Leveraging Data Relationships to Resolve Conflicts from Disparate Data Sources

Romila Pradhan, Walid G. Aref, Sunil Prabhakar

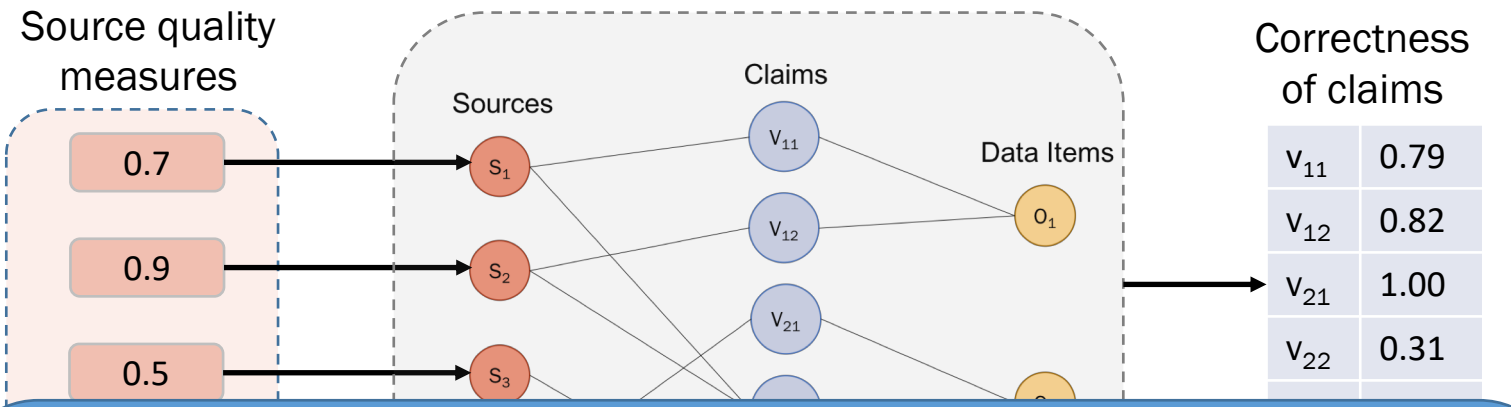
PURDUE UNIVERSITY

Fusing data from multiple sources

Data Item	S_1	S_2	S_3	S_4	S_5
Basera		745 9th Avenue	357 East 50th St.	Midtown East	
Alto	New York City	520 Madison Avenue	11 East 53rd St.	Midtown East	East 50s
A voce	New York	41 Madison Avenue	Flatiron/Union Square	Gramercy/ Flatiron	Flatiron

General principle of data fusion

Sources determine the correctness of claims

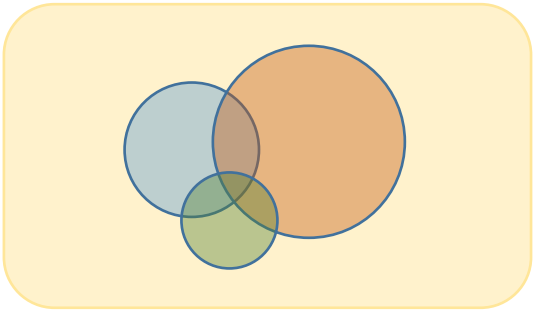


Failing to acknowledge claim relationships has been observed to account for as much as **35%** of **false negatives** in data fusion tasks[1].

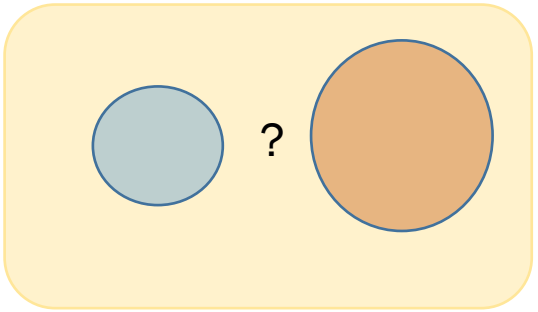
1. X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang, "From data fusion to knowledge fusion". PVLDB, 2014.

This talk

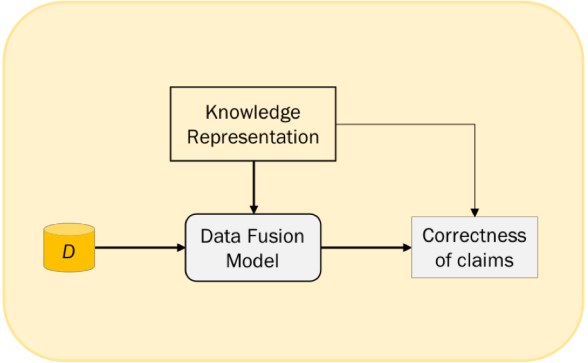
Observed relations



Representing claim relationships



Integration with data fusion models



Evaluation

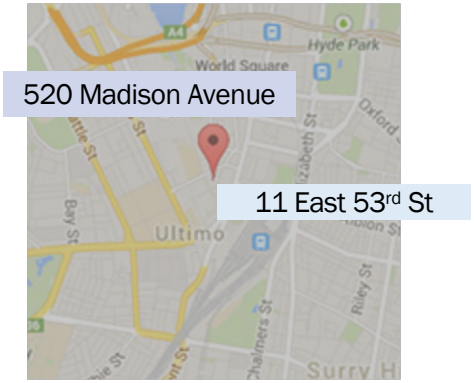
Relationships observed in data

1. Claims could be related through rich semantics, e.g., hierarchically
   - 'New York City' is a city in 'New York' state
   - 'Midtown East', 'Flatiron', 'Union Square', 'Gramercy' are neighborhoods in 'New York City'
   - 'Flatiron' is a part of the 'Flatiron/Union Square' and 'Gramercy/Flatiron' areas

Data Item	Claim
Basera	*745 9th Avenue*
Basera	*357 East 50th St*
Basera	*Midtown East*
Alto	*520 Madison Avenue*
Alto	*11 East 53rd St*
Alto	*Midtown East*
Alto	*East 50s*
Alto	*New York City*
Alto	*New York*
A voce	*41 Madison Avenue*
A voce	*Flatiron*
A voce	*Flatiron/Union Square*
A voce	*Gramercy/Flatiron*
A voce	*New York City*
A voce	*New York*

Relationships observed in data

2. There could be alternate representations for the same claim



Represent the same physical location!

