



Network Based Hard/Soft Information Fusion

Data Association Process

Gregory Tauer, Kedar Sambhoos,
Rakesh Nagi (co-PI), Moises Sudit (co-PI)



Objectives:

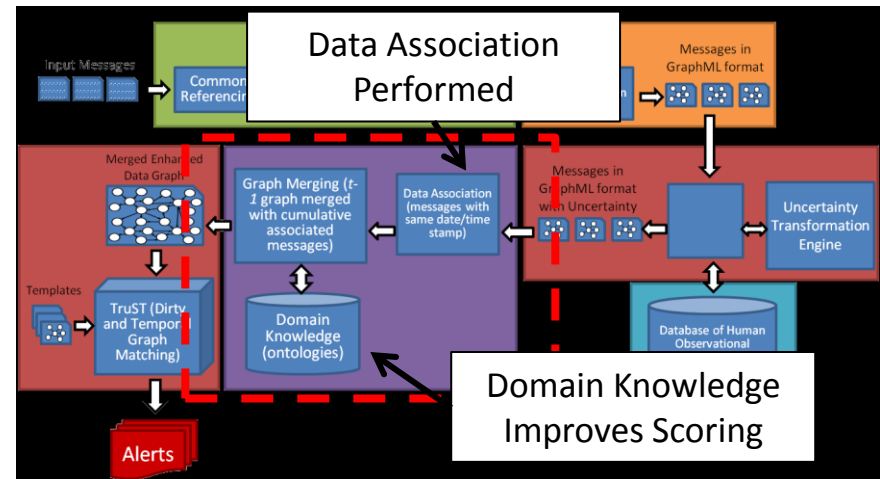
- Formulate and implement a workable, quantitatively-based association approach
- Integrate it into the overall Discovery Process Operations
- Conduct robustness testing and evaluation.

DoD Benefit:

- Heterogeneous data sources: How to merge entities and relationships in messages / sensor reports.
- Dimensionality Reduction and fusion gain.

Scientific/Technical Approach:

- Apply basic form of data fusion process on graph structured data.
- Convert hard data to a soft-compatible form for association.
- Formulate graph association as an IP.
- Use the formulation to study incremental updates.
- Utilize map-reduce framework to assist in distribution of algorithms.



Accomplishments:

- Developed Integer Programming Formulation to solve Graph Association problem.
- Designed a map-reduce Lagrangian decomposition-based heuristic for solving the resulting problem.
- Developed an incremental approach for solving a relaxed version of the problem.

Challenges:

- Scalable incremental approach.
- Divide-and-Conquer approach for large sets.



Project Statistics and Summary

Data Association Process



Degree Awarded

- M.S. – Megan Hannigan
- Ph.D. - Gregory Tauer (graduated 2012)

Students supported

- Number of Graduate Students: 3 (currently 1)
- Number of Undergraduate Students: 2 (currently 1)

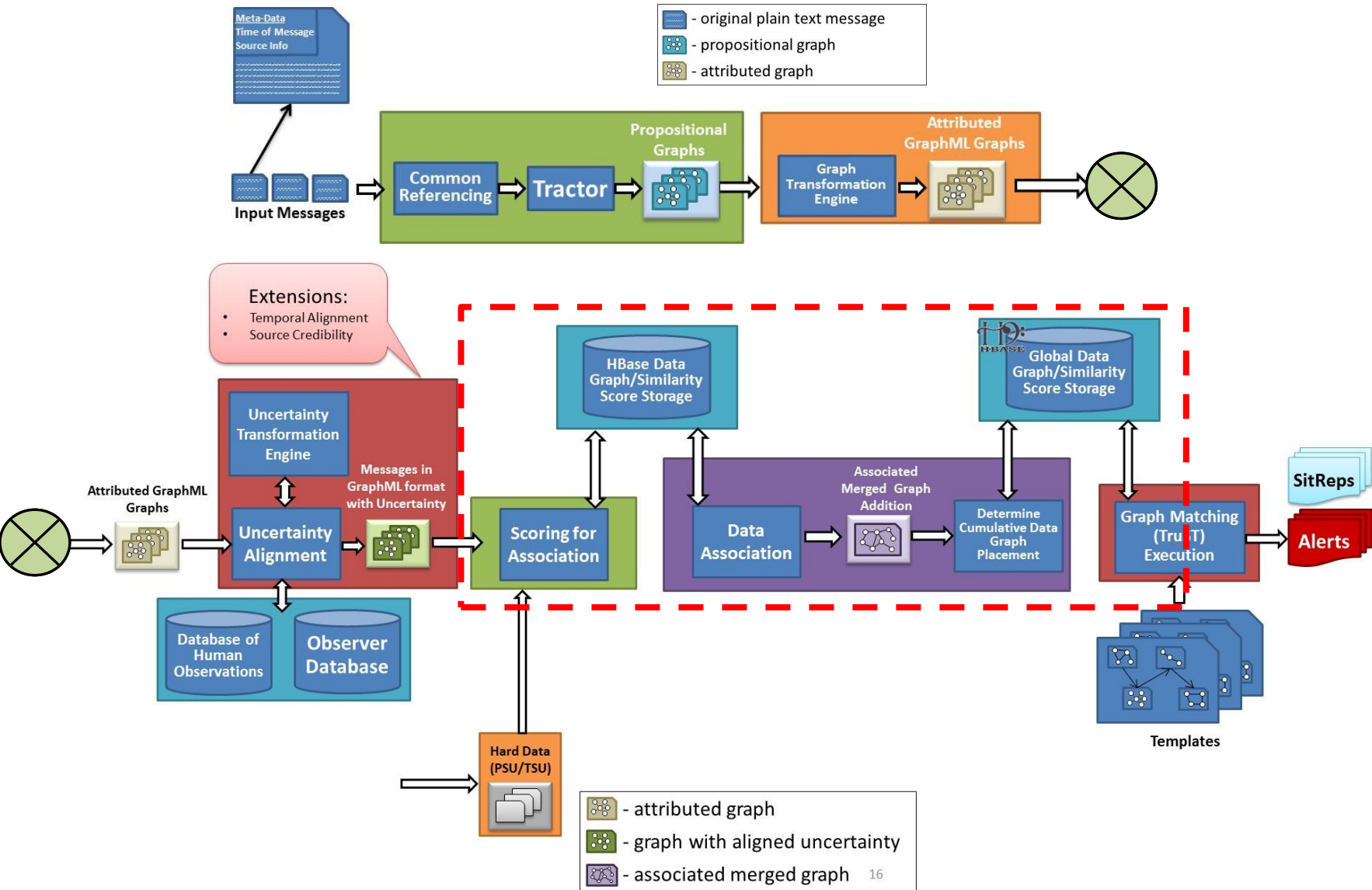
Publications

- Journal papers – 0 published (2 under review and 1 under preparation)
- Conference papers – 2 (1 published and 1 accepted)



The MURI Architecture

"In brief"





Main Scientific/Technical Accomplishments

Data Association Process



Motivation for Data Association:

- Information Gain
 - Provides a means to find answers that span multiple sources.
- Dimensionality Reduction
 - Reduction in duplicate entities improves the efficiency of analytics.



