Generating Preview Tables for Entity Graphs

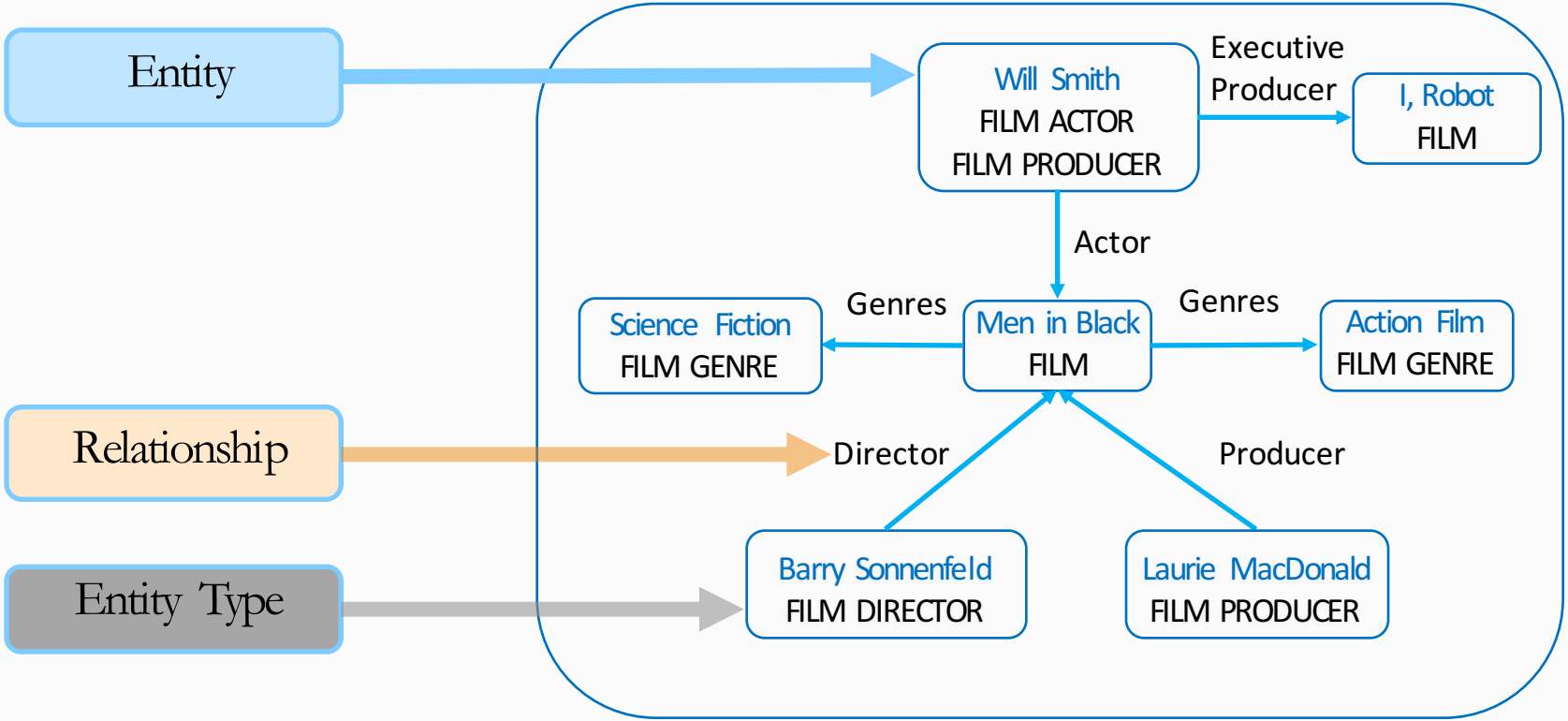
[#]Ning Yan, [*]Sona Hasani, [*]Abolfazl Asudeh, [*]Chengkai Li

[#]Huawei U.S. R&D Center

[*]University of Texas at Arlington, Innovative Database and Information Systems Research (IDIR) Laboratory

