Uncovering the plot: Detecting Surprising Coalitions of Entities in Multi-relational Schemas

Hao Wu
Jilles Vreeken
Nikolaj Tatti
Naren Ramakrishnan

MOTIVATION

Knowledge discovery from multi-relational data

Intelligence Analysis

Persons

Locations

Phone Numbers

⋮

MOTIVATION

Knowledge discovery from multi-relational data

DNA

Proteins

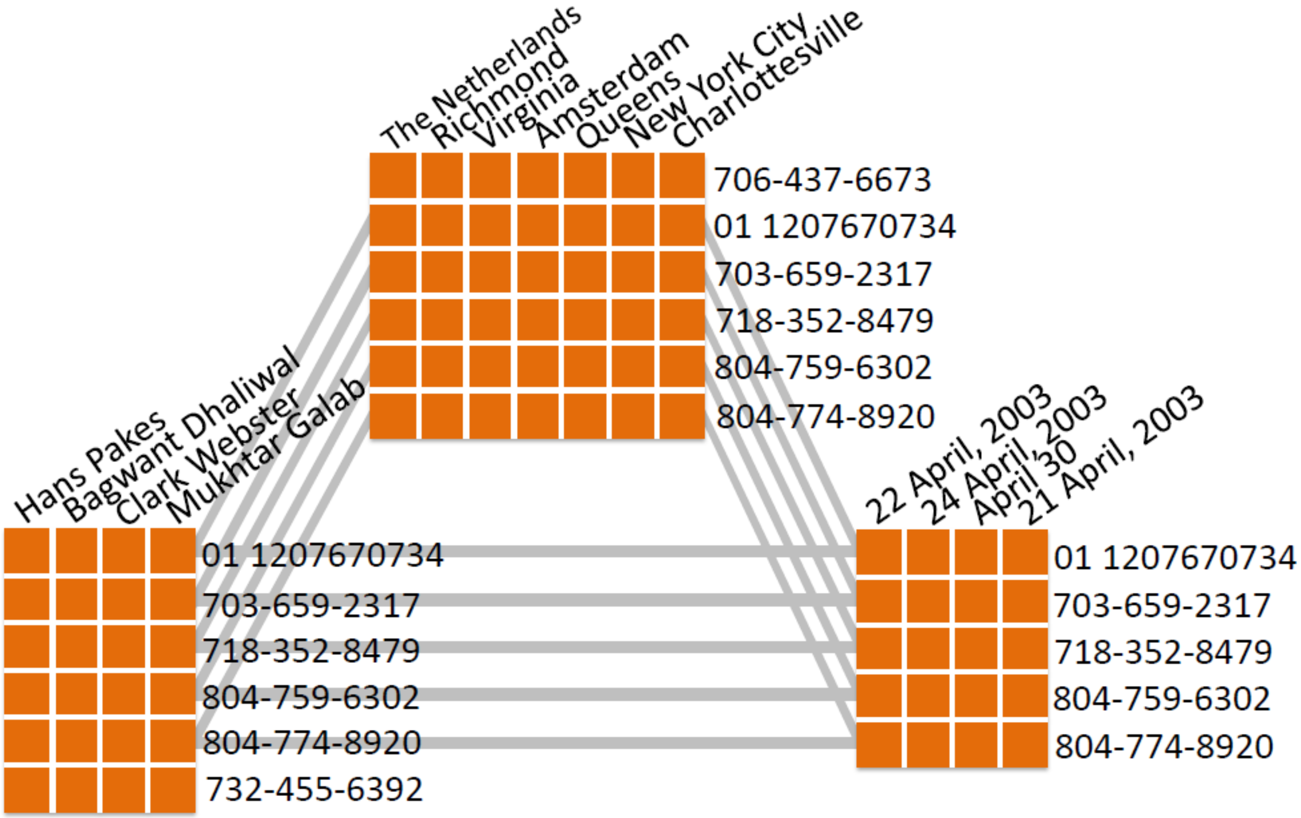
Pathways

⋮

Biological knowledge discovery

MOTIVATION

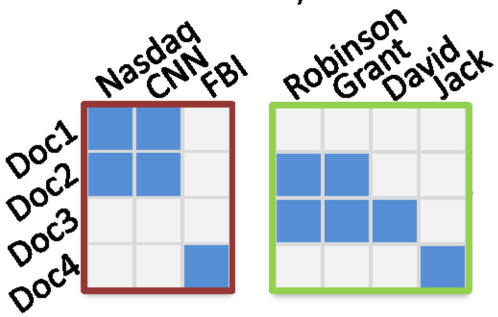
Automatically discover *surprising* multi-relational "3C" (coalitions, connections, & chains) patterns.



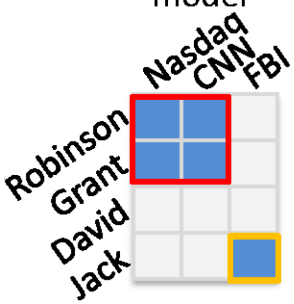
Structured and Unstructured Data

We consider **two** types of input data, or 'pattern spaces'