Data Item	Claim
Basera	*745 9th Avenue*
Basera	*357 East 50th St*
Basera	*Midtown East*
Alto	*520 Madison Avenue*
Alto	*11 East 53rd St*
Alto	*Midtown East*
Alto	*East 50s*
Alto	*New York City*
Alto	*New York*
A voce	*41 Madison Avenue*
A voce	*Flatiron*
A voce	*Flatiron/Union Square*
A voce	*Gramercy/Flatiron*
A voce	*New York City*
A voce	*New York*

Integrity constraints on claim correctness

3. Correctness of a claim may support that of another or, contradict another claim
   - If '745 9th Avenue' is not a part of the 'Midtown East' neighborhood, then both the claims cannot be be simultaneously correct
   - If 'Flatiron' is correct, 'Flatiron/Union Square' and 'Gramercy/Flatiron' should also be correct

Data Item	Claim
Basera	*745 9th Avenue*
Basera	*357 East 50th St*
Basera	*Midtown East*
Alto	*520 Madison Avenue*
Alto	*11 East 53rd St*
Alto	*Midtown East*
Alto	*East 50s*
Alto	*New York City*
Alto	*New York*
A voce	*41 Madison Avenue*
A voce	*Flatiron*
A voce	*Flatiron/Union Square*
A voce	*Gramercy/Flatiron*
A voce	*New York City*
A voce	*New York*

Likelihood of claim correctness

4. Correctness probability produced by data fusion models may not reflect the true likelihood of a claim being correct.

e.g., using TruthFinder[2], for item "*A voce*",

P(41 Madison Avenue = true) = 0.97

P(New York = true) = 0.88

However, in the ideal case,

P(New York) > P(41 Madison Ave)

Data Item	Claim
Basera	*745 9th Avenue*
Basera	*357 East 50th St*
Basera	*Midtown East*
Alto	*520 Madison Avenue*
Alto	*11 East 53rd St*
Alto	*Midtown East*
Alto	*East 50s*
Alto	*New York City*
Alto	*New York*
A voce	*41 Madison Avenue*
A voce	*Flatiron*
A voce	*Flatiron/Union Square*
A voce	*Gramercy/Flatiron*
A voce	*New York City*
A voce	*New York*

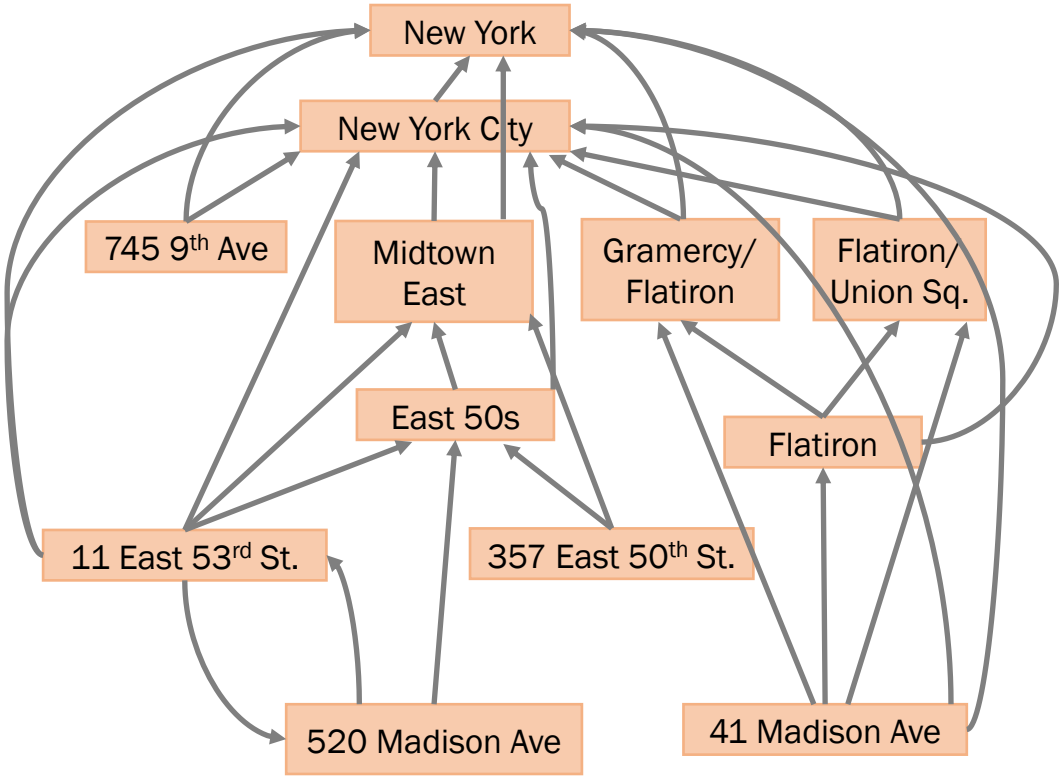
2. X. Yin, J. Han, and P. S. Yu, "Truth discovery with multiple conflicting information providers on the web," *TKDE*, 2008.

How do existing fusion models consider claim relationships ?