SIGMOD 2016, June 30th, 2016

Entity Graph

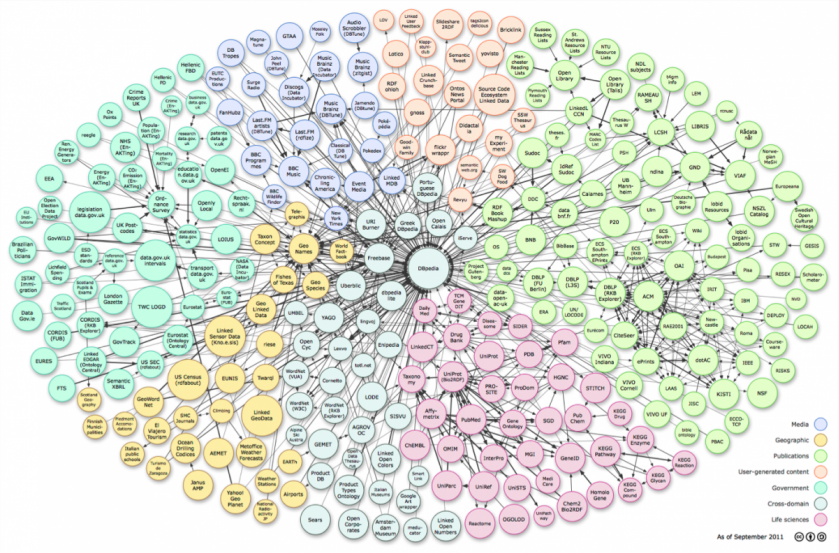


Ultra-heterogeneous Entity Graphs

Large and complex graphs capturing millions of entities and billions of relationships between entities.

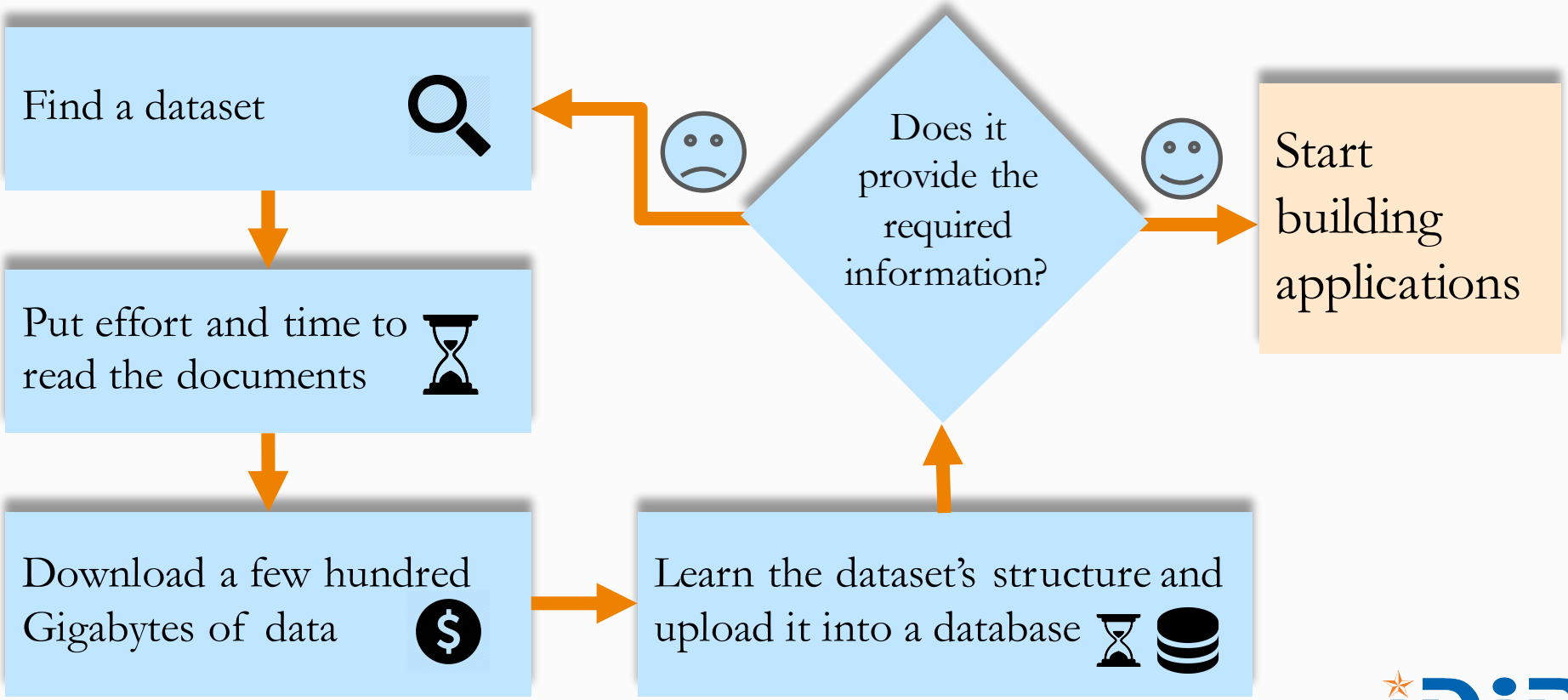
Applications: search, recommendation systems, business intelligence, health informatics, fact checking