Doc-Entity model



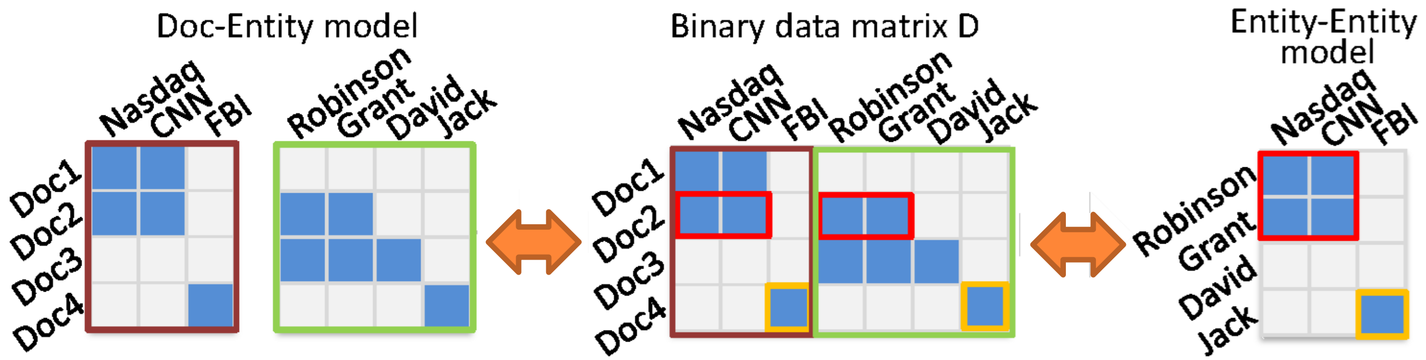
Entity-Entity model



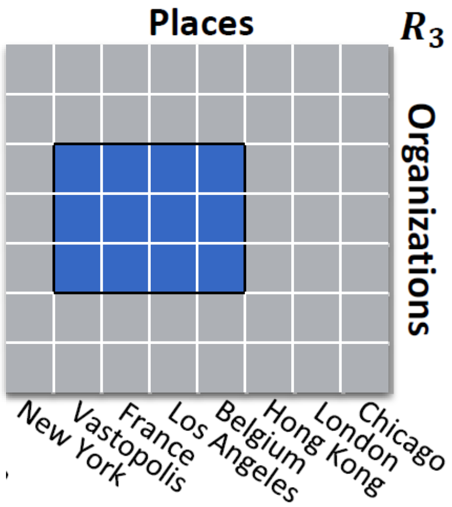
Structured and Unstructured Data

We consider **two** types of
input data, or 'pattern spaces'

by using a trick

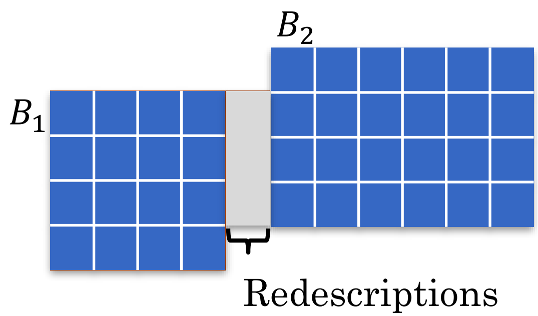


PATTERNS



Bicluster:
connected entity set

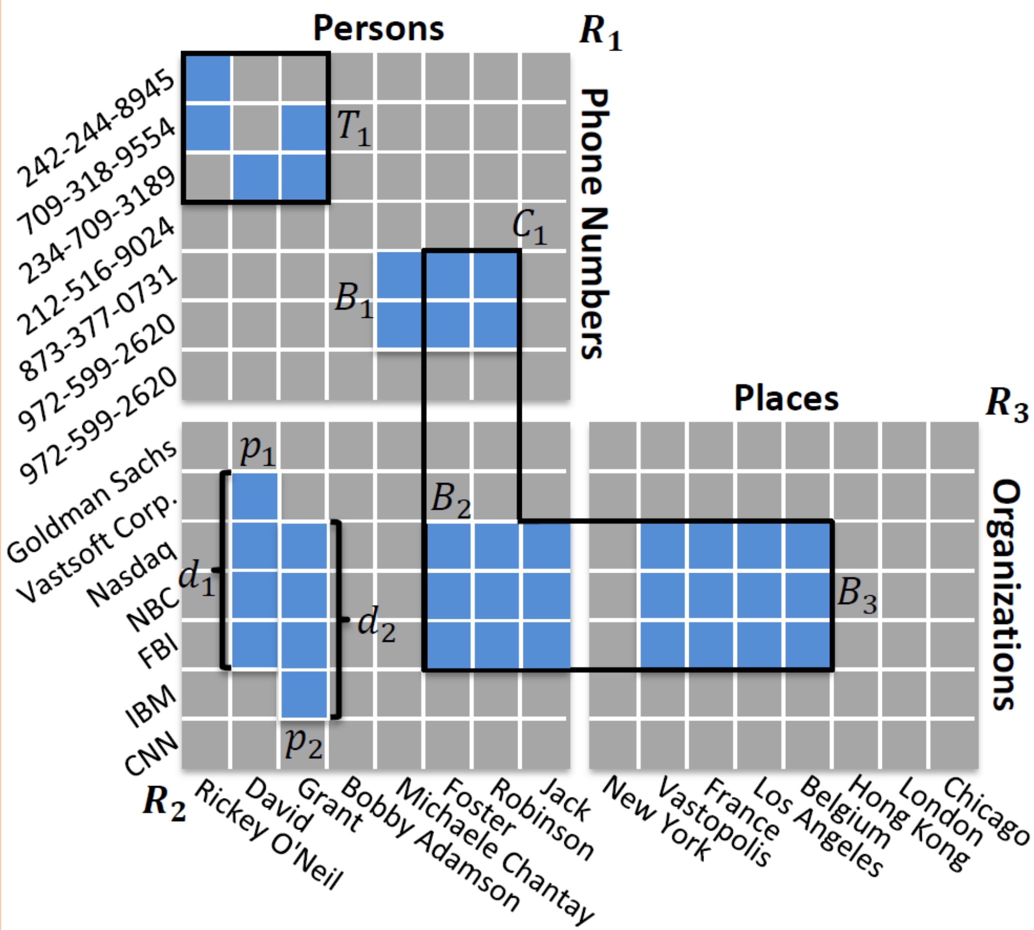
PATTERNS



Bicluster:
connected entity sets

Redescription:
bicluster pair identifying (roughly) the same entities for shared domain