Main Scientific/Technical Accomplishments

Data Association Process



Years 1 and 2

- Graph association for the association of richly relational data.
- Utilized location resolution to improve association of hard with soft.

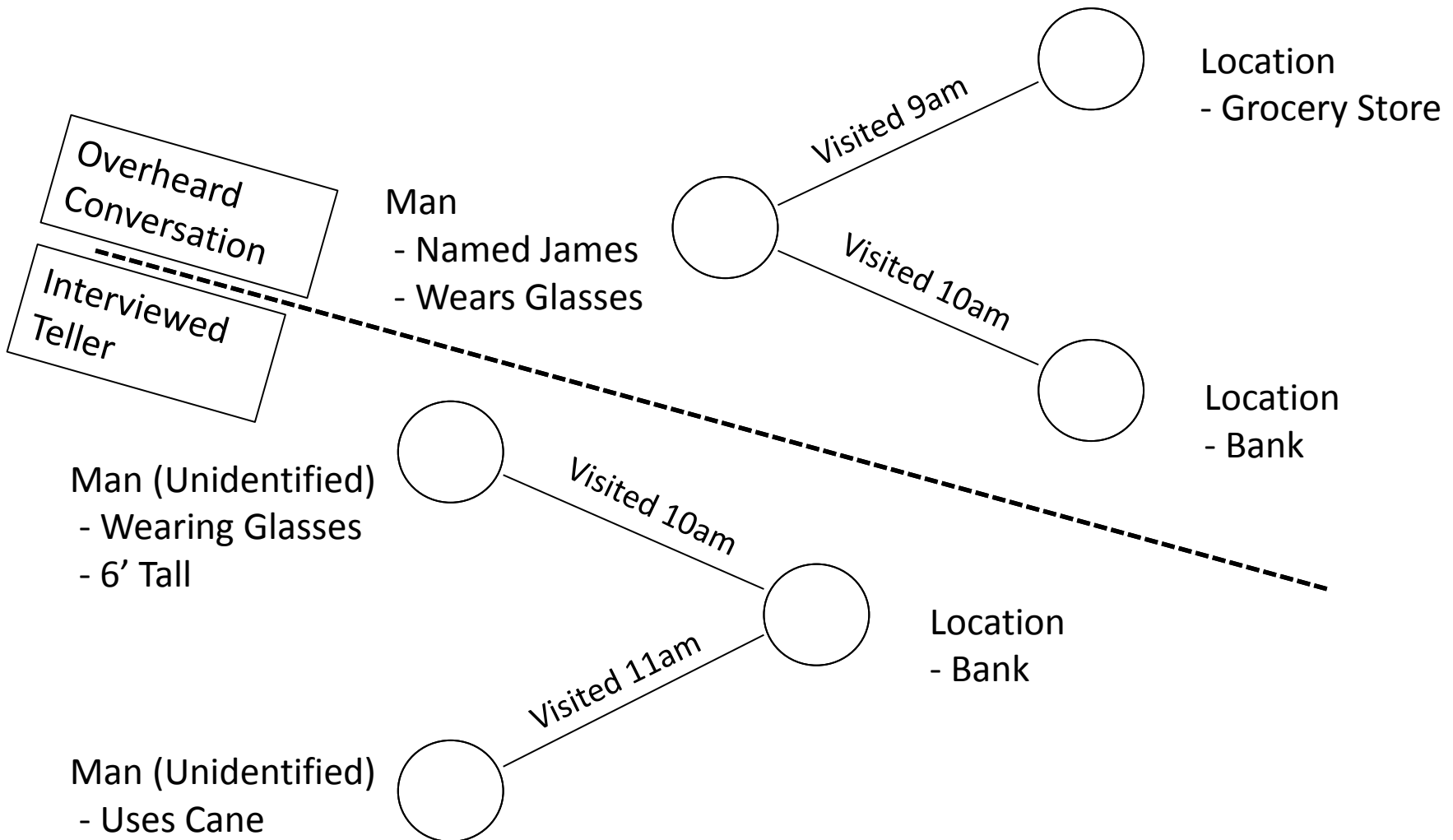
Year 3 Accomplishments

- Distributed (“Cloud”) version of the Lagrangian heuristic.
- Incremental association approach for streaming datasets.



Technical Approach

Graph Association



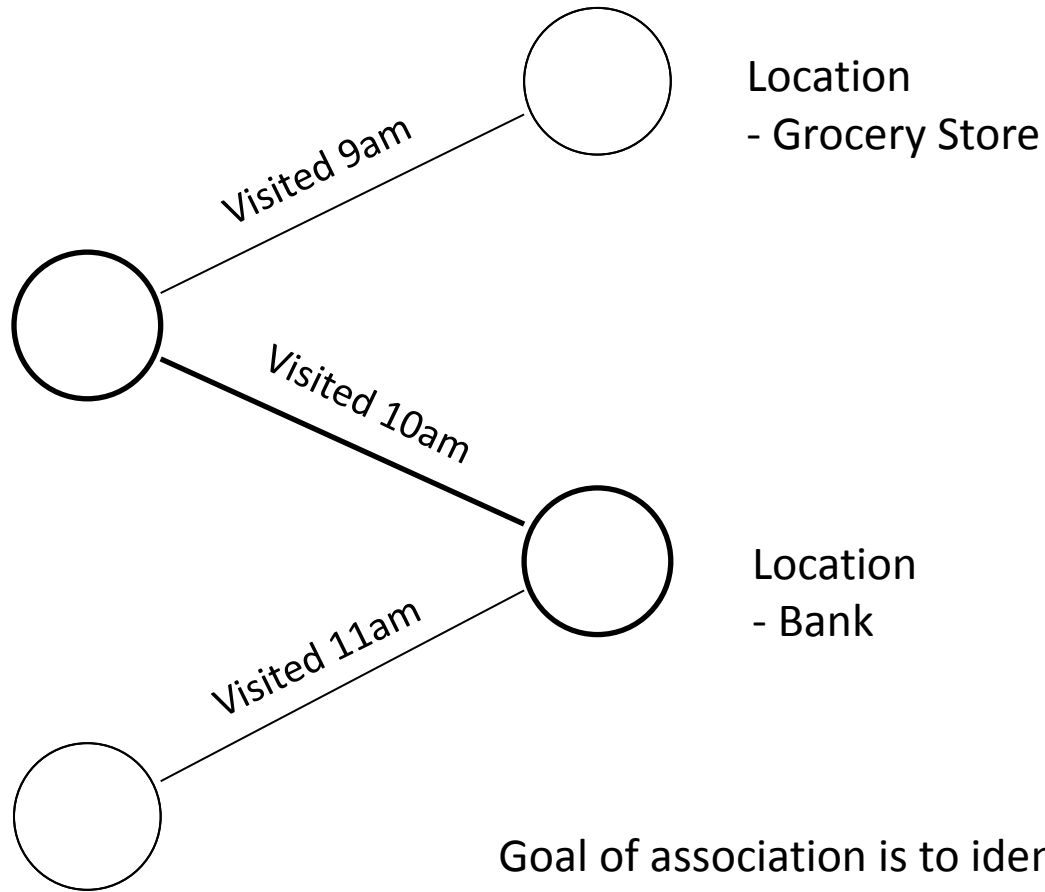


Technical Approach

Graph Association



Person
- Named James
- Wears Glasses
- 6' Tall



Person
- Uses Cane

Goal of association is to identify the common information in given graphs so they can be merged.