Freebase : 1.9 billion triples
DBpedia : 3 billion triples
YAGO : 120 million triples
Linked Open Data :
hundreds of datasets
52 billion RDF triples



http://linkeddata.org/

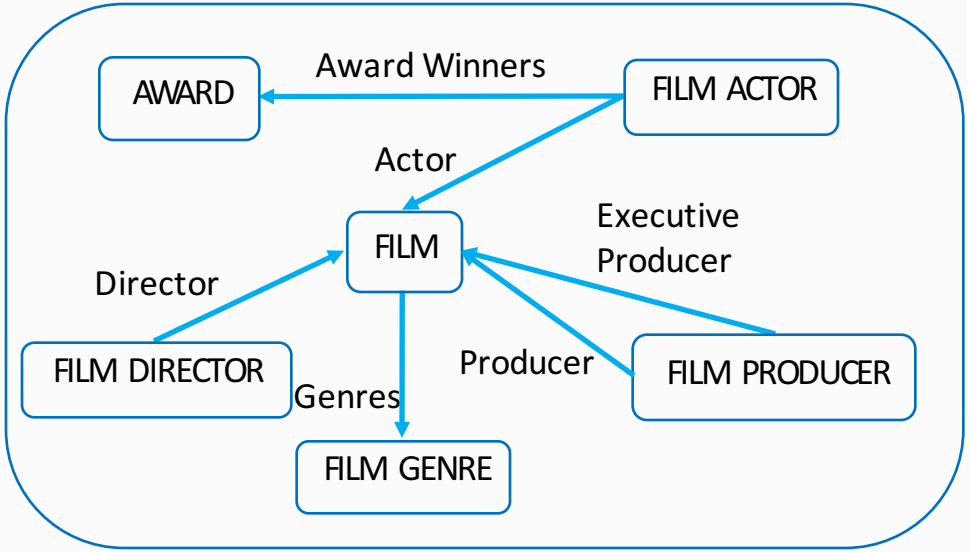
Steep Flag-Down Cost

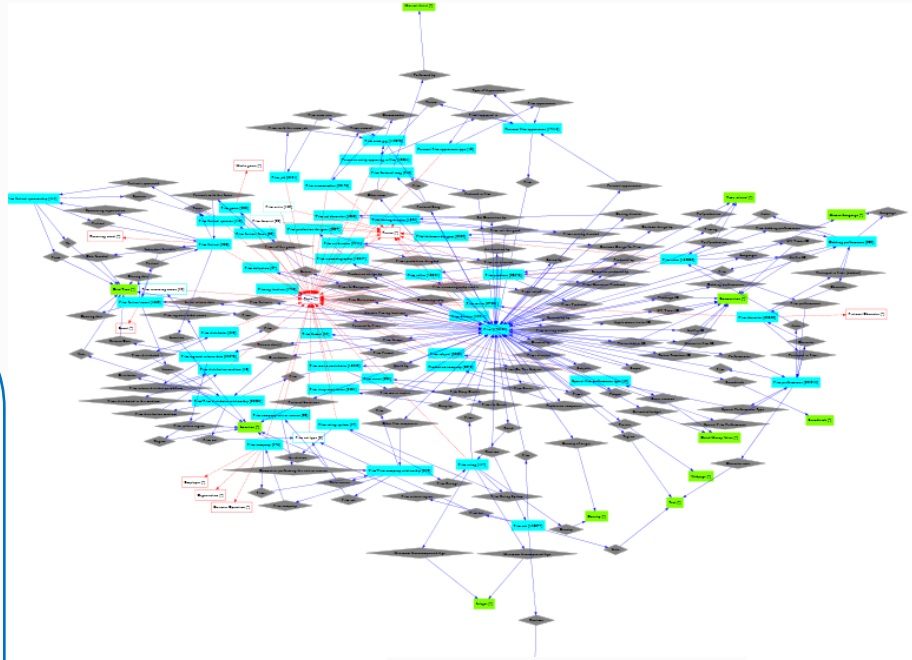


Need for a Quick Overview

Approach 1: Schema Graph

Schema graph itself can be too complex





Schema graph of "Film" domain in Freebase
Entity graph: 2M entities, 18 M edges
Schema graph: 63 entity types ,136 edges

Need for Quick Overview

Approach 2: Schema Summary

Schema summarization in relational database [Yang PVLDB09, Yang PVLDB11]

- Cluster tables in relational database by their semantic roles and similarities.
- Clusters tables, not relationships
- Detailed

XML summarization [Yu VLDB06]

- Provide a succinct overview of the entire schema graph

Graph summarization [Tian SIGMOD08, Zhang ICDE10]

- Group graph nodes based on their attribute similarity and allow users browse the summary from different grouping granularities.