PATTERNS

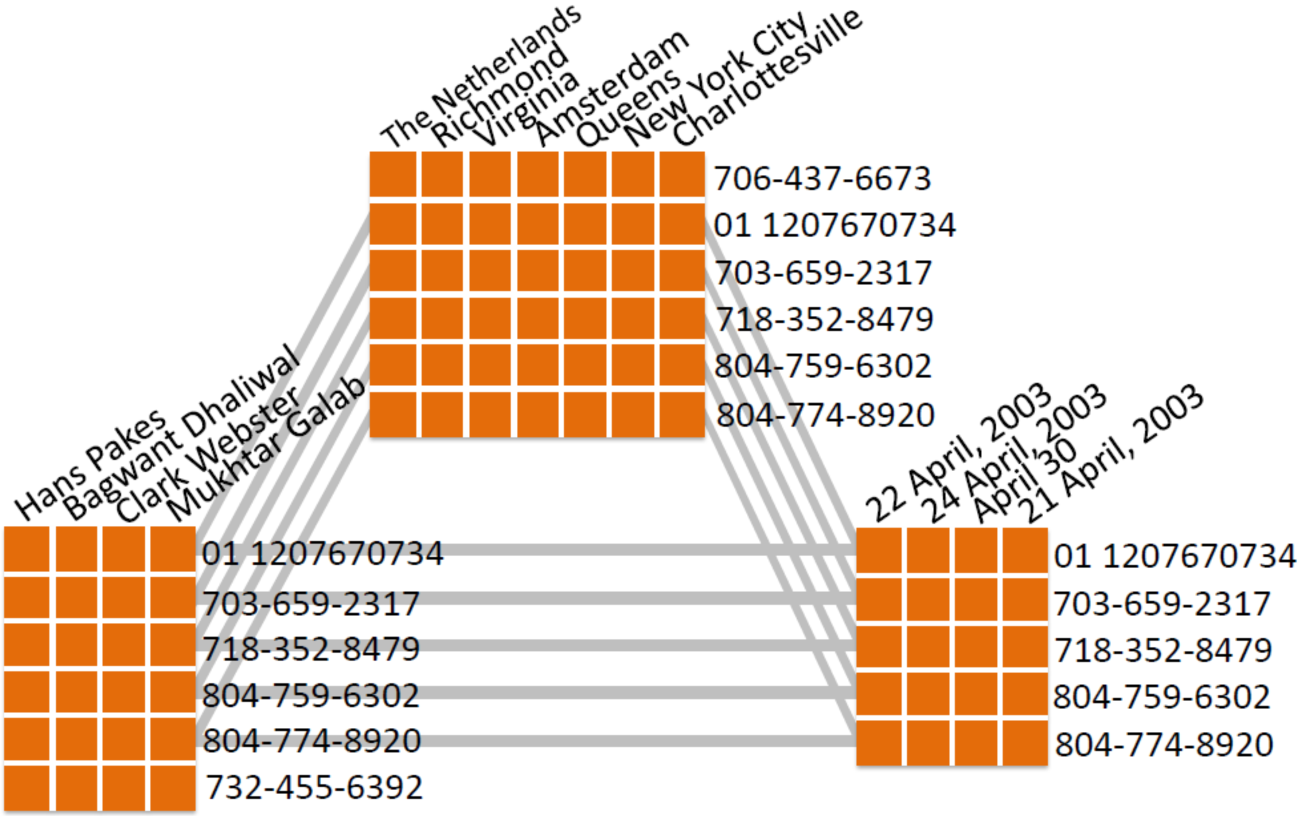


Bicluster:
connected entity sets

Redescription:
bicluster pair identifying (roughly) the same entities for shared domain

Bicluster Chain:
A **chain** of redescriptions

BICLUSTER CHAINS



SURPRISING PATTERNS

'Just mine biclusters!' – nope.

'Just mine redescriptions!' – better, but still nope.

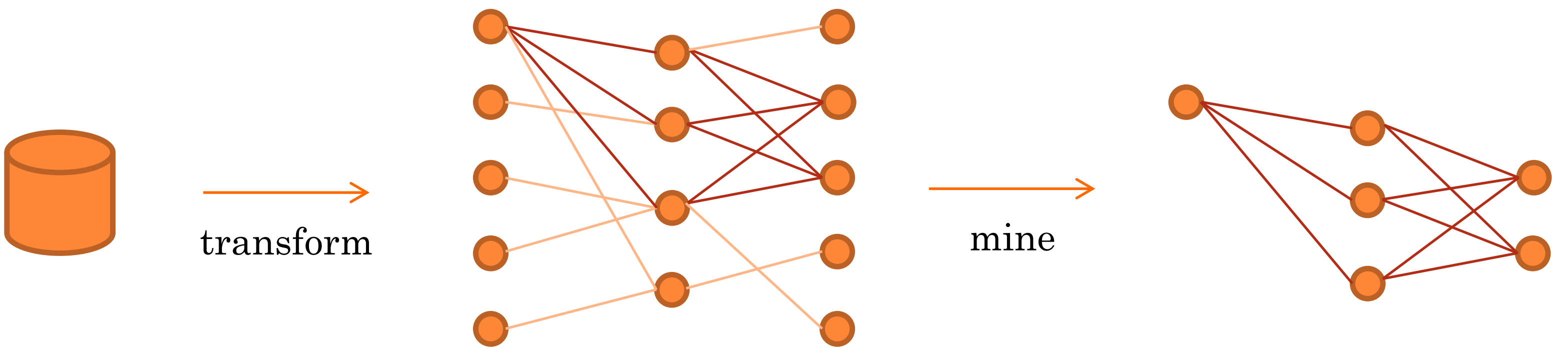
We are after *chains* of biclusters,
such that plots in the data are revealed

and, we want *only* those chains
that stand out
from what we already know

Related

Maximal Completely Connected Subgraphs

- Spyropoulou & De Bie (2011)