Single Truth	Multiple Truths
Each data item has one correct claim	Data items can have more than one correct claims. The correct claims do not need to be related.
Claims are often independent of each other but could be “similar” to other claims Implication/similarity based on ad hoc measures, e.g., Jaccard index, edit distance, numerical tolerance	Claims are independent of each other

Arbitrary directed graphs to represent claim relationships

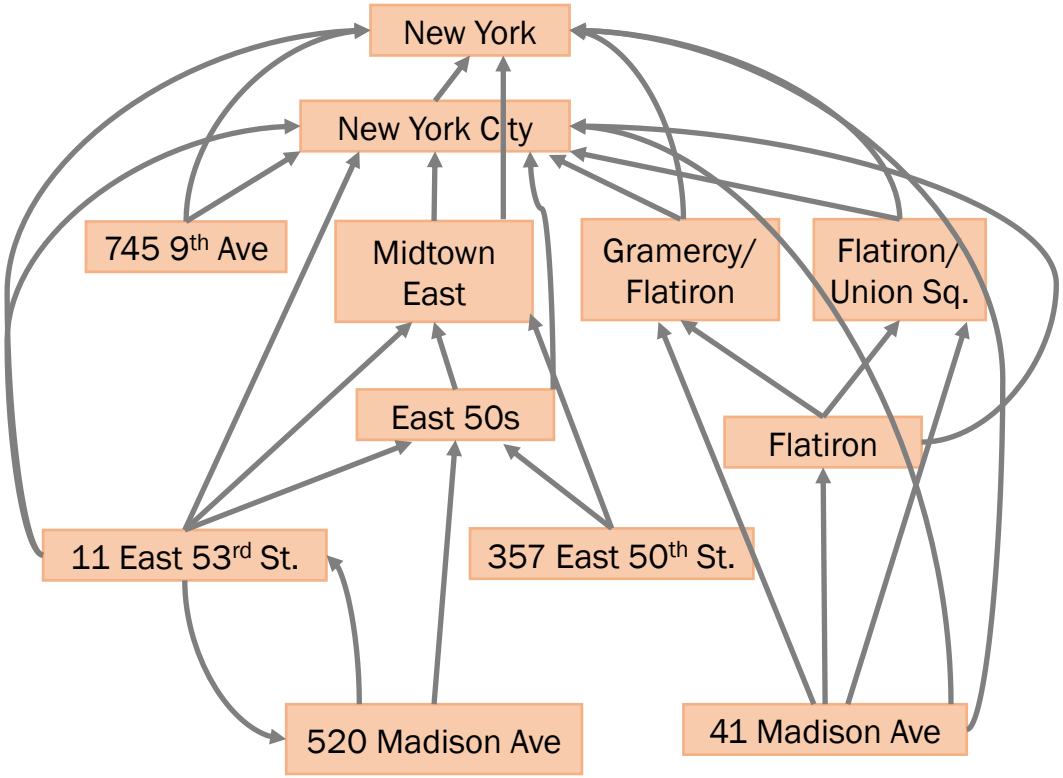
- Nodes represent claims
- Edge from a specific claim to a general claim



Directed graph in the context of data fusion

Relationships:

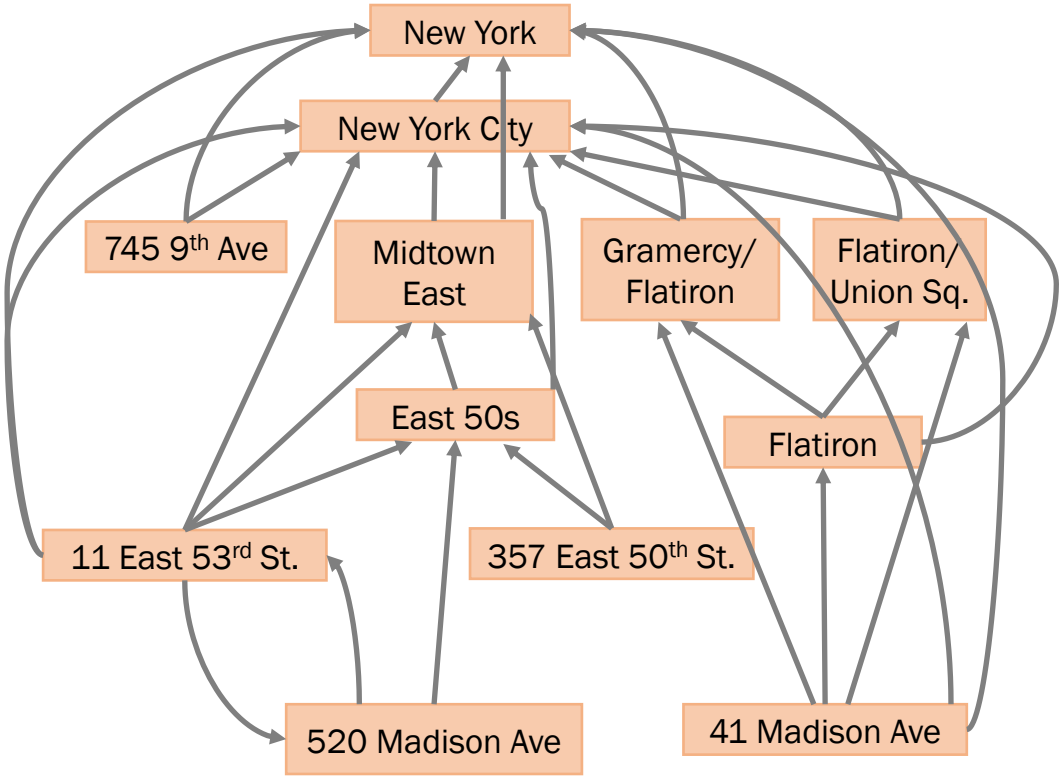
- Subsumption
- Equivalence
- Mutual Exclusion
- Overlaps between claims