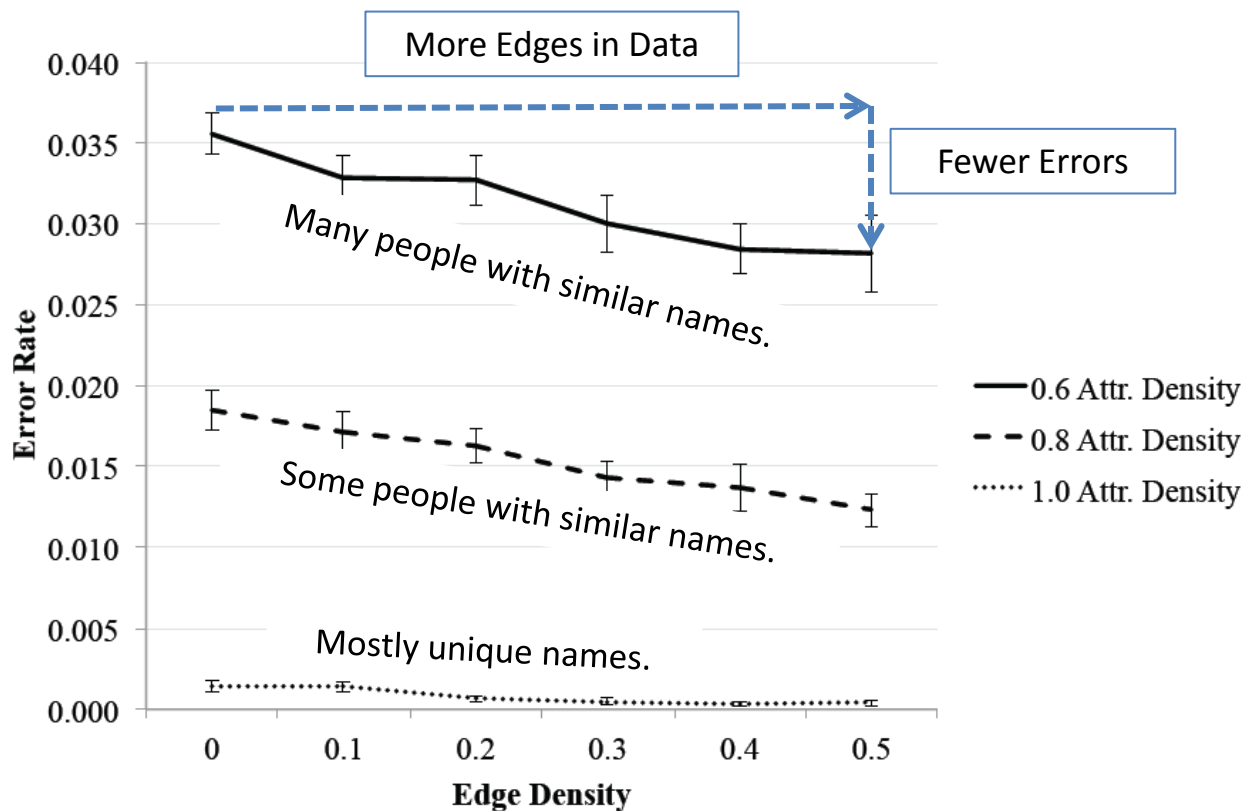


Technical Approach

Graph Association



- The primary advantage of graph association is that the relationship between observations can help resolve ambiguity, which improves accuracy.



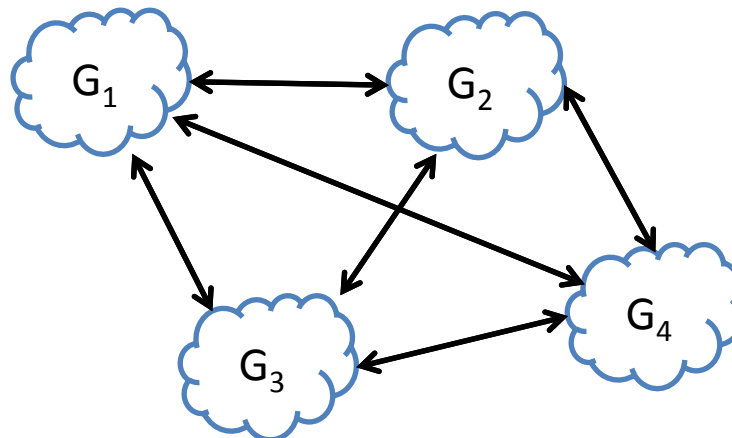
More Details Available in: G. Tauer, R. Nagi, and M. Sudit. The Graph Association Problem: Mathematical Models and a Lagrangian Heuristic. Re-submitted to Naval Research Logistics. 2012.



Technical Approach

Distributed ("Cloud") Algorithm

- Graph association is a **very hard problem**.
- We developed a heuristic based around solving sub-problems between all pairs of graphs instead of between all graphs simultaneously.
- Allows parallelization / distribution: *all sub-problems can be solved simultaneously on different computers.*