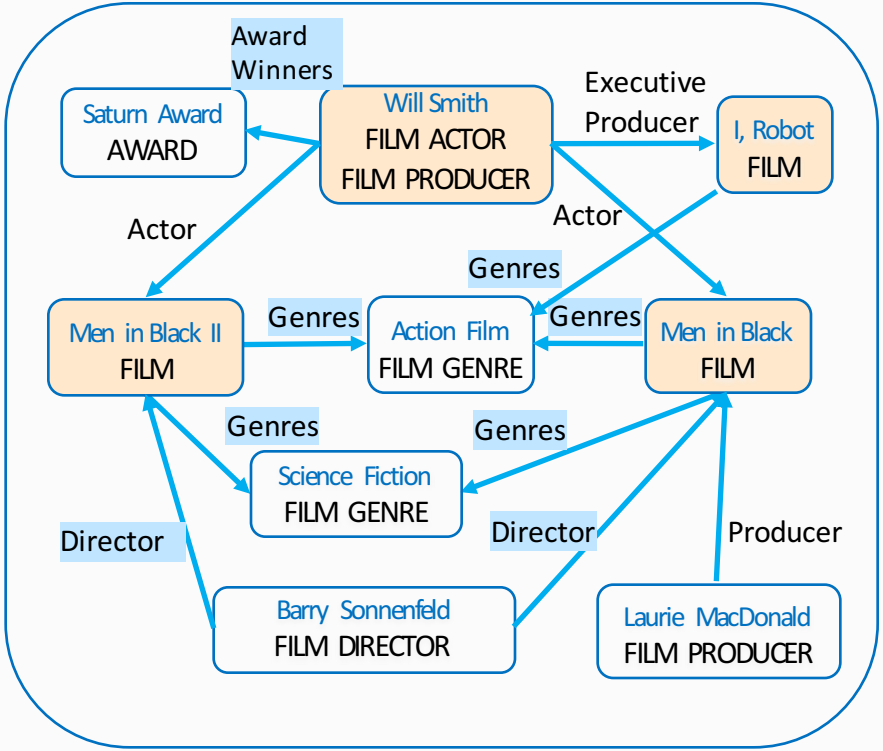
Preview Tables

Key attributes

Non-key attributes

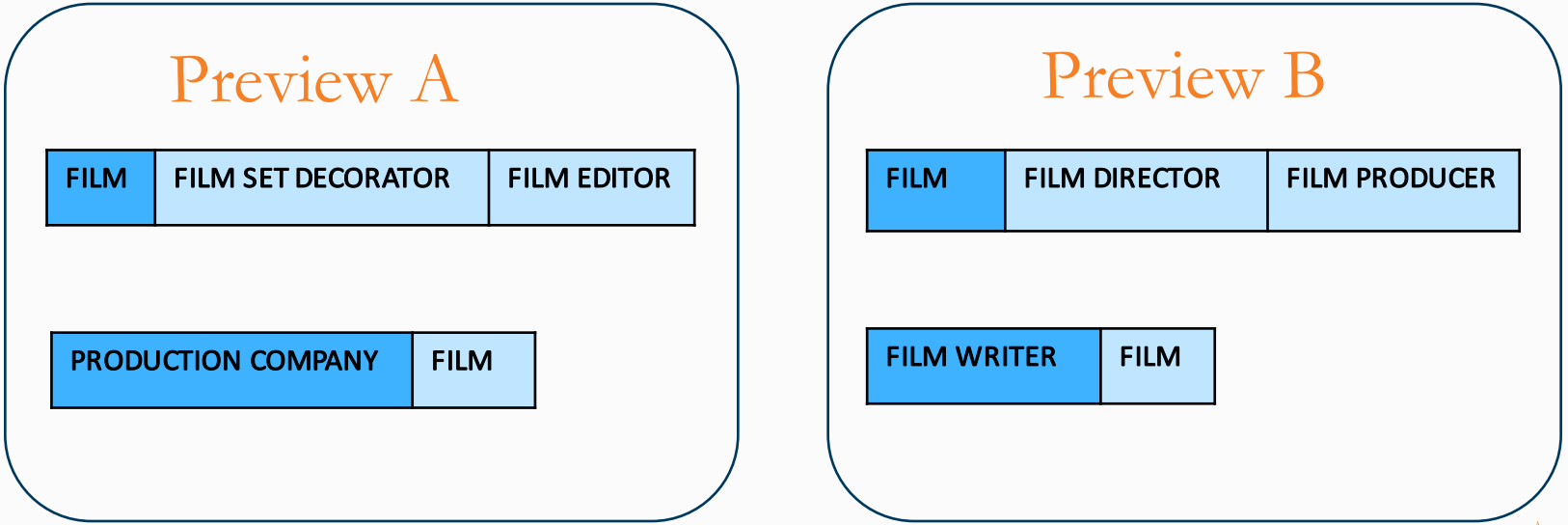
FILM ACTOR	Award Winners
Will Smith	Saturn Award

FILM	Director	Genres
Men in Black	Barry Sonnenfeld	{Action Film, Science Fiction}
Men in Black II	Barry Sonnenfeld	{Action Film, Science Fiction}
I, Robot	_	Action Film



Too Many Previews. Which one to Choose?

- Many possible previews
- Different choices



Aggregate Scoring

Score of the Table? Score of the Preview?

