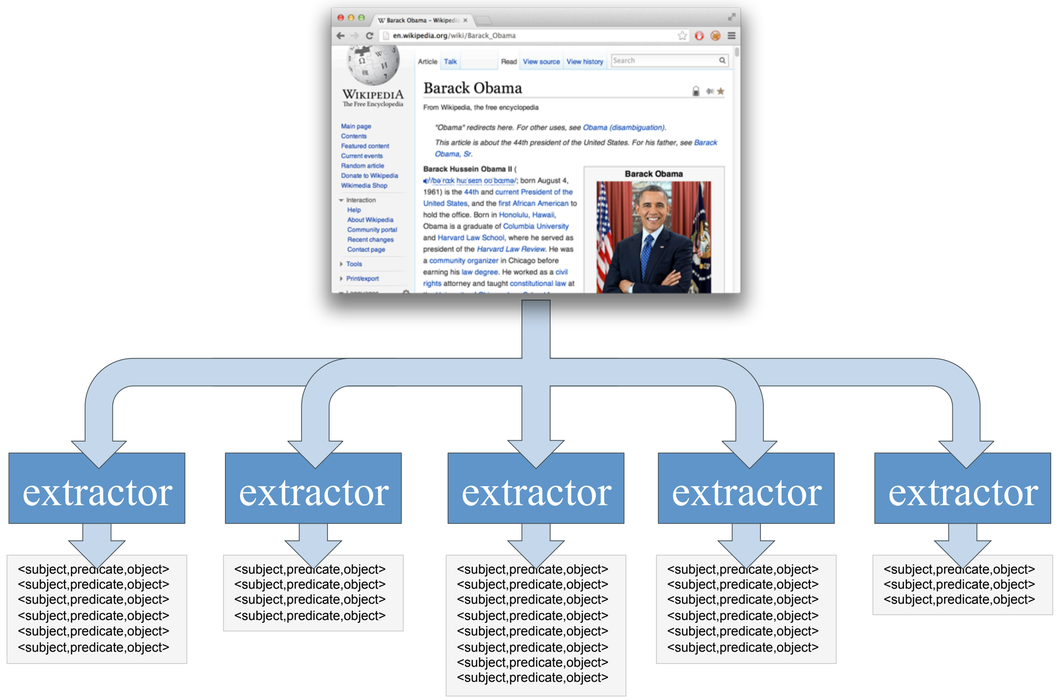
Fusing Data with Correlations

Ravali Pochampally — *UMass Amherst*
Anish Das Sarma — *Troo.ly*
Luna Dong — *Google*
Alexandra Meliou — *UMass Amherst*
Divesh Srivastava — *AT&T Research*



Knowledge base

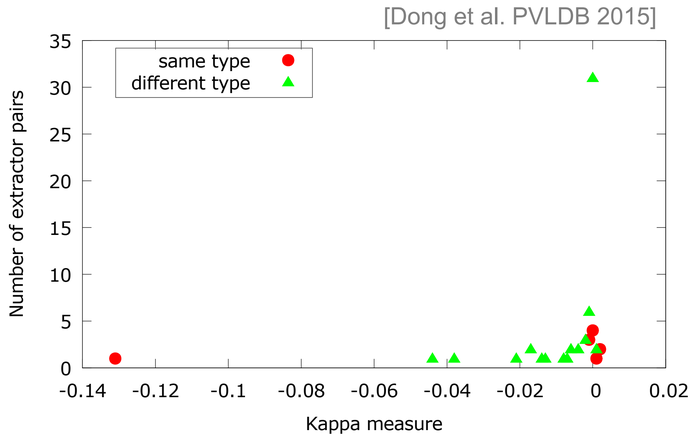
anti-correlated

ID	KnowledgeTriple	S_1	S_2	S_3	S_4	S_5	
t_1	$\{Obama, profession, president\}$	✓	✓		✓	✓	
t_2	$\{Obama, died, 1982\}$	✓	✓				✗
t_3	$\{Obama, profession, lawyer\}$			✓			
t_4	$\{Obama, religion, Christian\}$		✓	✓	✓	✓	
t_5	$\{Obama, age, 50\}$		✓	✓			✗
t_6	$\{Obama, support, White\ Sox\}$	✓			✓	✓	
t_7	$\{Obama, spouse, Michelle\}$	✓	✓	✓			
t_8	$\{Obama, administered\ by, John\ G.\ Roberts\}$	✓	✓		✓	✓	✗
t_9	$\{Obama, surgical\ operation, 05/01/2011\}$	✓	✓		✓	✓	✗
t_{10}	$\{Obama, profession, community\ organizer\}$	✓		✓	✓	✓	

correlated

Correlations in web extraction

[Dong et al. PVLDB 2015]