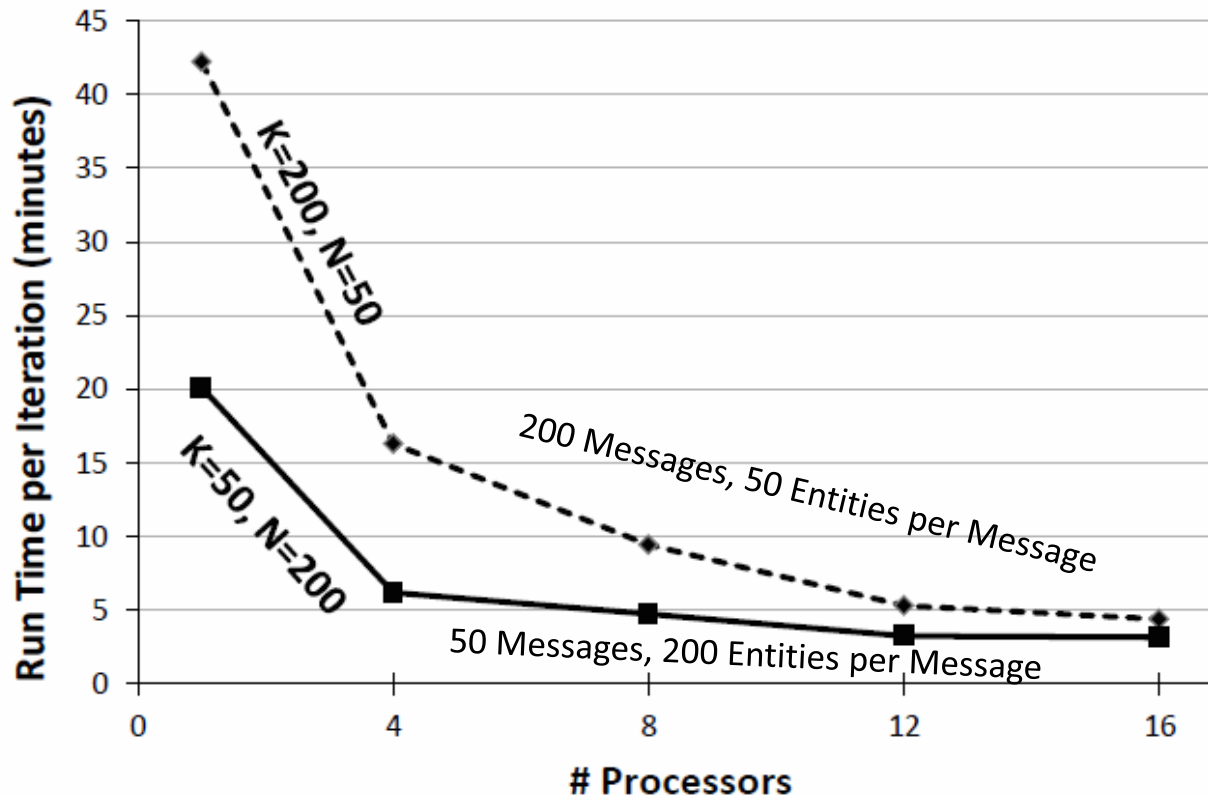




Technical Approach

Distributed (“Cloud”) Algorithm

- Heuristic was implemented using the Apache Hadoop framework.
- Our algorithm is iterative. It provides a preliminary solution after the first iteration and subsequent iterations improve upon it.





Technical Approach

Hard Location Data

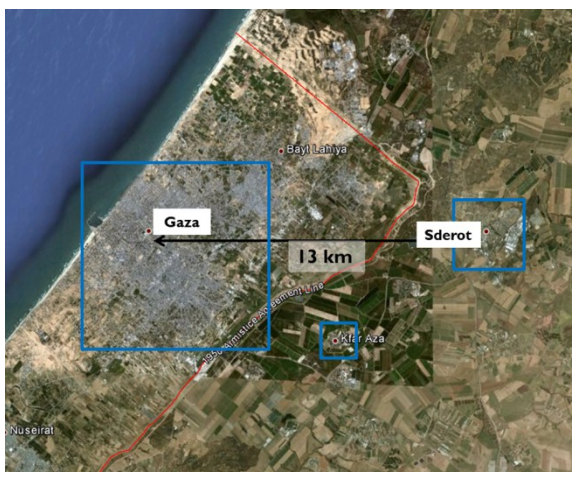
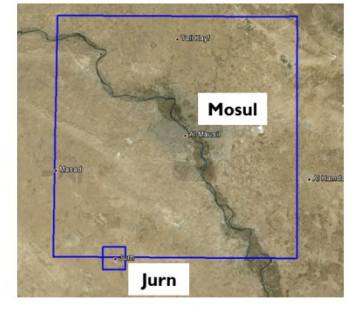
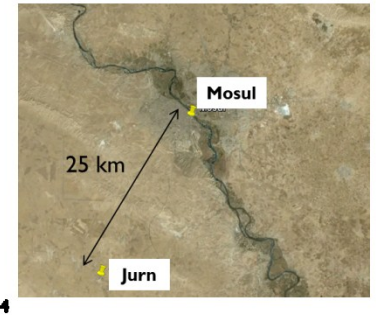
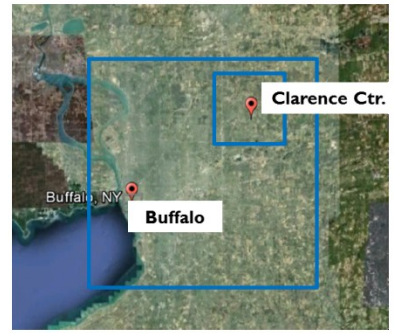
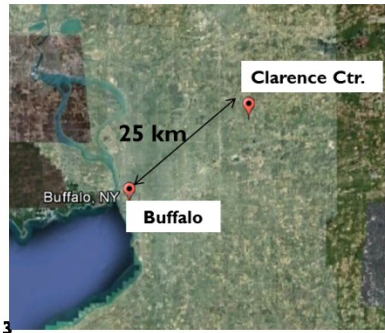


www.foxnews.com/
 February 13, 2009: A Continental Airlines commuter plane coming in for a landing in **Buffalo, N.Y.**, dove into a house in snowy, foggy weather, killing all 49 people on board and one person on the ground. The crash of Flight 3407 sparked a fiery explosion. Firefighters worked through the night to douse the flames...

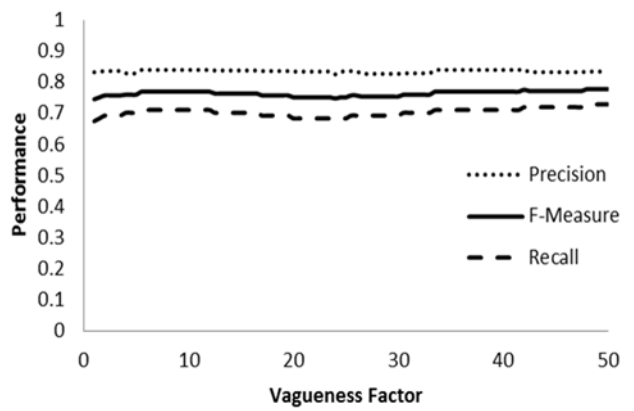
www.telegraph.co.uk
 Continental Airlines flight 3407 crashed into the town of **Clarence Center, New York**, about five minutes before it was due to land on Thursday night at Buffalo airport. Clarence is located 20 miles north-east of Buffalo. All 48 people aboard a commercial plane and one person on the ground have been killed after the aircraft crashed into a house near Buffalo, New York state, and burst into flames...

GTD: 200804060007
 04/06/2008: On Sunday, unknown gunmen set up a fake checkpoint and intercepted two college buses, one carrying male students and one carrying female students, in **Mosul, Nineveh province, Iraq**. The bus carrying the female students managed to escape but the gunmen held the 42 male college students at gunpoint and ...

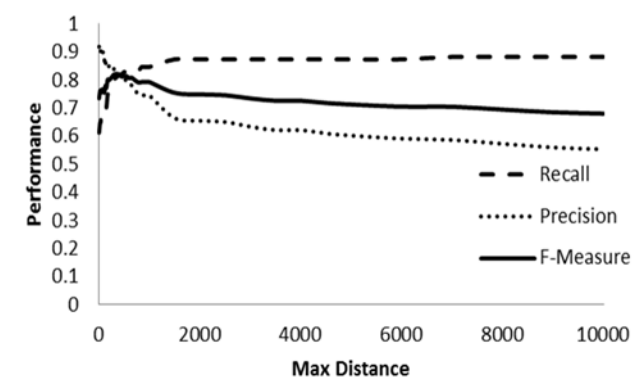
WITS: 200804509
 On 6 April 2008, in the morning, in **Jurn, Ninawa, Iraq**, armed assailants stopped two school buses carrying students to Mosul University at a fake checkpoint. The assailants then fired upon one of the busses as it managed to escape, wounding three students and damaging the bus. Assailants kidnapped all 42 students on board the second bus...