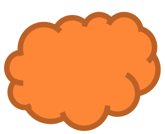
multi-relational database

K-partite graph

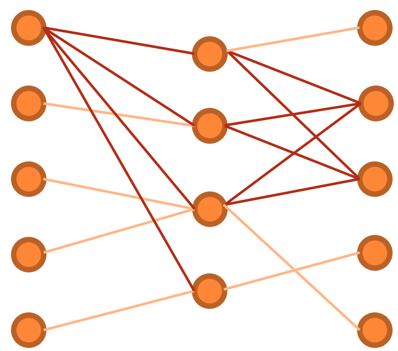
MCCS *surprising* wrt margins

CONNECTING TO MCCS

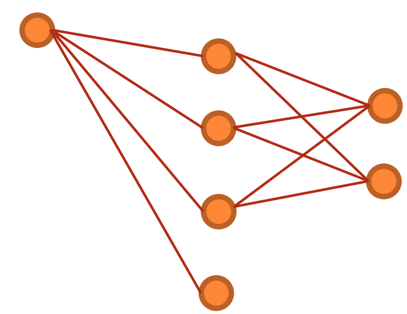
We mine chains of *redescriptions*



transform



mine



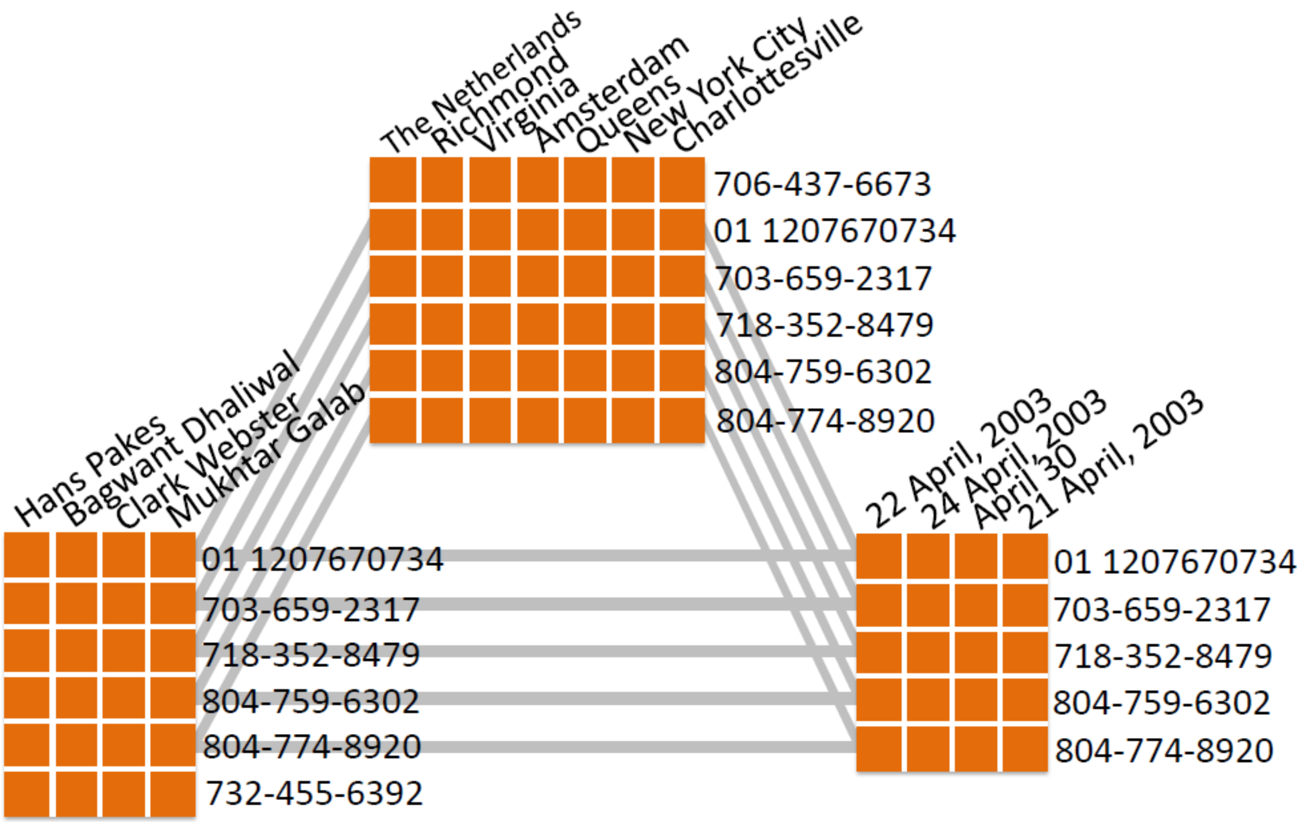
probabilistic model of the data

entity-entity graph
or
document-entity db

Chain of **redescriptions**
surprising wrt margins
and all mined chains

Motivation

Automatically discover *surprising* multi-relational "3C" (coalitions, connections, & chains) patterns.



ITERATIVE MINING

Knowledge *changes* during data analysis

- **interestingness** of chains changes depending on what results we study/reject

S*tatic* ranking of results is overly simplistic

- leads to redundancy – hides interesting results

How can we score results based on (accumulated) background knowledge?

What prior should we use?

MAXIMUM ENTROPY MODELLING

'the best distribution p^* satisfies the background knowledge, but makes **no further** assumptions'

very useful for data mining:

unbiased measurement of **subjective interestingness**

MaxEnt For Binary Data

Tiles

- A tuple of row IDs and column IDs from the given binary data matrix D.
- Frequency of a Tile

$$\gamma_T = fr(T;D) = \frac{1}{|\sigma(T)|} \sum_{(i,j)\in\sigma(T)} D(i,j)$$

where $D(i,j)$ represents the (i,j) entry in D, and $\sigma(T)$ represents the set of all the entries in tile T.

MaxEnt For Binary Data

Needed: MaxEnt model for tiles

- we use the model by Tatti & Vreeken (2011), De Bie (2011)

$$p^*_{\mathcal{T}} = \arg\max_{p \in \mathcal{P}} H(p)$$

where

$$\mathcal{P} = \{p \mid fr(T;p) = \gamma_T, \forall T \in \mathcal{T}\}$$

$$H(p) = -\sum_{D \in \mathcal{D}} p(D) \log p(D)$$

$$fr(T;p) = \frac{1}{|\sigma(T)|} \sum_{(i,j) \in \sigma(T)} p((i,j) = 1)$$

Background Knowledge

Background information in terms of Tiles

- $\mathcal{T}_{col}$: a set of column margin tiles
- $\mathcal{T}_{row}$: a set of row margin tiles *per entity domain*
- $\mathcal{T}_{dom}$: a set of entity domain tiles

$$\mathcal{T}_{back} = \mathcal{T}_{row} \cup \mathcal{T}_{col} \cup \mathcal{T}_{dom}$$

MEASURING SURPRISINGNESS

Evaluating a bicluster chain

1) Convert the chain into a set of tiles
   (depends on data model, see paper)
2) Infer the MaxEnt model
3) Calculate surprisingness through divergence

$$s_{global}(B) = KL(P_B || P_{back})$$

$$s_{local}(B) = -\sum_{T \in \mathcal{T}_B} \sum_{(i,j) \in \sigma(T)} \log p^*((i,j) = D(i,j))$$

GLOBAL VS LOCAL SCORE

$$s_{local}(B) = -\sum_{T \in \mathcal{T}_B} \sum_{(i,j) \in \sigma(T)} \log p^*((i,j) = D(i,j))$$

tile $\mathcal{T}_B$

.73	.94	.82	.89	.82	.46	.73	.61
.58	.88	.70	.80	.70	.30	.58	.45
.73	.94	.82	.89	.82	.46	.73	.61
.30	.70	.42	.55	.42	.12	.30	.20
.30	.70	.42	.55	.42	.12	.30	.20
.44	.80	.56	.69	.56	.19	.44	.31
.44	.80	.56	.69	.56	.19	.44	.31
.18	.54	.27	.39	.27	.06	.18	.11
.30	.70	.42	.55	.42	.12	.30	.20

tile $\mathcal{T}_B$

$$s_{global}(B) = KL(P_B || P_{back})$$

.86	.98	.92	.85	.77	.40	.67	.55
.75	.96	.85	.73	.62	.24	.50	.37
.86	.98	.92	.85	.77	.40	.67	.55
.44	.85	.60	.42	.30	.08	.21	.13
.20	.63	.31	.61	.48	.15	.36	.25
.30	.76	.45	.74	.63	.25	.50	.37
.30	.76	.45	.74	.63	.25	.50	.37
.11	.47	.19	.45	.32	.09	.22	.15
.20	.63	.31	.61	.48	.15	.36	.25

SEARCHING GOOD CHAINS

Super Naïve Strategy:

1) Mine all the biclusters!
2) Construct all the chains!
3) Evaluate all subsets of k chains!
4) Choose the most surprising set.

