Handling Uncertainty in Information Extraction

Anish Das Sarma
Google Research

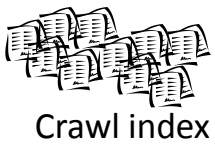
Joint work with Alpa Jain, Divesh Srivastava

Data Uncertainty

- General-purpose modeling of uncertainty in the context of databases
  - Data integration, extraction, scientific, sensor data
  - Trio project @ Stanford
  - Several other projects: MystiQ, MayBMS, Orion, BayesStore, MCDB, ProbView, …
- Uncertainty in Data Integration
  - Uncertain schemas & schema mappings, erroneous data, …
- Uncertainty in Information Extraction
  - This Talk [SIGMOD 2010]

Information Extraction (IE)

- IE systems automatically generate relations from text documents (e.g., web pages)



Information extraction

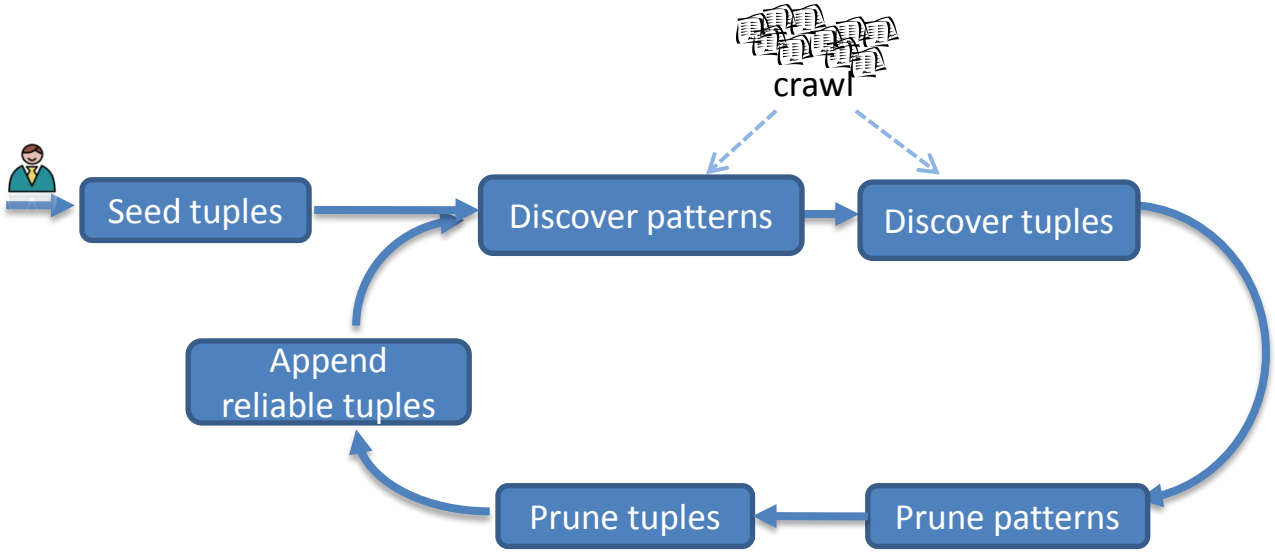
Acted-in

Movie	Actor
Top Gun	Tom Cruise
24 hours	Nicole Kidman
Hollywood	Brad Pitt
…	….

?

- IE users or developers may perform a post-mortem analysis of output, e.g., to understand unexpected tuples
- **Goal: Build an (interactive) investigation tool for IE**

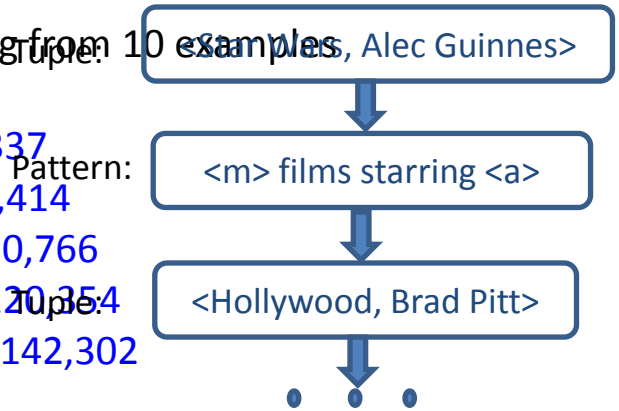
Iterative Information Extraction from Text