Spatial Performance vs Increasing Vagueness



Euclidean Performance vs Increasing Max Distance

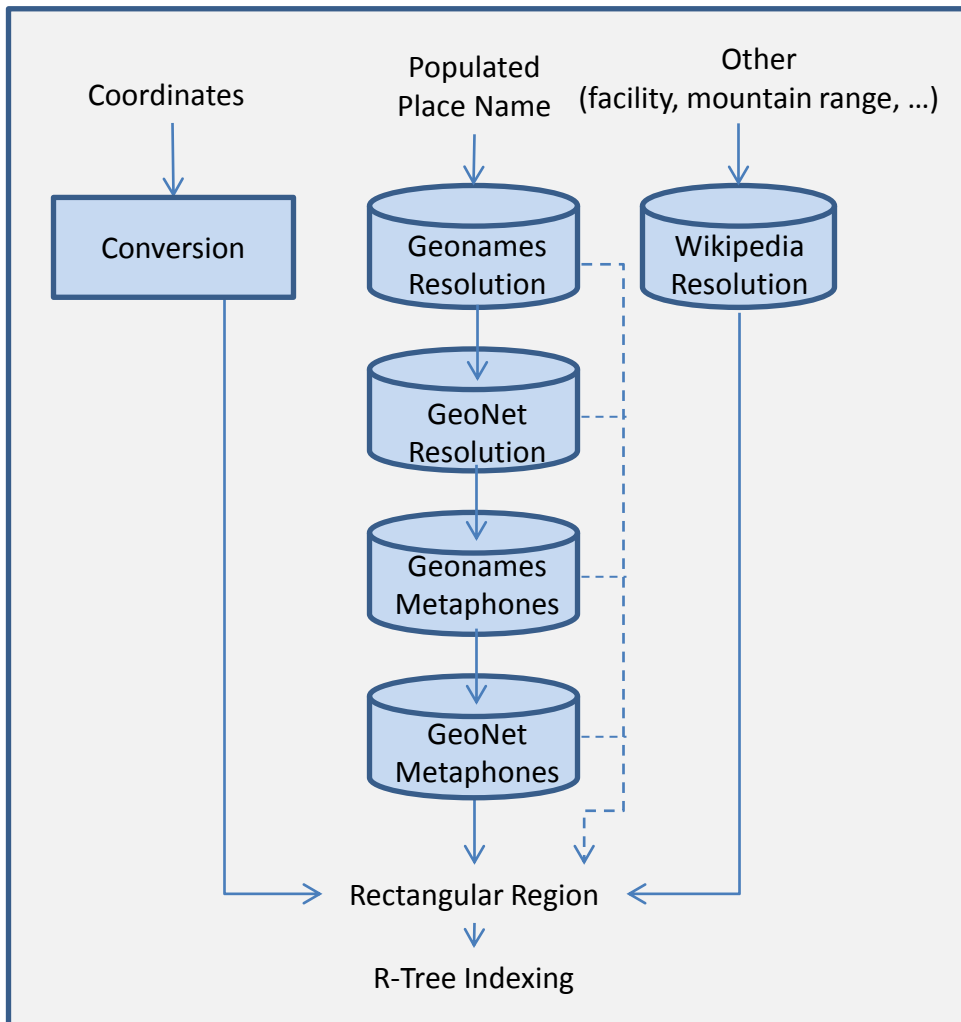




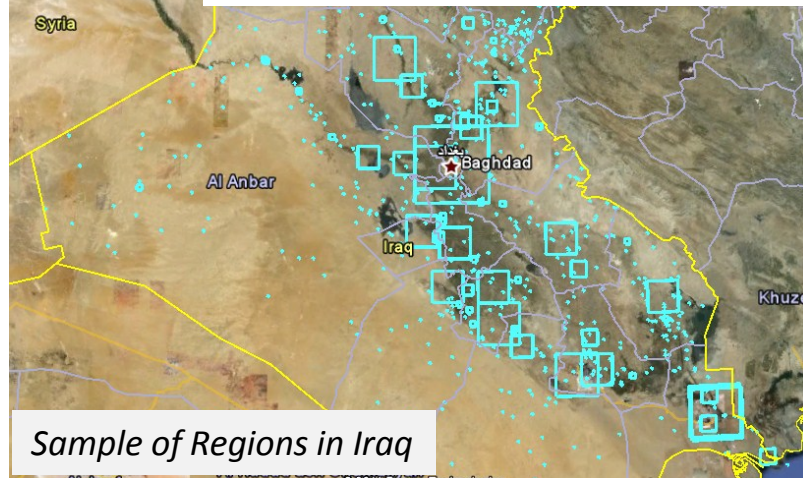
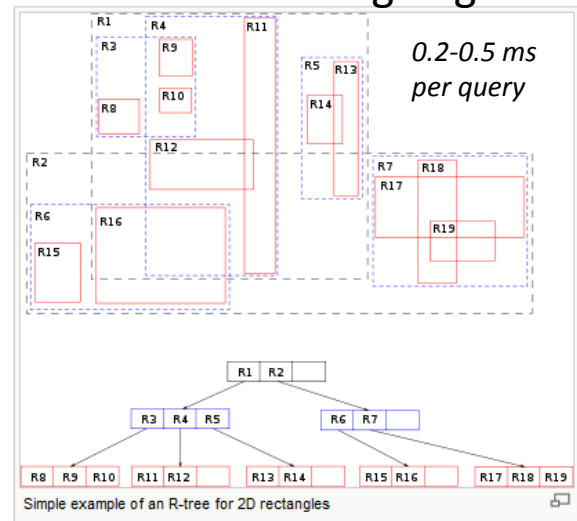
Technical Approach

Hard Location Data

Location Normalization (McConky et al. 2012)



R-Tree for Indexing Regions





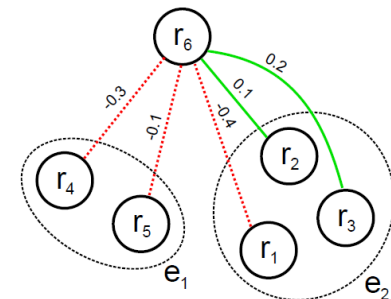
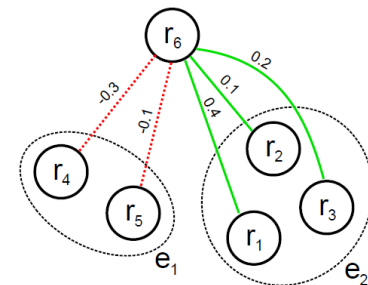
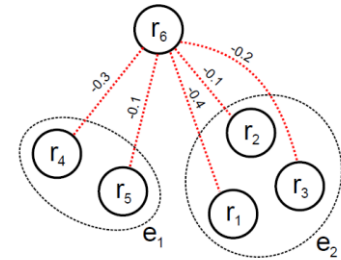
Technical Approach

Incremental Algorithm

Think of association as clustering nodes by their true identity.

- When a new node arrives to a previously associated problem, one of three cases will happen:

- The new observation will be assigned to its **own new cluster**. This should happen if the observation represents an observation that **has not** previously been encountered.
- The new observation will be **added to an existing cluster**. This should happen if the new observation represents an entity that **has** previously been encountered.
- The new observation will be **added to an existing cluster** and **some of the other cluster assignments will change**. This case allows the algorithm to **undo a previous mistake** when new evidence arrives.