SEARCHING GOOD CHAINS

Slightly Less Naïve Strategy:

1) Mine all the biclusters!
2) Construct all the chains!
3) While not yet chosen k chains:
   evaluate each chain C against P_{back}
   greedily choose most surprising C
   $back \leftarrow back + C$, and infer P_{back}

SEARCHING GOOD CHAINS

Our strategy:

1) Mine all the biclusters!
2) **while** not yet mined k chains:
   find most surprising bicluster B_0,
   while there is a redescription B_i of $B_{i\text{-}1}$
      add most surprising B_i to chain
   $back \leftarrow back + C$, and re-infer P_{back}

Experiment Results

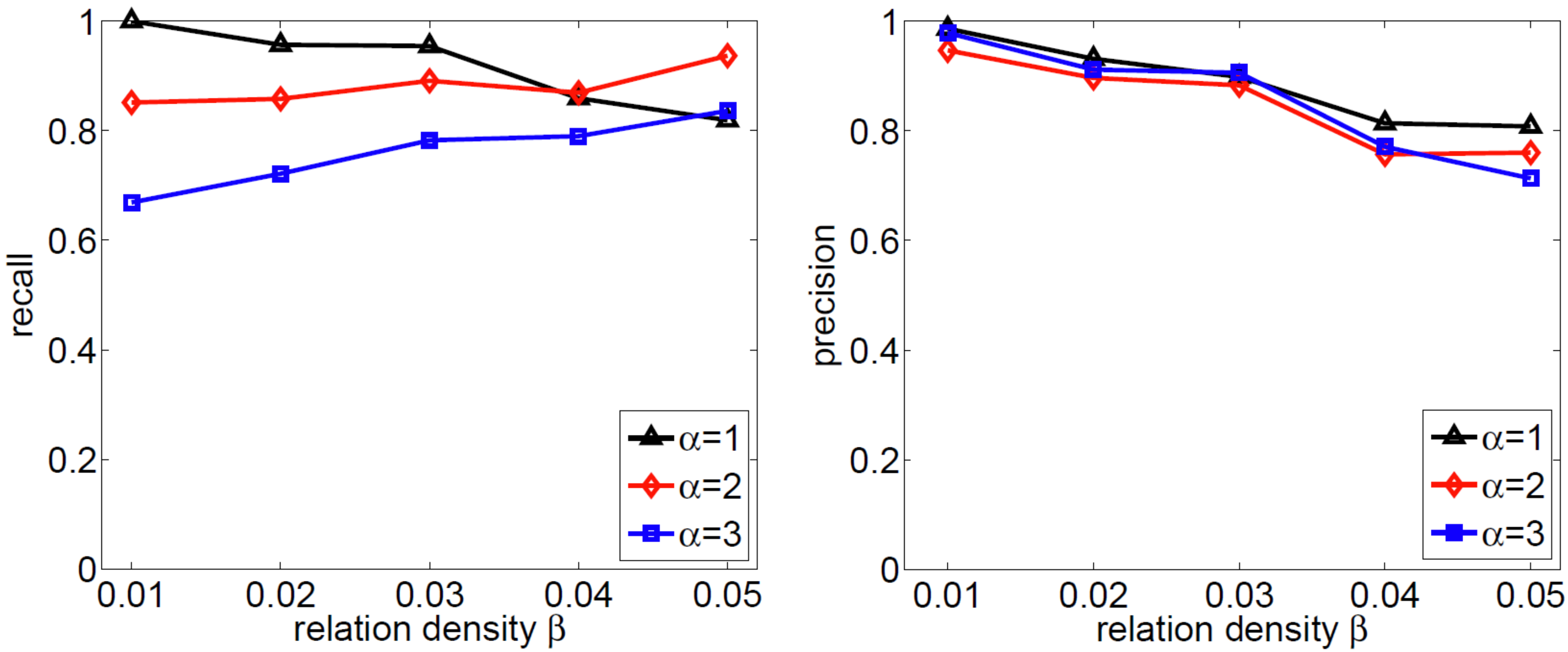
Datasets Statistics

Dataset	Number of Documents	Number of Entities	Doc–Entity	Entity–Entity	
			%1s	min %1s	max %1s
Synthetic 1k	1000	1000	0.01 — 0.05	0.01	0.05
Synthetic 2k	2000	2000	0.01 — 0.05	0.01	0.05
Synthetic 3k	3000	3000	0.01 — 0.05	0.01	0.05
Synthetic 5k	5000	5000	0.01 — 0.05	0.01	0.05
Synthetic 10k	10000	10000	0.01 — 0.05	0.01	0.05
Atlantic Storm	111	716	0.0179	0.0261	0.0608
Crescent	41	284	0.0425	0.0357	0.136
Manpad	47	143	0.0299	0.0385	0.0714