Unexpected tuples in output

Example extractions: Starting from 10 examples

- {Book, Author} ➔ 122,337
- {Movie, Author} ➔ 237,414
- {Movie, Director} ➔ 210,766
- {Musician, Band} ➔ 129,354
- {Organization, CEO} ➔ 142,302



Roadmap

- Three phases of interactive investigation
- Algorithms for investigation within an iteration
- Chaining: investigations spanning multiple iterations
- Experiments over real data

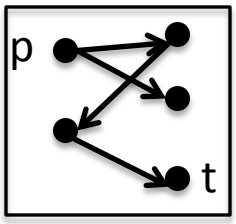
Three Phases of Investigation

- **Explain**
  - Keep **track** of what you've extracted and integrated
  - Why is a data item (not) present in the result?
    - e.g., "*<hollywood, brad pitt> was generated by pattern p*"

- **Diagnose**
  - What would happen if an operator (e.g., threshold) was modified?
  - What is the **impact** of modifying input data (e.g., pattern)?
    - e.g., "*pattern p gave the most number of tuples*"

- **Repair**
  - Fix the IE execution when you have **feedback** on output
    - e.g., "*I know tuples <t1>, <t2> are wrong, and <t3> is correct; fix output*"
  - **Suggest** potential problems automatically, to guide debugging

Explanation Queries



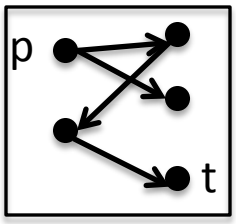
Q1-3: Given tuple *t*, determine:

1. Set of patterns *that contributed* to *t*
2. Pattern that *contributed to t the most*
3. *First iteration* that discovered *t*

Q4: Determine **most influential patterns** in IIE:

- Helps in focusing human feedback
- *Rank* patterns in influence order
- Find *set of K* patterns with maximum *combined influence*

Diagnosis Queries



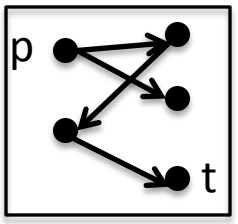
Q1-3: Given pattern *p*, determine:

1. Set of tuples *produced by p*
2. All *tuples that get eliminated* on removing *p*
3. *First iteration* that discovered *p*

Q4: Determine ***K most influential tuples*** in IIE (to get feedback)

- tuples that are contributed to by the largest number of patterns

Repair Queries



Incrementally revise IIE when:

Q1: One or more patterns are deleted

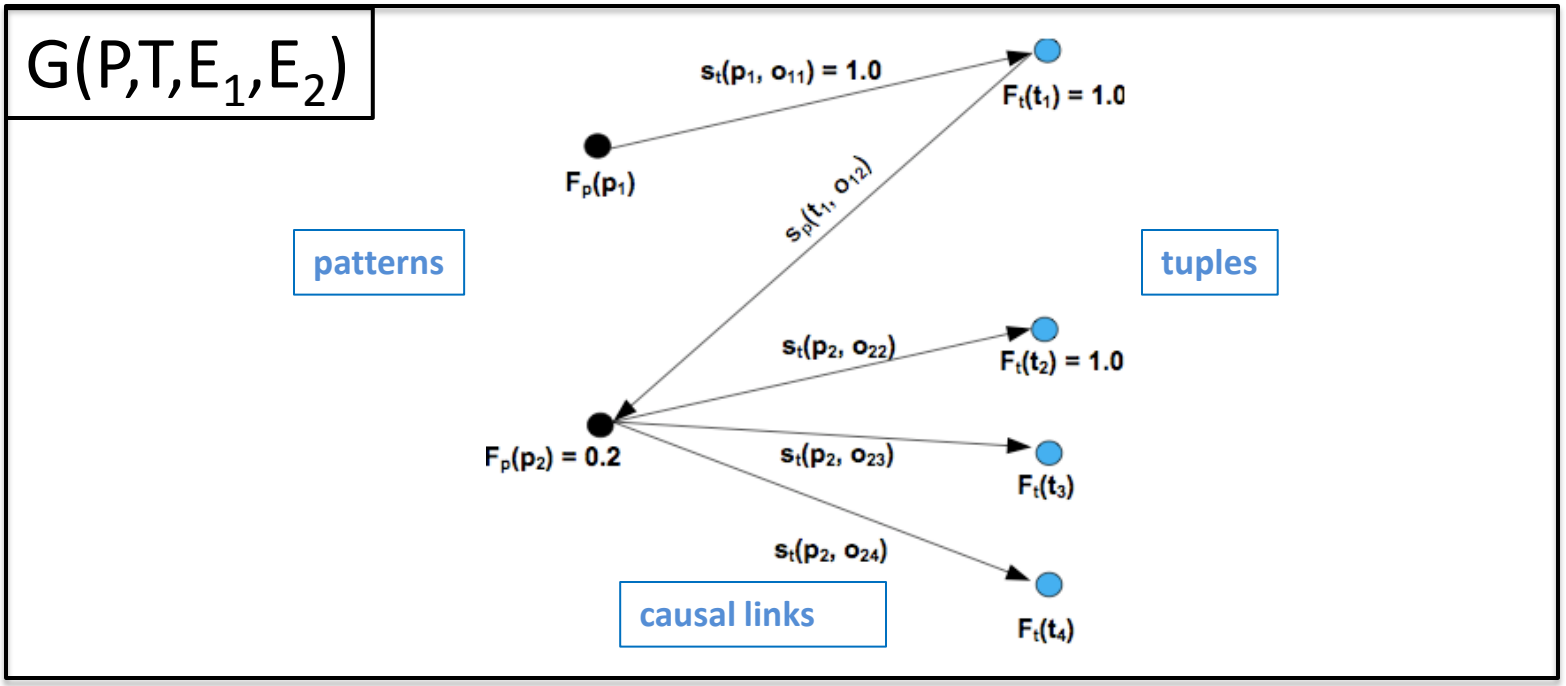
Q2: Score of one or more patterns is revised

Q3: Some thresholds on tuples and patterns are modified

Q4: A (small) set of tuples is annotated (by a user) as correct or incorrect

EBG: Data-structure per Iteration

- We characterize each iteration using an **Enhanced Bipartite Graph** (EBG)



- As IIE progresses, **maintain** tracing information in EBG
- Answer **questions on an iteration** using corresponding EBG
- Answer **questions across iterations** by chaining EBGs

Roadmap

- Identify three phases of interactive investigation
- **Algorithms for investigation**
- Chaining: investigations spanning multiple iterations
- Experiments over real data

Algorithm Performance Summary

Given EBG with M patterns and N tuples:

Complexity	1	2	3	4
Explain	✓	✓	✓	✗ Next
Diagnose	✓	✓	✓	✗ NP-complete(N,M), (1-1/e)-approx M=O(log N) => $N^{\log \log N}$
Repair	✓	✓	✓	✓

✓ Tractable questions: near-linear time algorithms

✗ Potentially intractable

E4: Identify Influential Patterns

- Not all patterns have the same impact on extraction
- Limited editorial resources
- Best patterns to seek feedback on

Influence Measure
Confidence score of *p* (naïve)
Number of tuples produced by *p*
Number of tuples *only p* produced
Total score contribution of *p* over all tuples

Influence Measure: Number of tuples produced by *p*

- **NP-complete: Direct reduction from set cover.**
  - **Instance of set cover**
    - Universe U={1,2,…,n}
    - Subsets S_1, …, S_m
  - **Construct EBG:**
    - Each S_i forms a pattern
    - Element in U forms a tuple
    - Pattern P_i produces tuple t_j iff: $j \in S_i$
- **Constant-factor approximation**
  - Greedy algorithm