Technical Approach

Incremental Algorithm

We model this problem as the clique partition problem (CPP):

$$\begin{aligned} CPP(\mathbf{w}) = \max & \sum_{i=1}^{N-1} \sum_{j=i+1}^N w_{ij} x_{ij} \\ \text{st: } & x_{ij} + x_{ik} - x_{jk} \leq 1, & \forall i = 1, 2, \dots, N-2 \\ & -x_{ij} + x_{ik} + x_{jk} \leq 1, & \forall j = i+1, i+2, \dots, N-1 \\ & x_{ij} - x_{ik} + x_{jk} \leq 1, & \forall k = j+1, j+2, \dots, N \\ & x_{ij} \in \{0, 1\}, \\ & 0 \leq x_{ij} \leq 1 \end{aligned}$$

- This facilitates a clear understanding of when, and in what manner, each of the three previous cases should be applied.
- Details of this algorithm are available in:
 - G. Tauer, R. Nagi, and M. Sudit. An Incremental Graph-Partitioning Algorithm for Entity Resolution. Working paper. 2012.
 - G. Tauer. Data Association on Large Quantities of Complex Data. Ph.D. dissertation. State University of New York at Buffalo, 2012.



Technical Approach

Incremental Algorithm



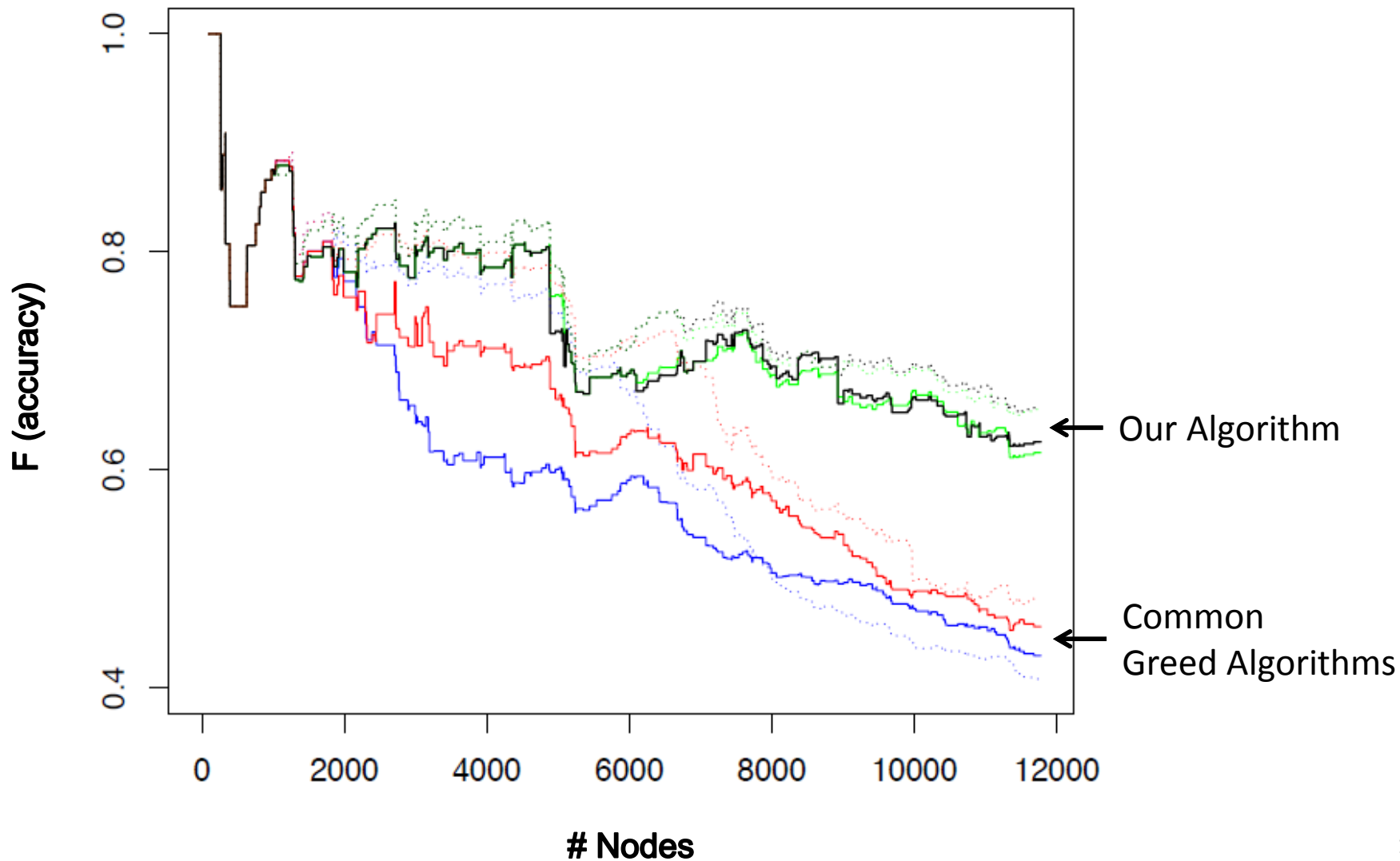
- The incremental algorithm is tested on benchmark datasets:
 - “John Smith”
 - 197 NY Times articles each containing a person named “John Smith”.
 - 35 different John Smiths, 11,775 total nodes.
 - arXiv
 - Database of high-energy physics publications originally used in KDD Cup 2003.
 - 58,515 entities to associate.