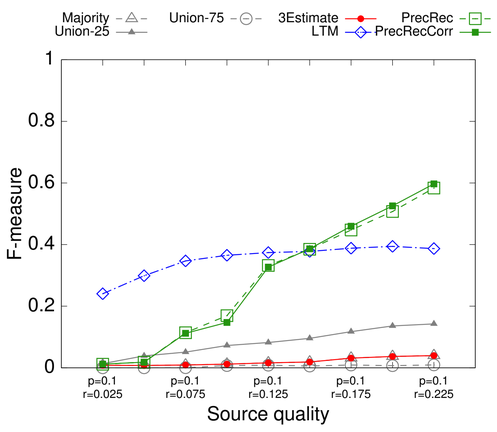


Correlations are richer than copying relationships

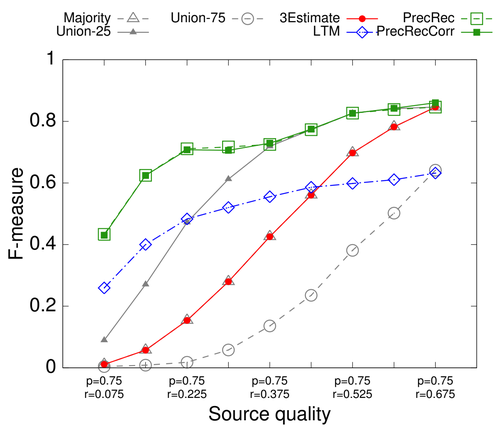
There is an exponential number of correlation parameters, but we provide a scalable solution