Chaining Investigations over Iterations

- **So far:** One-step, i.e., per iteration
- **Goal:** *Chain* multiple iterations
- EBG allows easy extension of algorithms for chained investigation – effectively "unfolding" the bipartite graph*****

- **Examples:**
  - Explain: Find all patterns that (directly or indirectly) contributed to tuple *t*
  - Diagnose/Repair: Find all tuples that would get deleted/modified if pattern 'p' were deleted/score changed

* **Details in the paper**

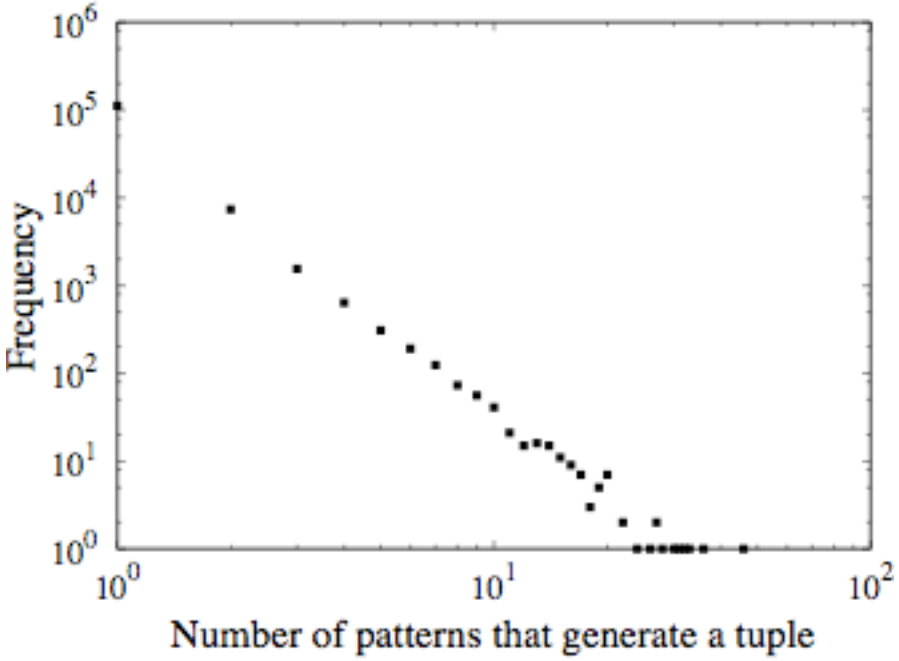
Experimental Evaluation

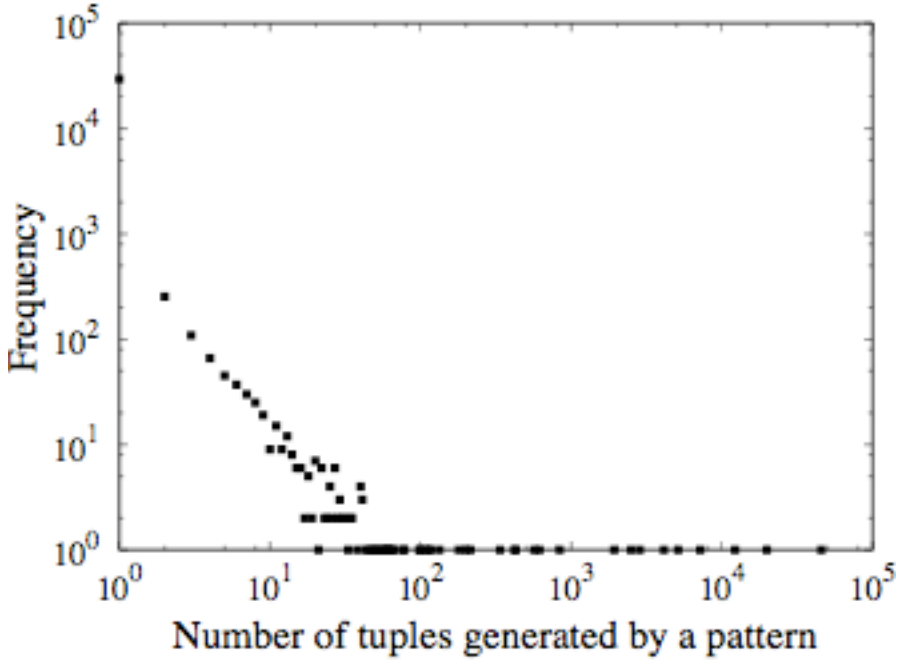
- Extraction using Wiper (Y! Labs) on 6 domains
  actors, books, directors, mayors, senators, political-party
- Web pages in a Y! index sample of ~500M documents
- Size of relations varying from 2,000 to 250,000