EXPERIMENT RESULTS

First things first: Synthetic Data

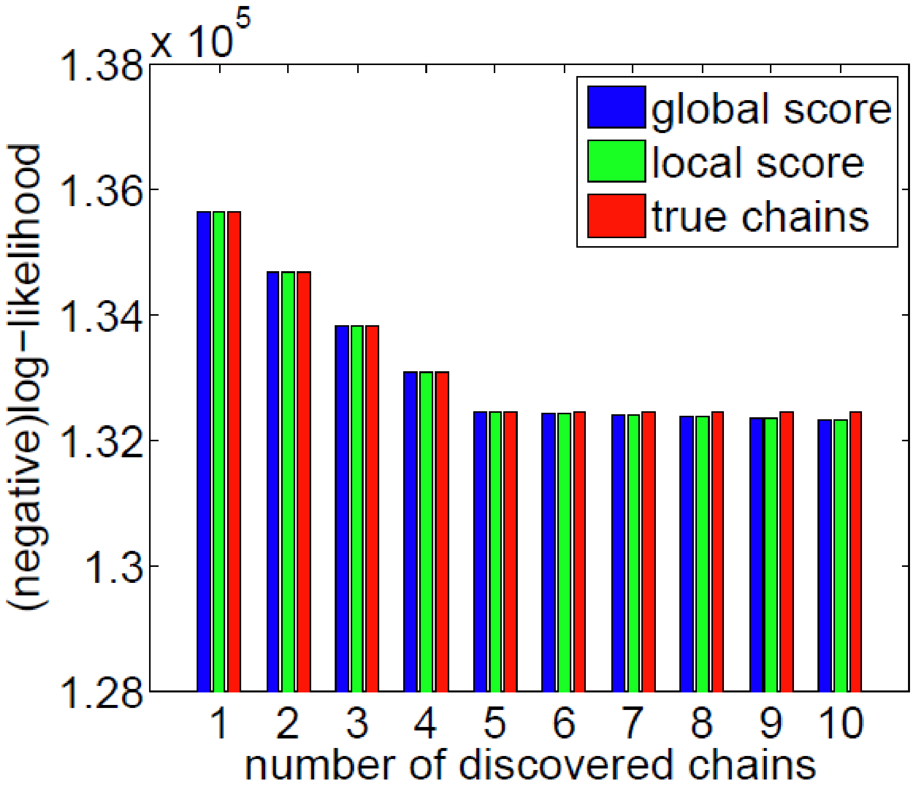
- can we uncover the plot?



Experiment Results

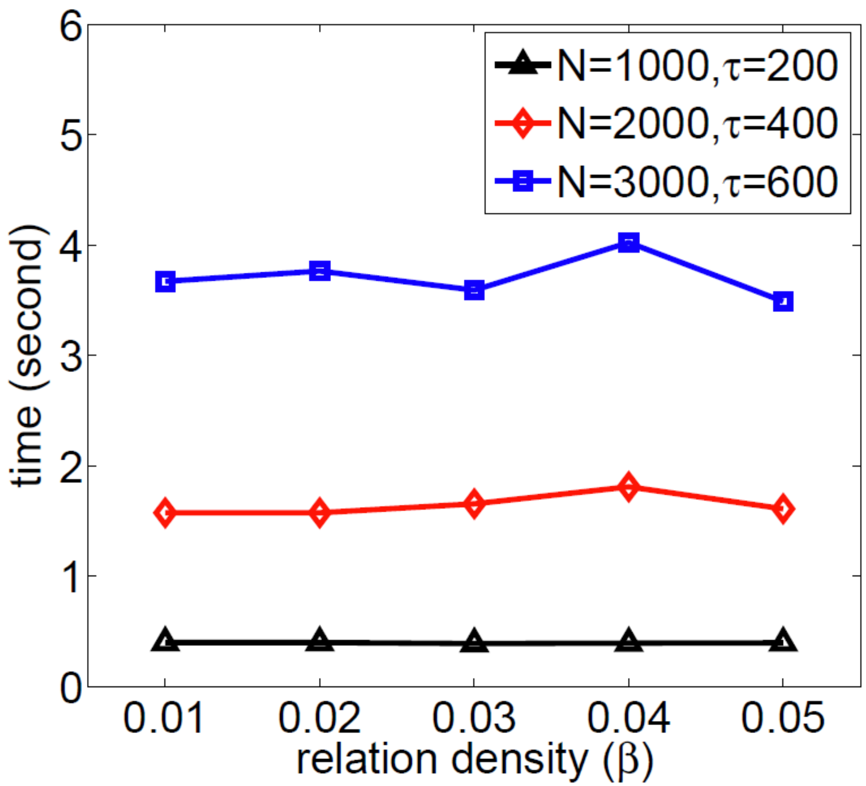
Second things second: Synthetic Data

- can we tell when to stop?