Notion of support among claims

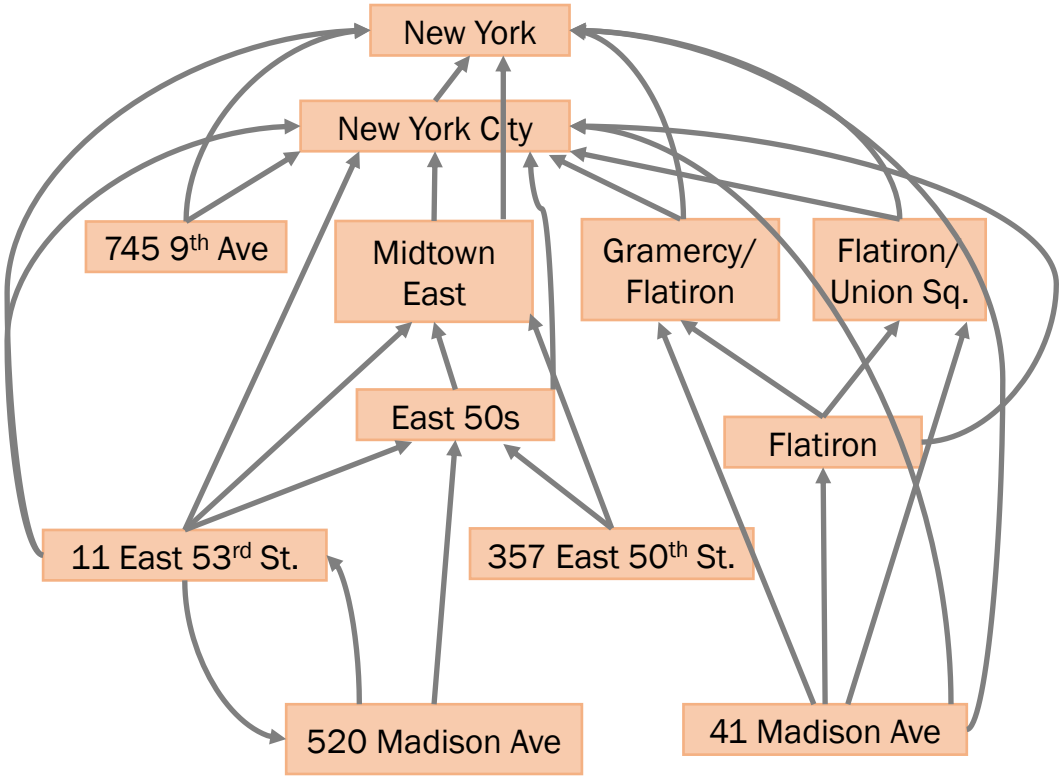
Expressed through reachability:

- A claim is supported by all claims that can reach it
- A claim supports all claims that are reachable from it

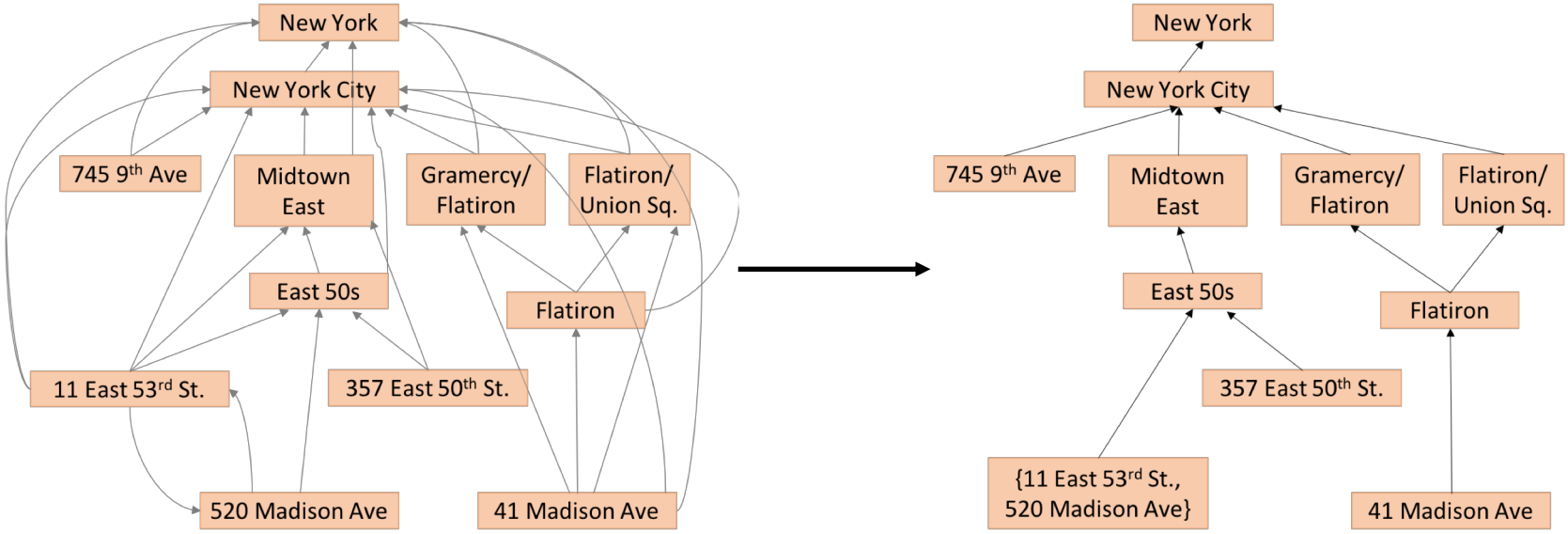


Pre-processing the graph for effective representation

- Transitive reduction
  - Remove redundant edges
  - By using the property of transitivity
- Condensation
  - Represent alternative representations
  - By identifying strongly connected components
- Final result: Directed Acyclic Graph (DAG)



