a, H_c.a)}{\sum_{a \in \mathcal{A}} M[s_r, a]}$$

Discount records belonging to multiple clusters

Lower the impact of erroneous values on our matching decision