- **Goal:** Initial feedback on the utility of the I4E framework
  - Is **influence** a useful measure?
  - Is **overhead** manageable?

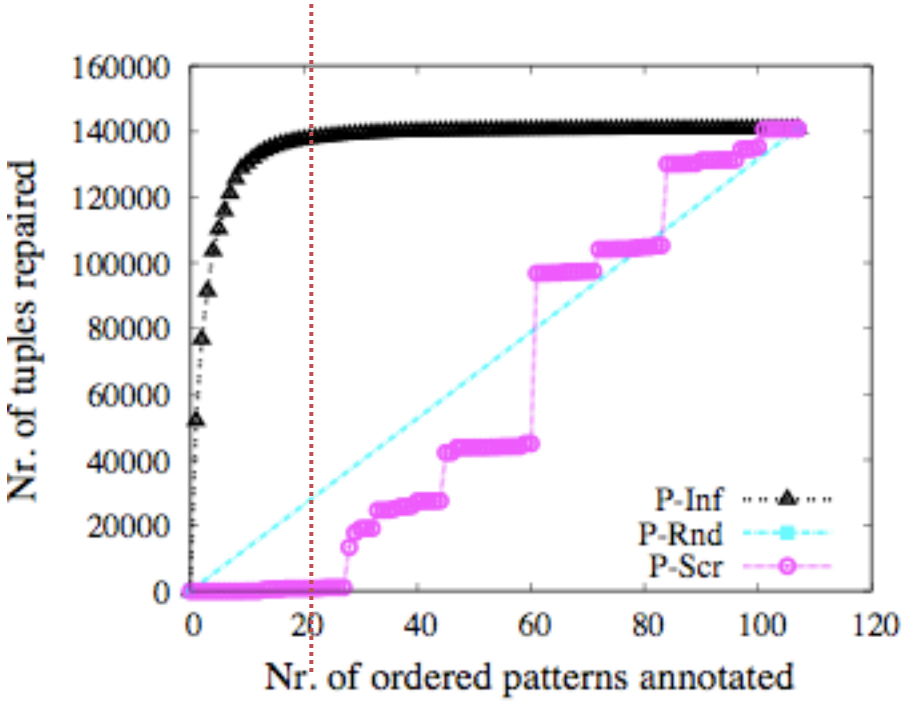
- Presenting results on books (paper has more)