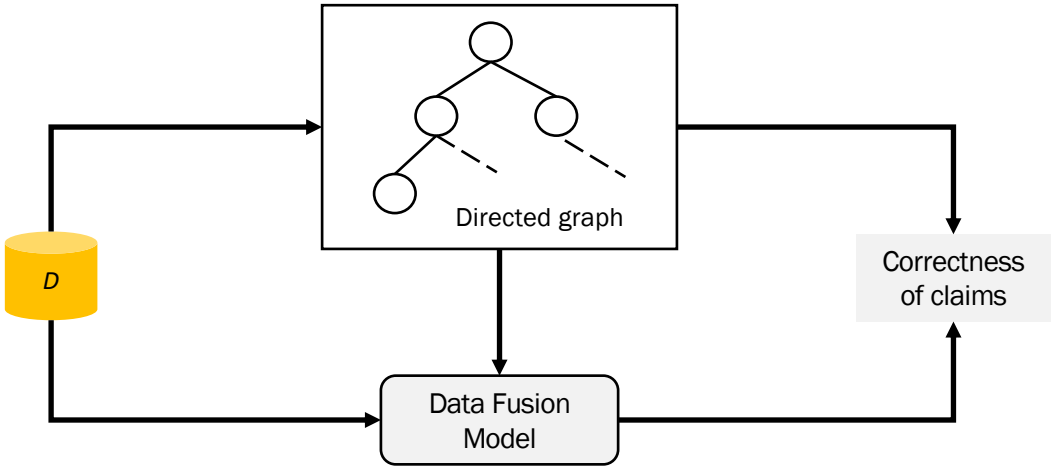
Pre-processing the graph for effective representation



Original directed graph

Directed acyclic graph without redundant edges and with collapsed nodes

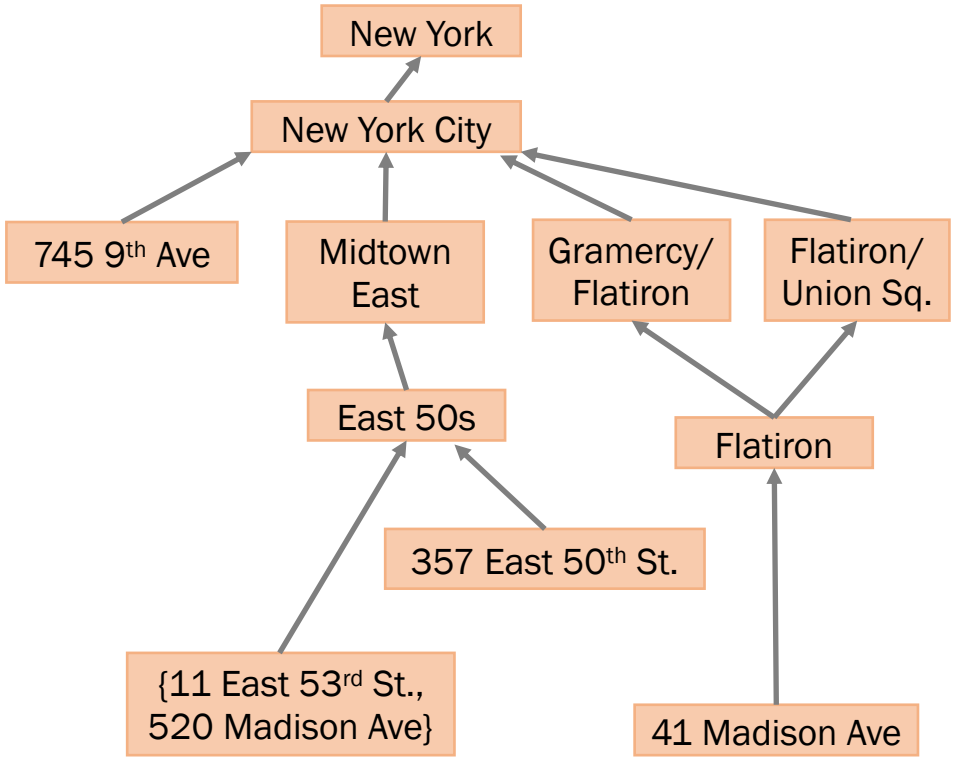
How to integrate directed graphs with data fusion models?



Source properties change

A source that provides claim "A" now *implicitly* supports claims that "A" supports