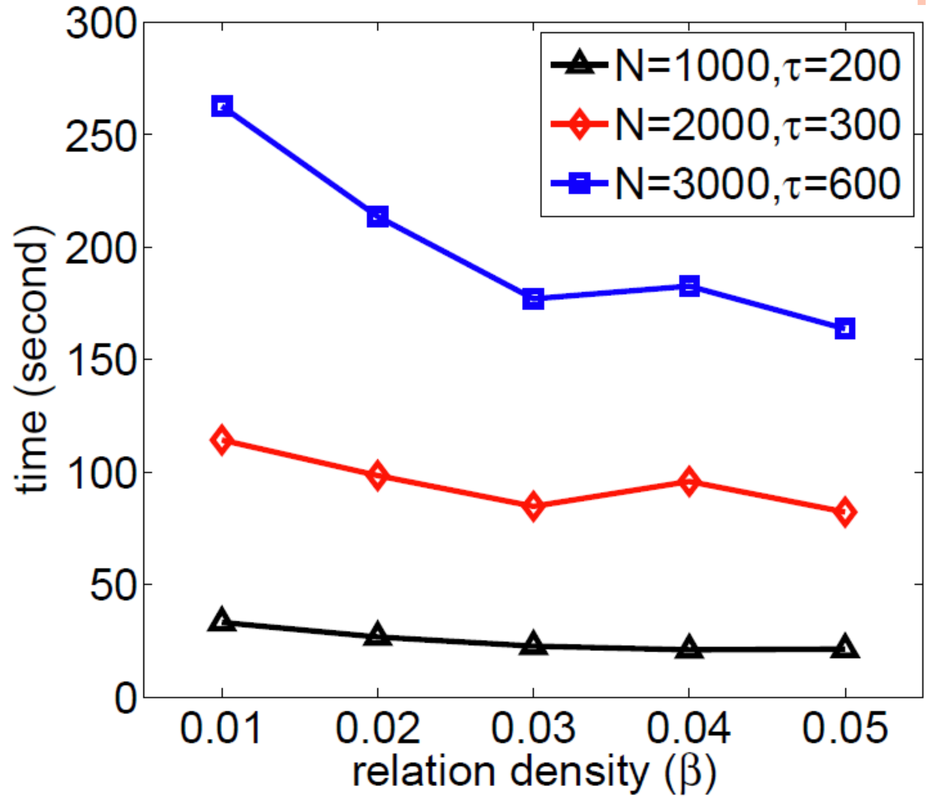


Experiment Results

Runtime Performance



Background model training time

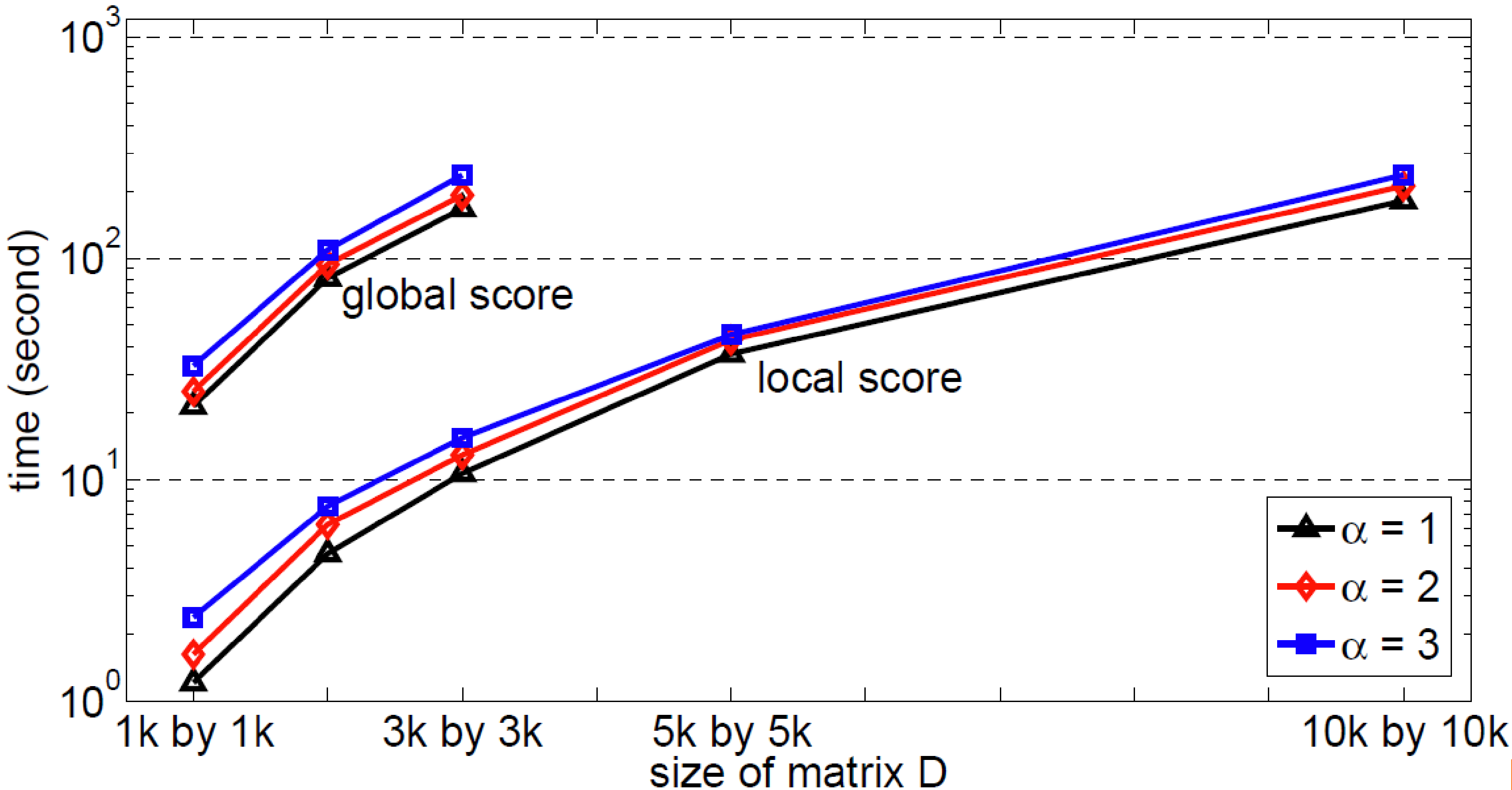


Total time

Discovery Analytics Center

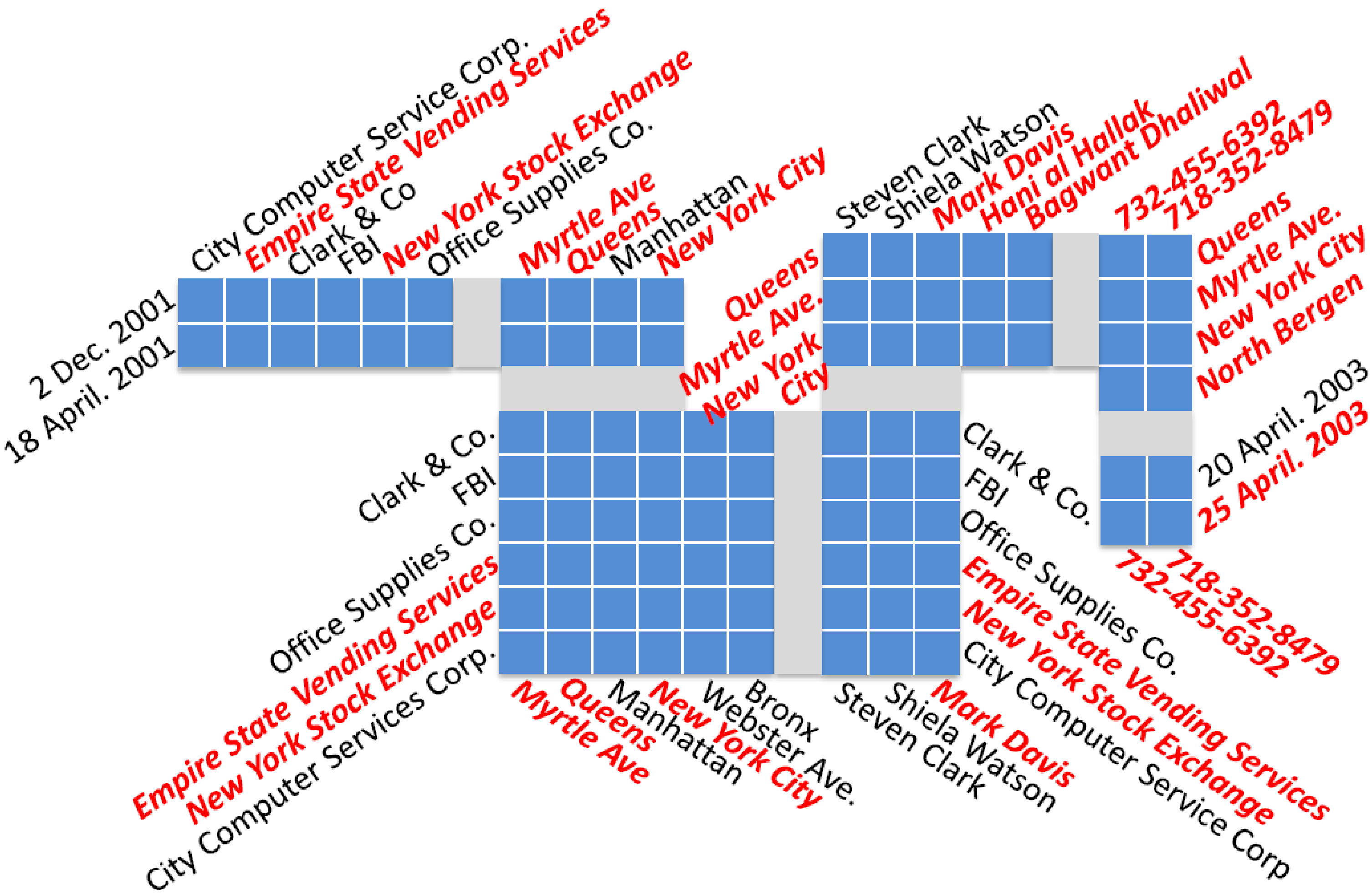
Experiment Results

- Global Score vs. Local Score



Experiment Results

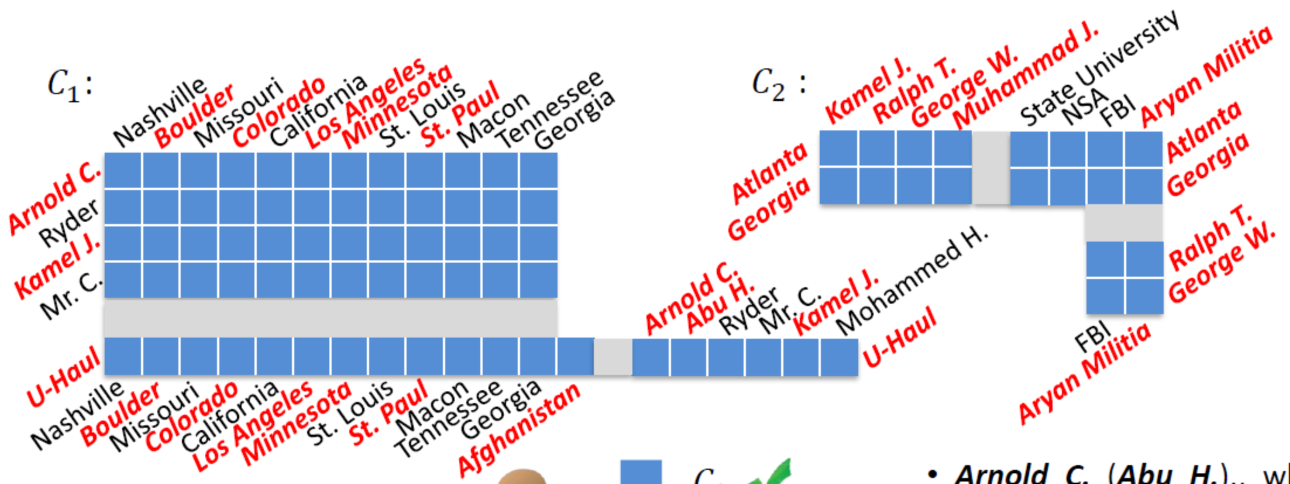
Real Data



Intelligence Analysis Dataset: *Crescent*

Experiment Results

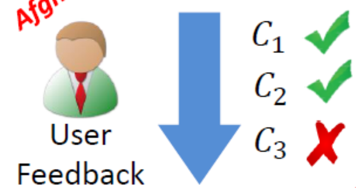
Iterative Knowledge Discovery

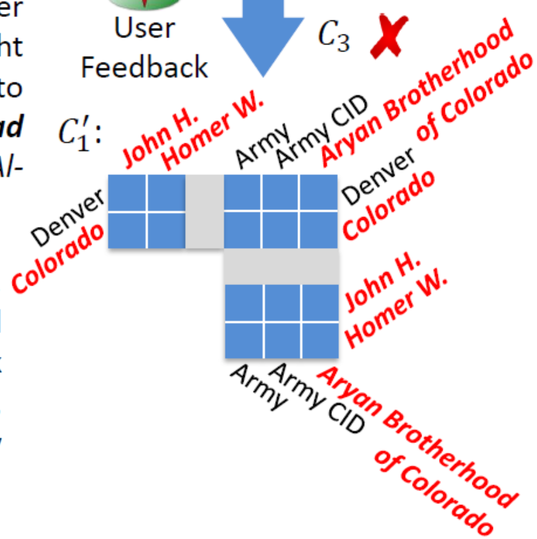


C_3: . . .

- ***Ralph T.***, who is a member of ***Aryan Militia***, bought weapons and sells them to ***George W.*** (***Muhammad J.***) who is a member of *Al-Queda*.

- ***Ralph T.*** meets ***Kamel J.*** in **Atlanta**, **Georgia**, and ***Kamel J.*** drives a truck from **Atlanta** to **St. Paul**, **Minnesota**. He probably transports weapons.





- ***Arnold C.*** (***Abu H.***)., who was a suspect of the 9/11 attack and spent time in **Afghanistan**, rents a **U-Hual** truck and drives it from **Boulder**, **Colorado** to **Los Angeles**. He probably transports the weapons.

- ***Homer W.***, who is a member of ***Aryan Brotherhood of Colorado***, sells the weapons to ***John H.***, who is a member of *Al-Queda*, in **Colorado**.

Conclusion

- Applicable to analyze multi-relational unstructured or discrete data
- Discover surprising entity coalitions with new data modeling primitives and algorithms
- Experiments on both synthetic and real datasets show that elaborate 'plots' can be detected
- Support human-in-loop iterative knowledge discovery

Thanks!

- Applicable to analyze multi-relational unstructured or discrete data
- Discover surprising entity coalitions with new data modeling primitives and algorithms
- Experiments on both synthetic and real datasets show that elaborate ‘plots’ can be detected
- Support human-in-loop iterative knowledge discovery