Synthetic datasets

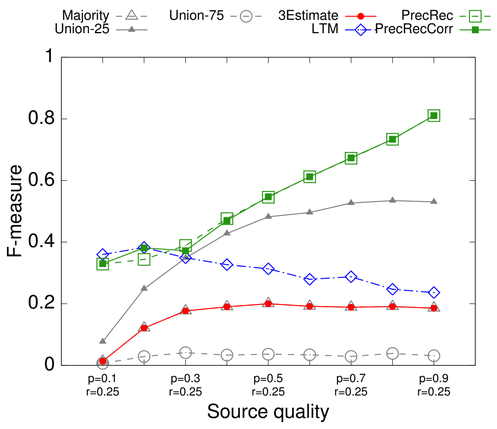
low precision



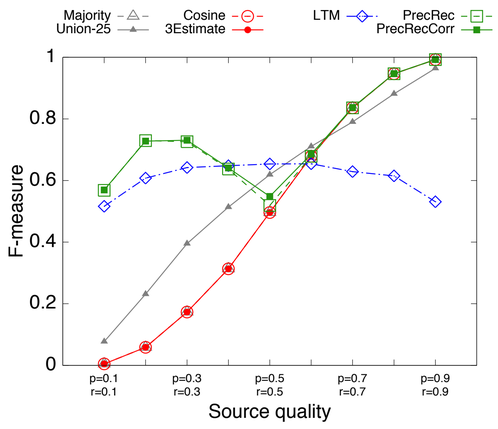
high precision



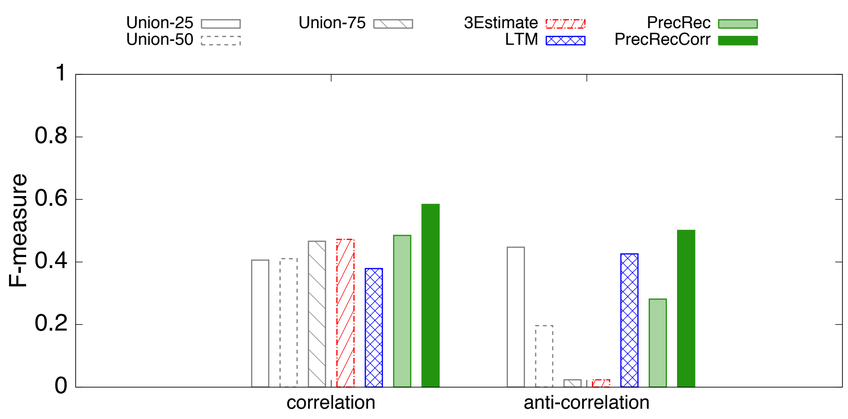
low recall



bad to good quality

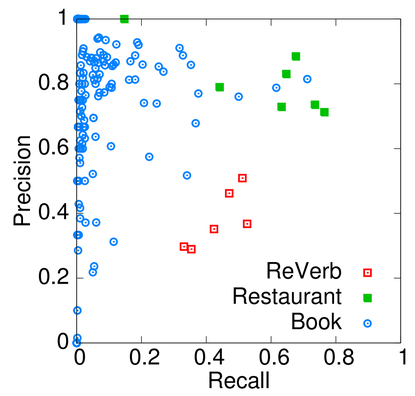


correlation



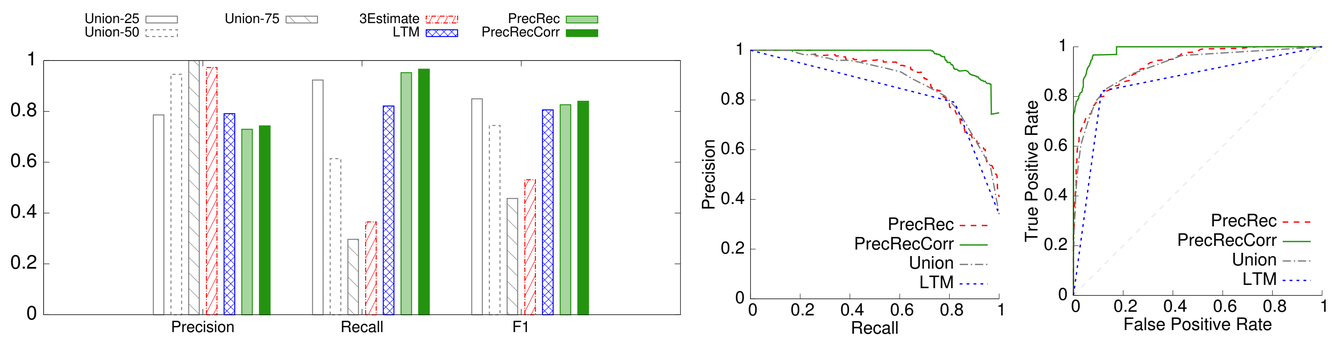
Real-world datasets

- Union-K
  - –a triple is true if at least K sources provide it
- 3-Estimate [Galland et al. WSDM 2010]
  - –iteratively computes trustworthiness
- LTM [Zhao et al. PVLDB 2012]
  - –uses graphical models and Gibbs sampling
- PrecRec / PrecRecCorr — our methods

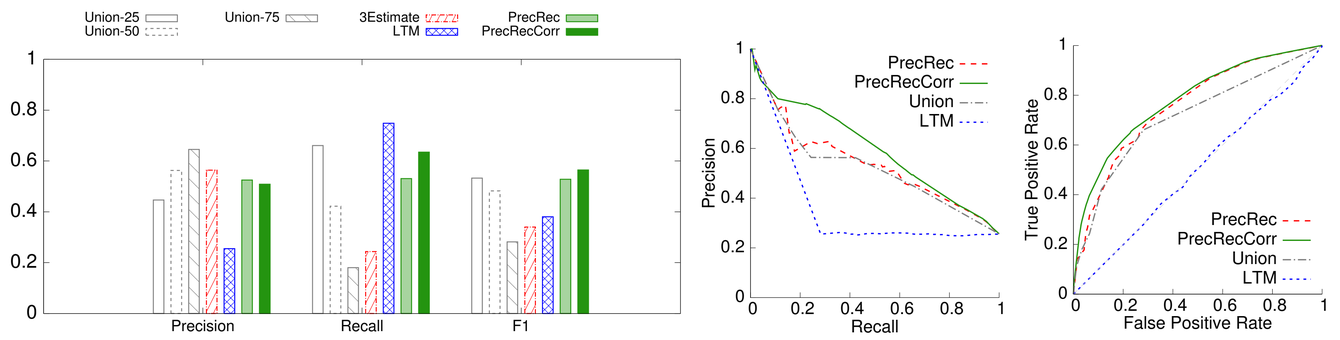


Restaurant: 7 sources, 93 triples

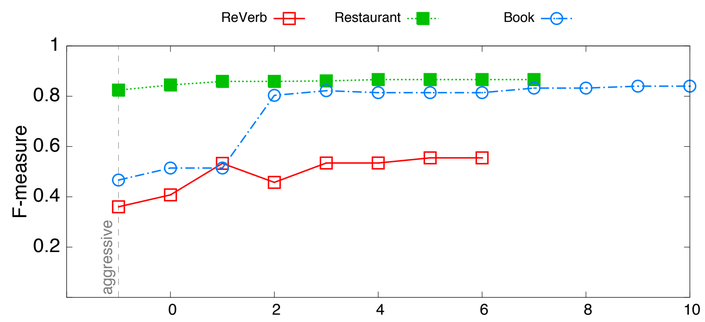
Book: 879 sources, 225 triples



ReVerb: 6 extractors, 2407 triples



elastic approximation



runtime comparison

time(sec)	ReVerb	Restaurant	Book
Union-25	0.39	0.56	3.86
Union-50	0.14	0.32	3.71
Union-75	0.11	0.35	3.00
3-Estimate	0.7	0.06	39
LTM (10 iter)	49	5.3	3791
PrecRec	2.6	0.3	35
PrecRecCorr	124	5.4	6786
PrecRecCorr-lvl3	79	2.25	2452