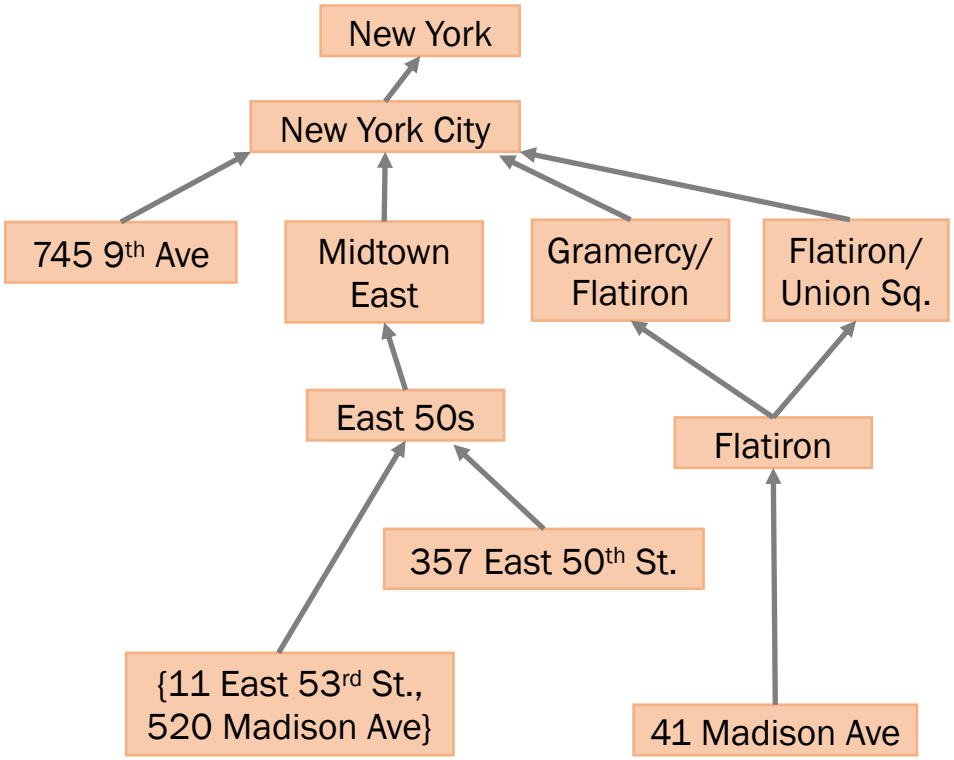
- e.g. the accuracy of source S_2 that provides "41 Madison Ave" will depend on the correctness of claims "41 Madison Ave", "Flatiron", "Gramercy/Flatiron", "Flatiron/Union Sq.", "New York"



Correctness probabilities of claims change

The correctness of a claim is affected by sources that provide claims supporting it

- e.g., correctness of claim "Flatiron" depends on source S_5 and also depends on source S_2 that provides claim "41 Madison Ave"



Modified Data Fusion

Input: Database *D*, directed acyclic graph *G*, data fusion model *F*

Output: *P* correctness probabilities of claims

1. Estimate quality of sources

   Q_F = EstimateSourceQuality(*D,G,F,P*)

2. Estimate correctness of claims

   P = EstimateClaimCorrectness(*D, G, F,* Q_F)

Determining correct claims

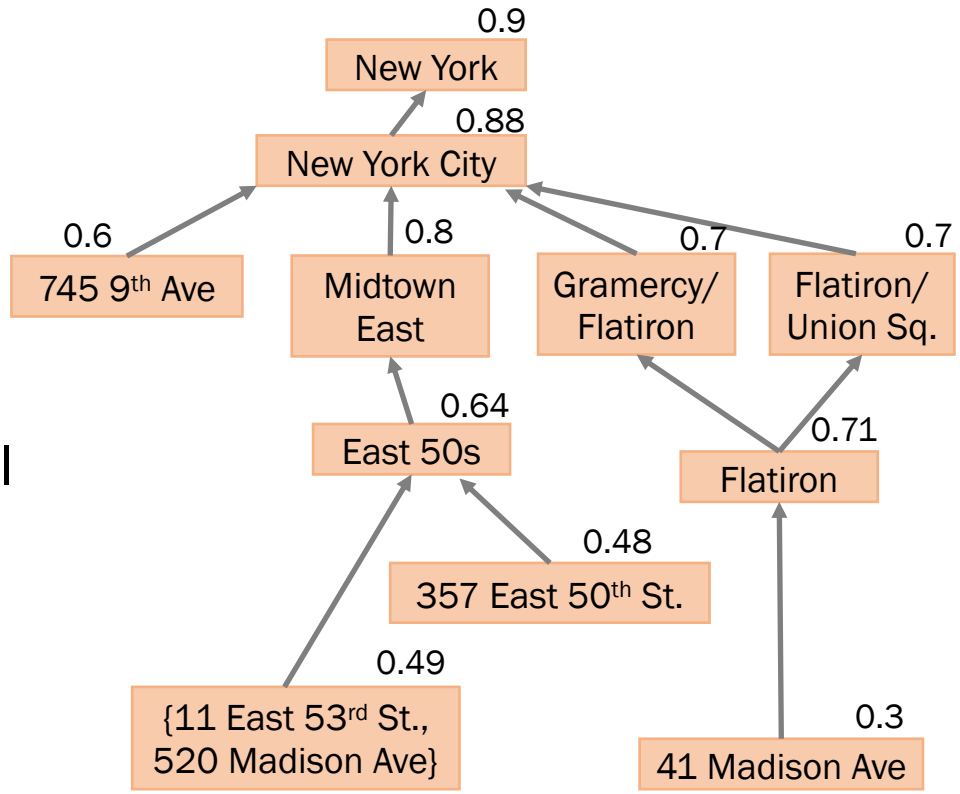
Given the correctness probabilities of claims,

- Single-truth fusion models will consider the claim with the highest probability as correct.

  Lose granularity

- Multi-truth fusion models will consider claims having probability greater than a threshold correct.