Technical Approach

Incremental Algorithm

"John Smith" Results

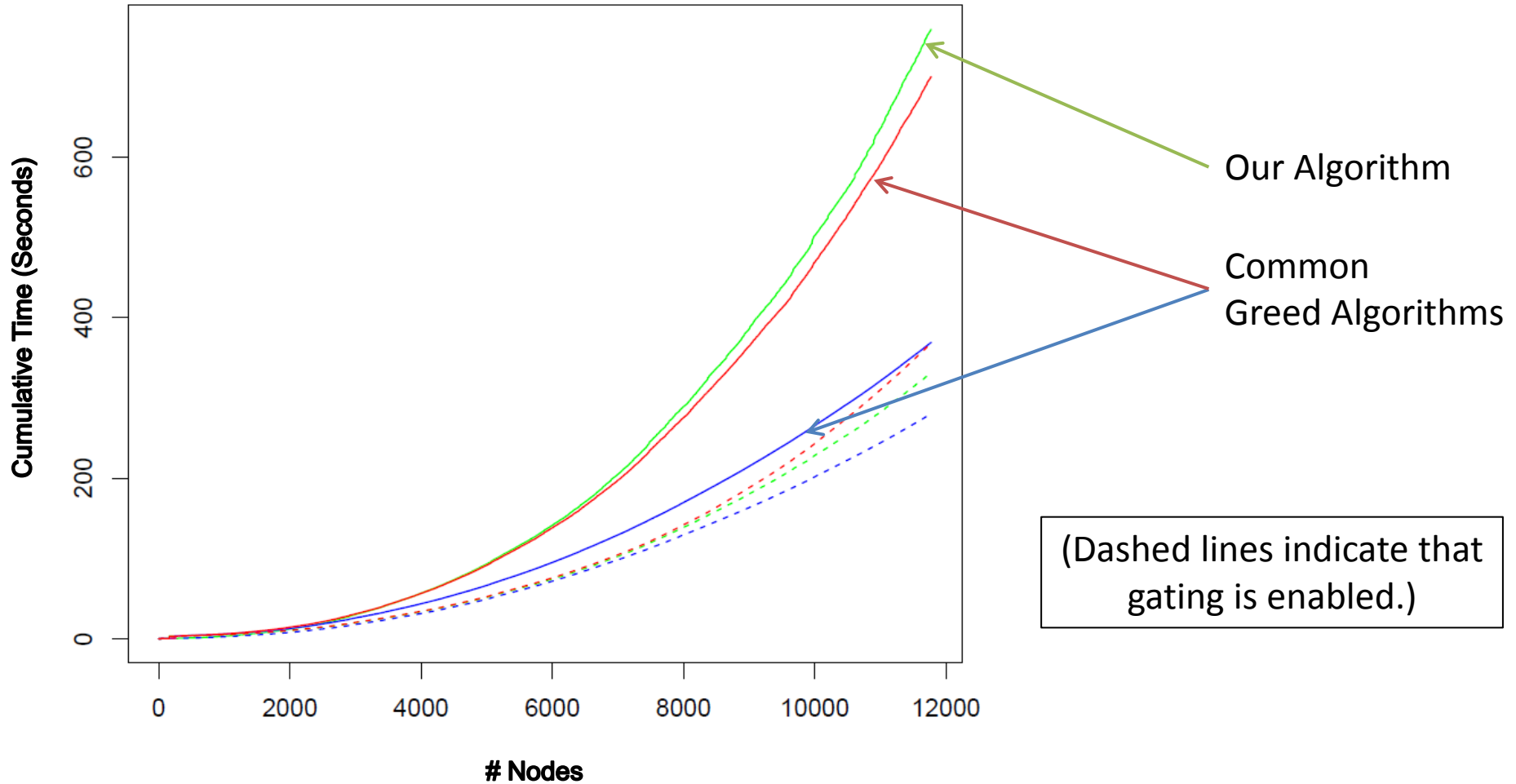




Technical Approach

Incremental Algorithm

“John Smith” Results

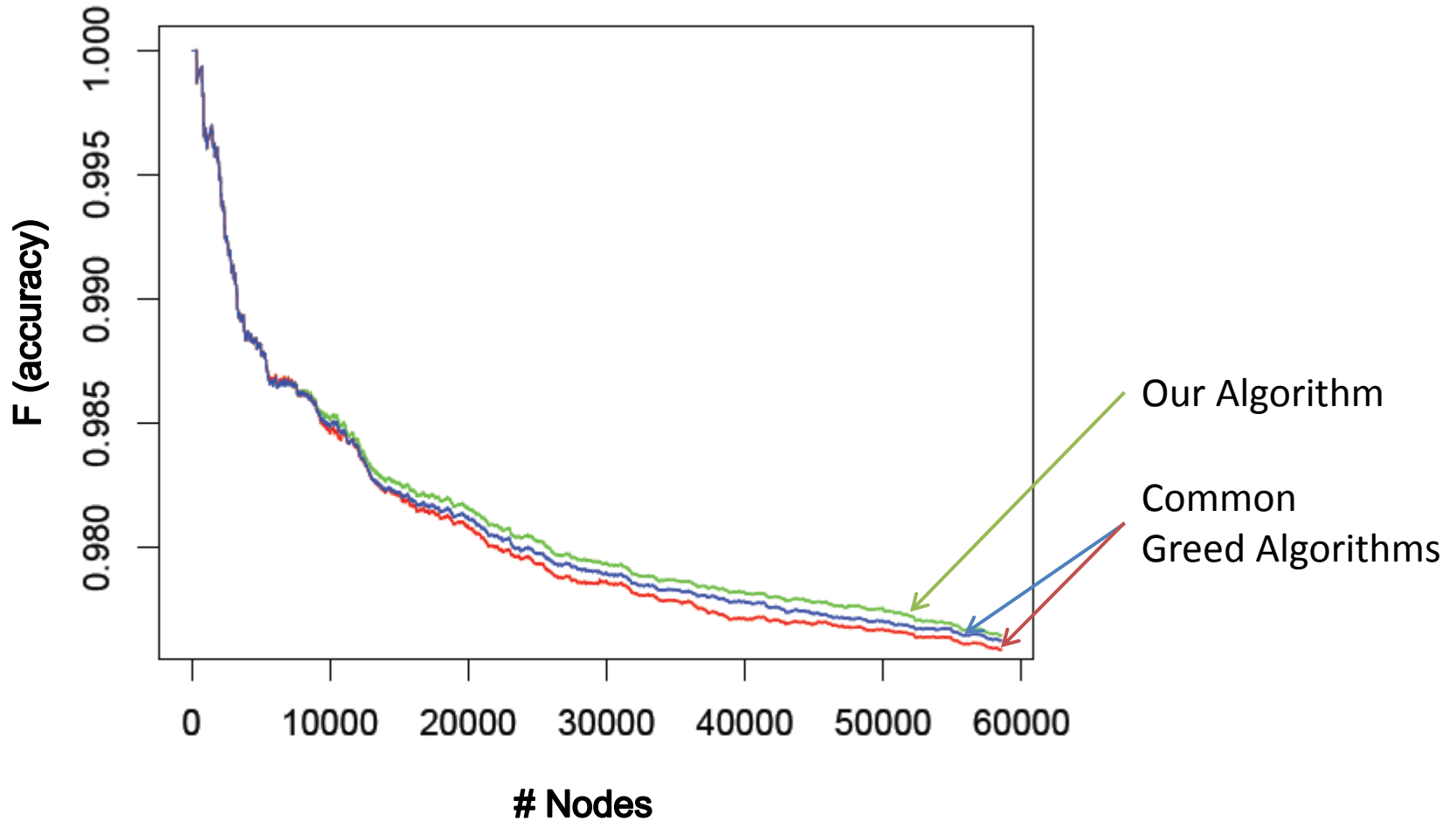




Technical Approach

Incremental Algorithm

“arXiv” Results



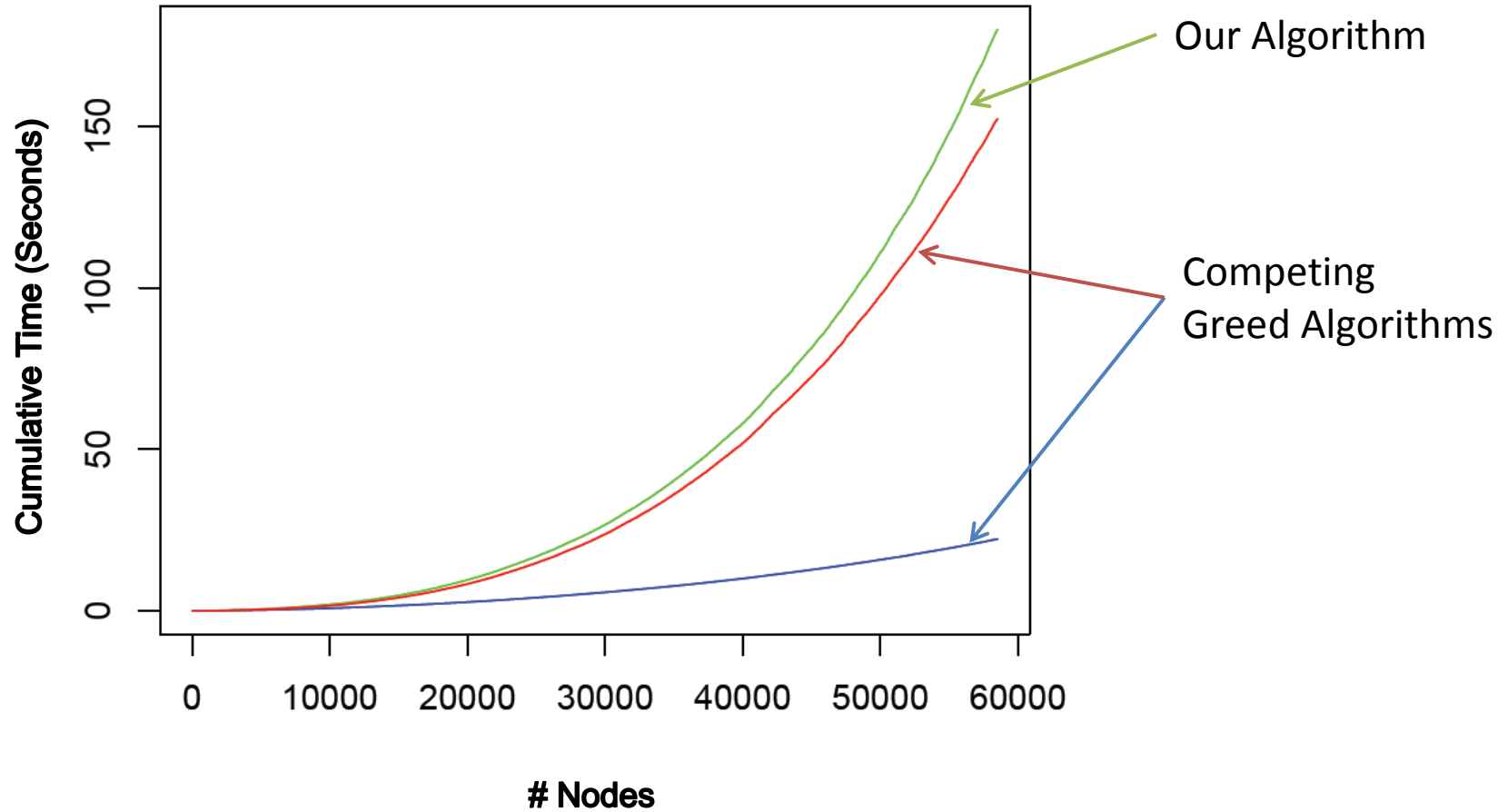
- Greedy algorithms are popular in data association / entity resolution because they often work very well.



Technical Approach

Incremental Algorithm

"arXiv" Results





Option Year Plans

Data Association Process

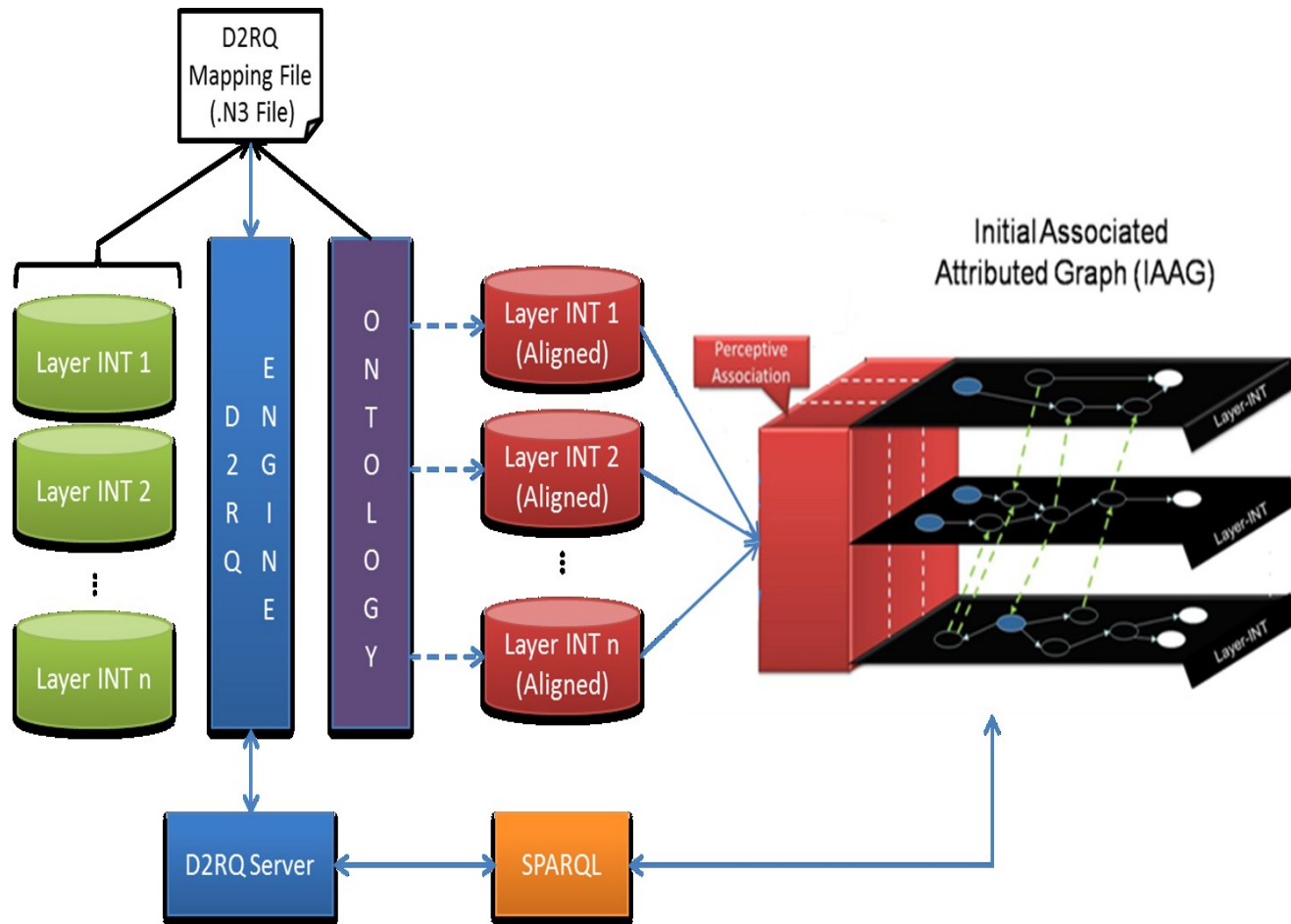


- **Capability Goal:**
 - Will build capacity to handle extremely large and high bandwidth datasets in preparation for transition.
- **Research Goals:**
 - Explore local search / meta-heuristics.
 - Adapt approximate graph matching heuristics for association.
 - Divide and conquer strategy for very large input data sets.



Option Year Plans

Data Association in a Layering Approach



“Divide and Conquer”

Partition data into subsets which are likely to be highly connected.



Process these subsets independently.



Iteratively merge associated subsets until all data is associated.