Statistics

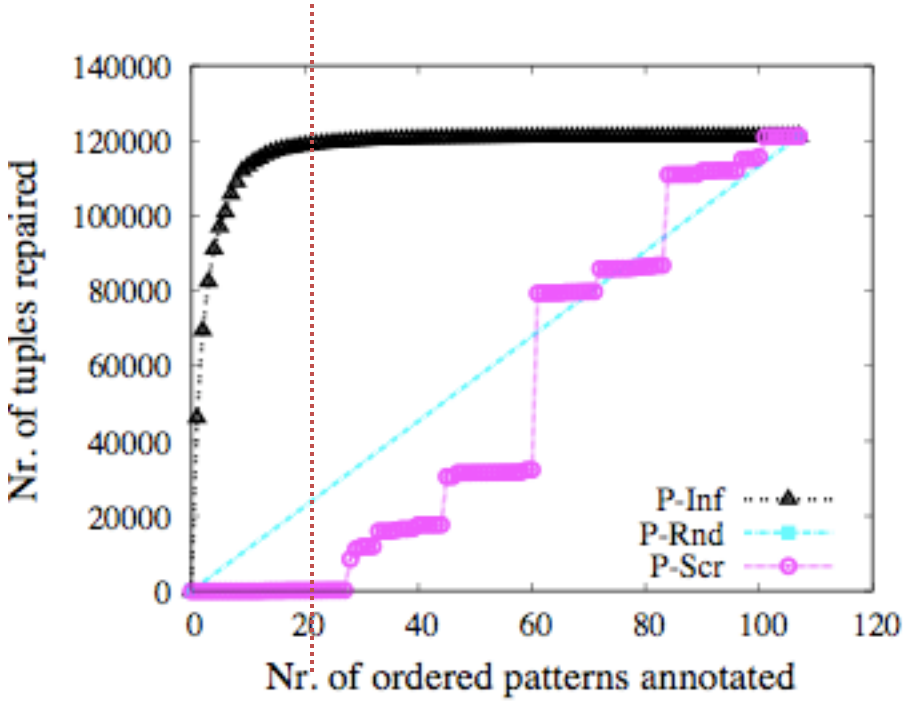

Frequency
10^6
10^5
10^4
10^3
10^2
10^1
10^0
10^0
10^1
10^2
Number of patterns that generate a tuple


Frequency
10^5
10^4
10^3
10^2
10^1
10^0
10^0
10^1
10^2
10^3
10^4
10^5
Number of tuples generated by a pattern