  May generate claims that are inconsistent

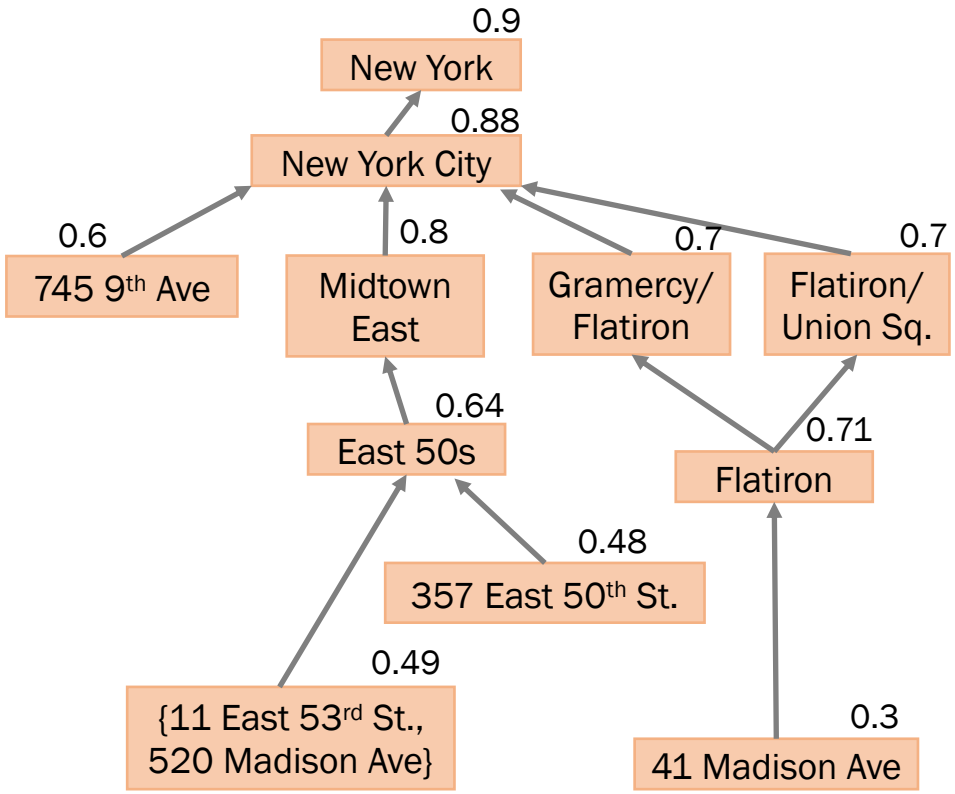


Determining correct claims

Top-down approach where we follow the *maximum probability path*

1. Identify root nodes
2. Select the claim with the highest probability of being correct
3. Follow the path from the selected node down the hierarchy and repeat

Output: Consistent claims

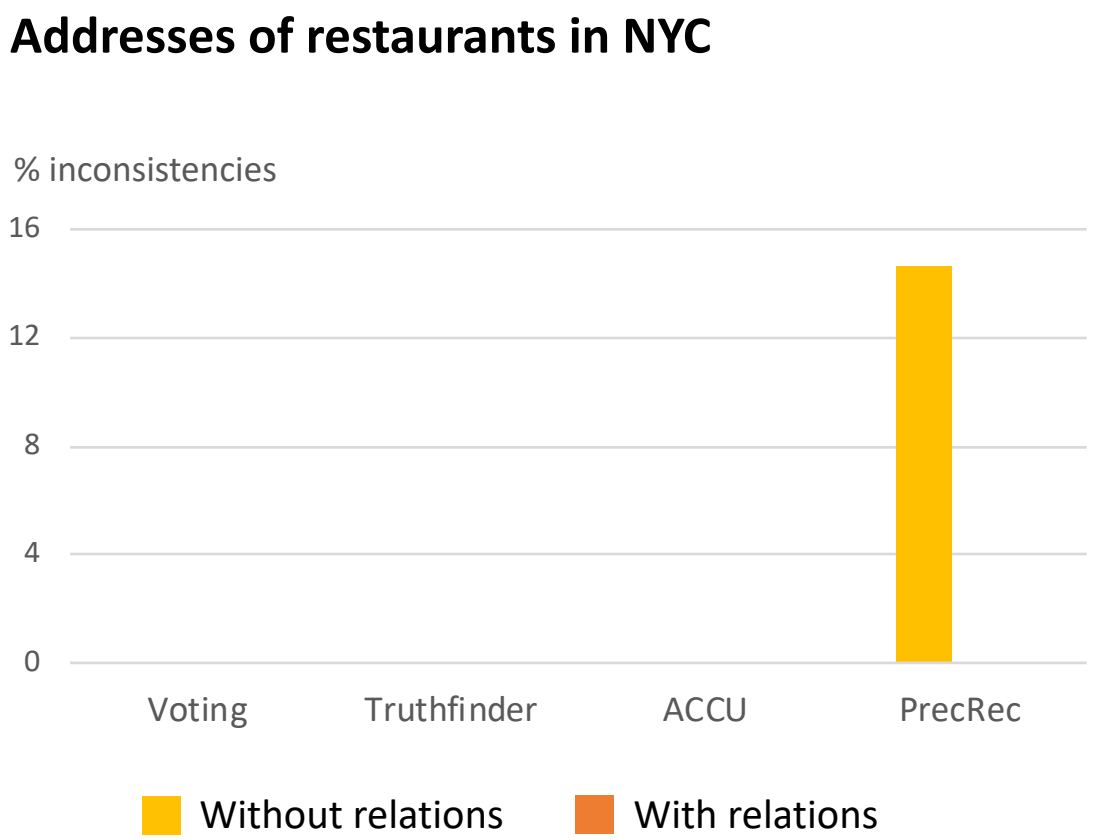


Experimental Setup

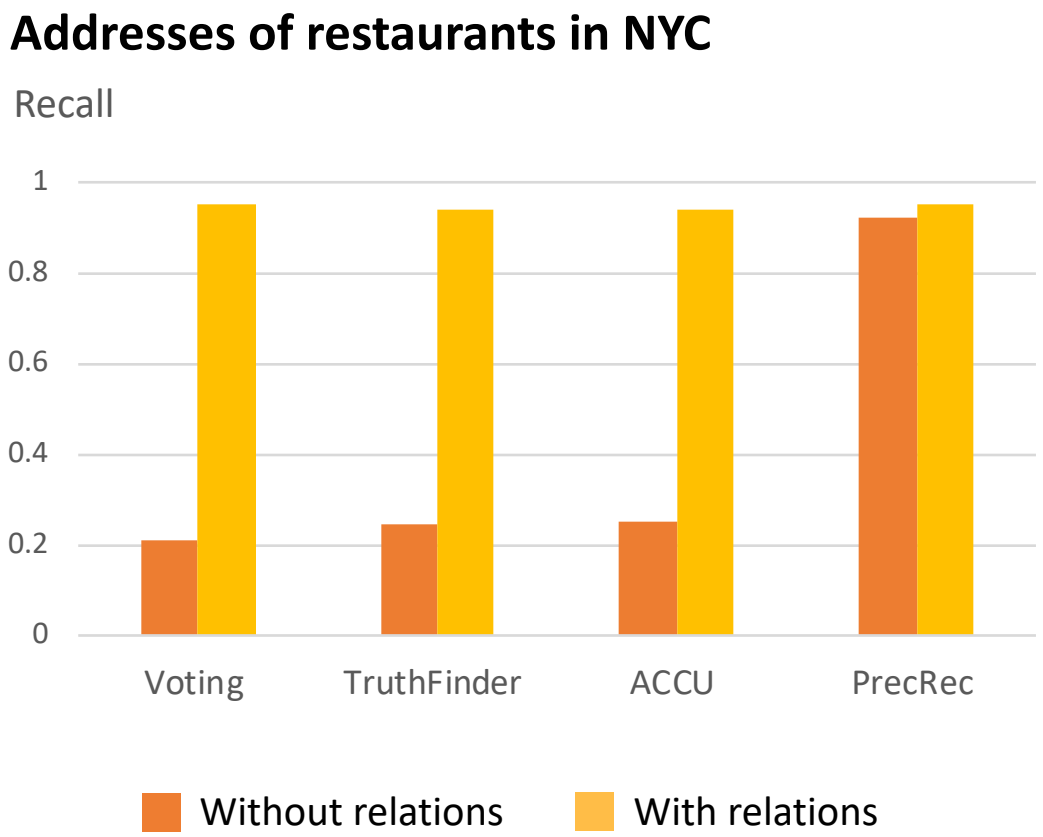
- Datasets
  - Addresses of 11, 589 restaurants in New York's Manhattan area provided by 12 sources[3]
- Relations among claims
  - Obtained from DBPedia and Google Maps
- Ground truth
  - Manually obtained for 500 restaurants
- Data fusion models
  - Single-truth: Voting, ACCU, Truthfinder
  - Multi-truth: PrecRec
- Performance Metrics
  - Precision, Recall, F1-score

3. X. L. Dong, L. Berti-Equille, and D. Srivastava, "Truth discovery and copying detection in a dynamic world," PVLDB, 2009.

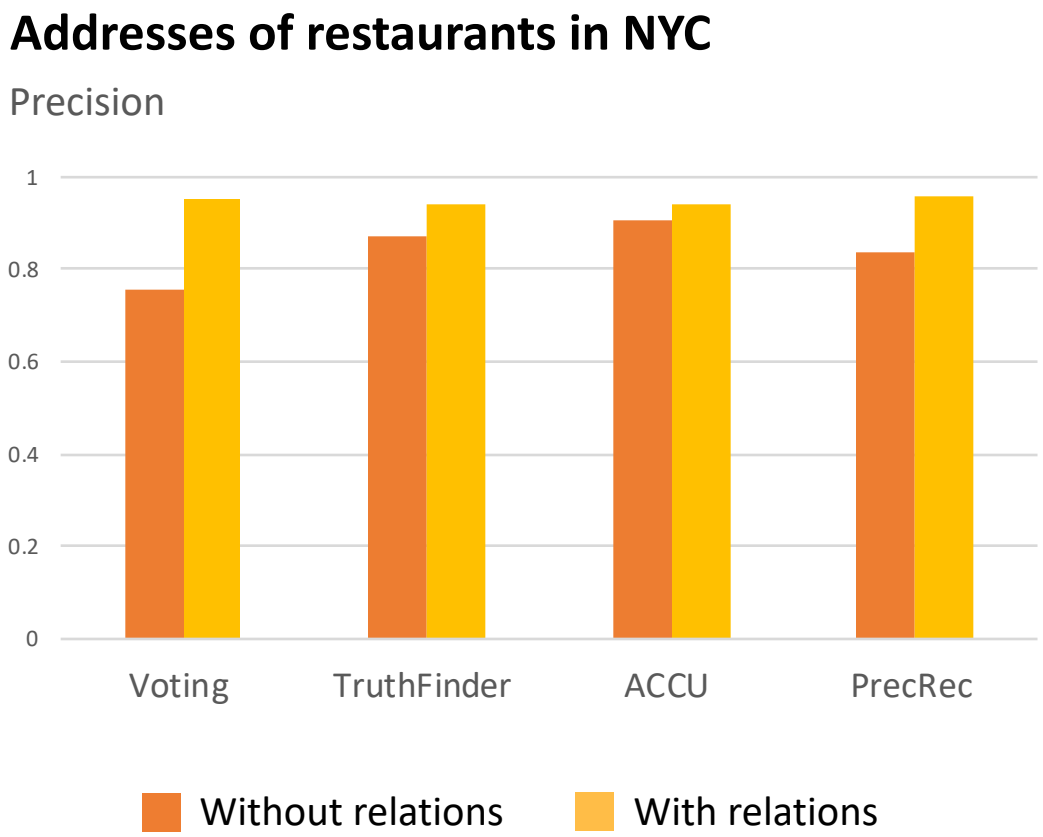
Integrating knowledge of relations removes inconsistent output



Knowledge of relations among claims improves fusion



Knowledge of relations among claims improves fusion



Takeaways

Leveraging the knowledge on relations among claims improves data fusion

- Removed inconsistencies in output claims
- Single-truth data fusion models converted to multi-truth models that perform comparable to multi-truth models relying on training data