Repaired Tuples for Annotated Patterns

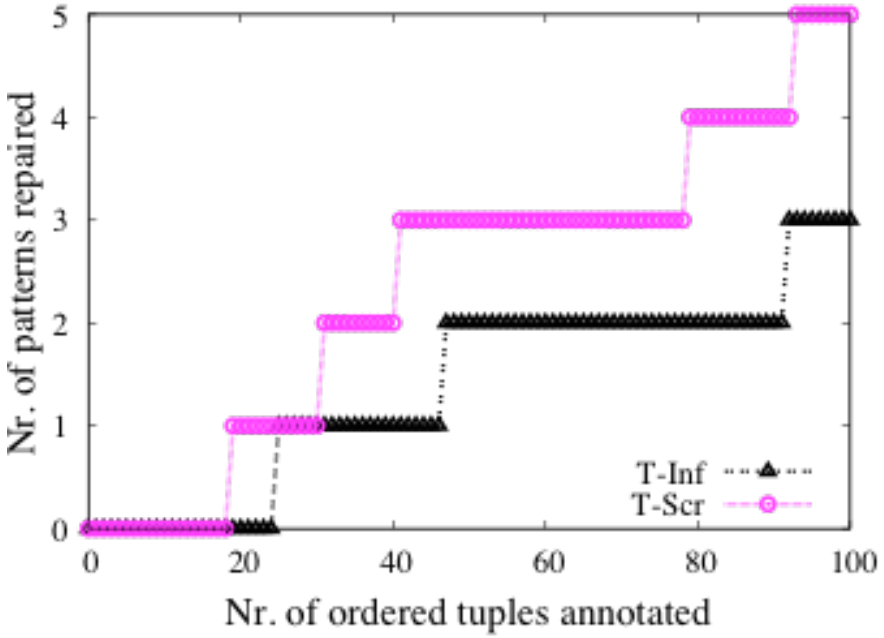


When pattern is correct

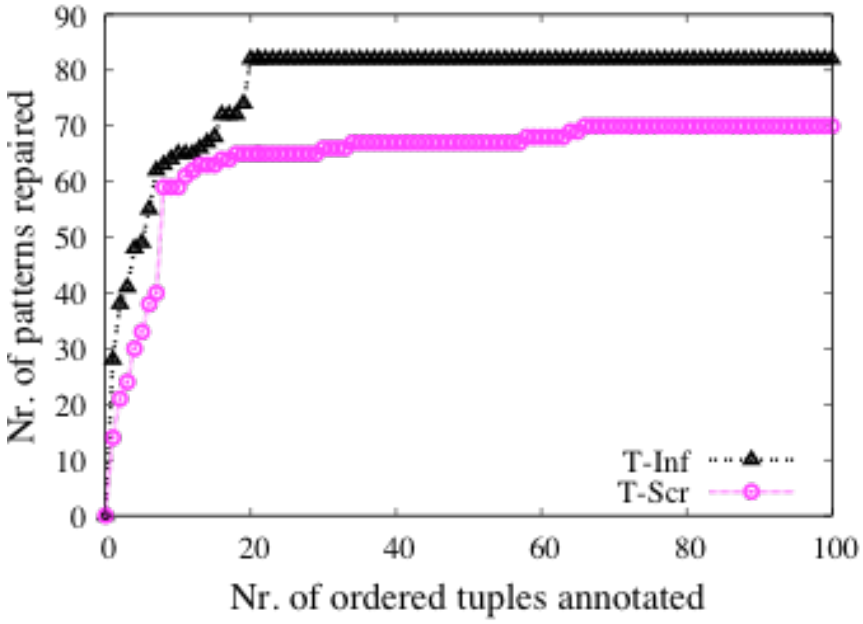


When pattern is wrong

Repaired Patterns for Annotated Tuples



When tuple is correct



When tuple is wrong

Space Overhead

- **Two scenarios:**
  - B1: Baseline needs to keep track of pattern-tuple edges for score recomputation
  - B2: Baseline doesn't keep track of edges at all (unrealistic?)

B1: ~5-15%

#iterations

Domain	5	10	15
actors	14.1	6.67	4.31
books	13.22	6.66	4.1
directors	13.00	6.21	4.04
mayor	13.13	6.23	4.13
sen-party	15.31	7.21	4.71
sen-state	14.23	6.70	4.40

B2: ~30-100%

#patterns

Domain	5	10	15	all
actors	30.2	52.5	63.9	113.2
books	34.3	55.6	61.2	98.4
directors	33.7	46.8	55.7	94.9
mayor	37.3	56.2	59.7	97.1
sen-party	45.2	60.1	69.1	138
sen-state	21.5	41.7	52.7	115.2

Time Overhead

#iterations

Domain	5	10	15
actors	2.39	1.01	0.65
books	2.37	1.28	0.8
directors	7.3	6.51	1.3
mayor	1.71	0.91	0.62
sen-party	12.40	6.22	4.12
sen-state	2.89	1.33	0.86

B1: <10%

#patterns

Domain	5	10	15	all
actors	5.61	12.05	17.05	21.27
books	2.75	9.29	13.02	22.66
directors	3.85	4.54	15.9	19.56
mayor	0.37	1.05	12.71	21.31
sen-party	30.1	49.1	61.8	71.2
sen-state	1.23	2.25	16.64	23.32

B2: <25%

Space-Coverage Tradeoff

Domain/#patterns	Top-5		top-15		All patterns	
	Overhead	Coverage	Overhead	Coverage	Overhead	Coverage
actors	30.2	72.7%	63.9	92.2%	113.2	100%
books	34.3	78.3%	61.2	96.3%	98.4	100%
directors	33.7	79.0%	55.7	93.5%	94.9	100%
sen-party	45.2	71.4%	69.1	84.4%	138	100%
sen-state	21.5	77.7%	52.7%	83.2%	115.2	100%

Space vs. Coverage: 15 patterns cover ~85%

Summary: Interactive Investigation of IIE

- Identify investigative operations for debugging
- Maintain auxiliary information using EBG
- Algorithms for efficient investigation

- **Ongoing work:**
  - Debugging for generic information extraction
  - Non-iterative pipelines
  - Under limited information about extraction operations

Thanks!

Anish Das Sarma
anish.dassarma@gmail.com