FILM	Actor	Genres
4	6	5

$4 \times (6+5) = 44$

FILM ACTOR	Actor	Award Winners
2	6	2

$2 \times (6+2) = 16$

$44+16= 60$

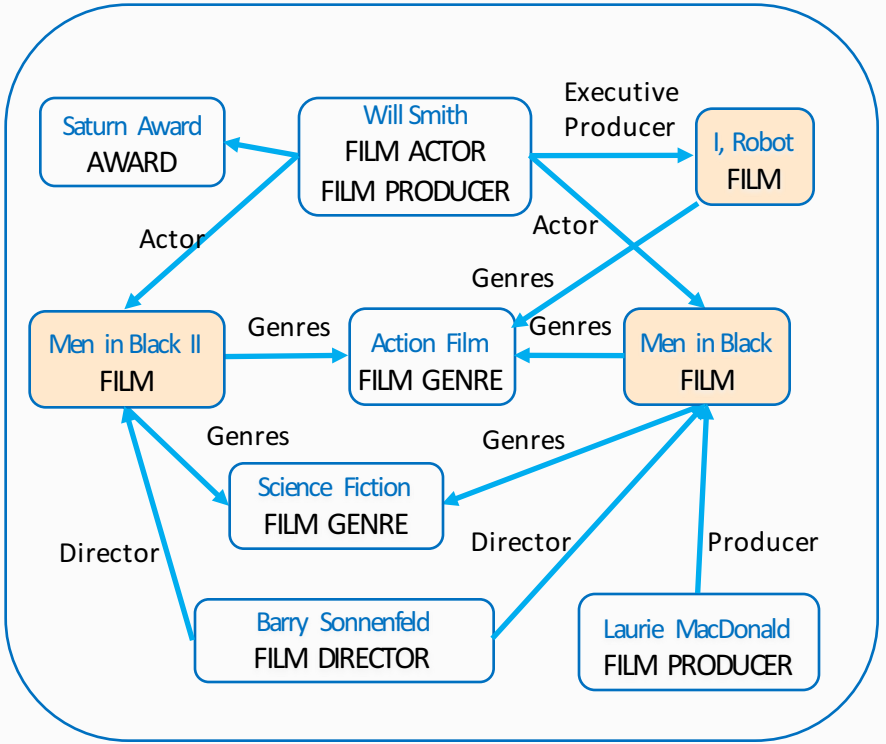
Key Attributes Scoring

Coverage-based method

Coverage(FILM) = 3

Random walk-based method

Stationary distribution of a random walk process defined over the schema graph



Non-key Attributes Scoring

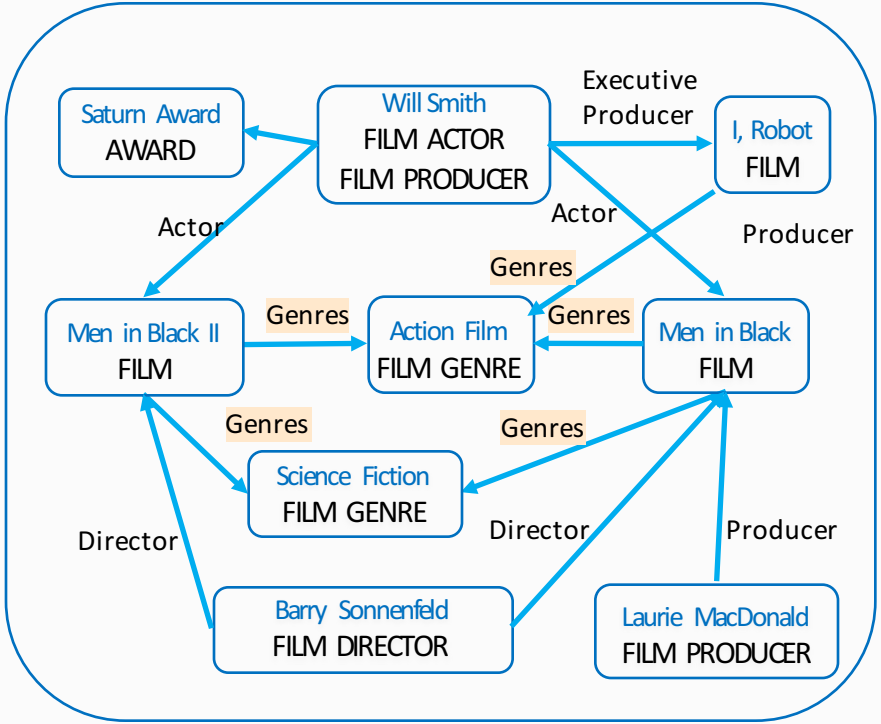
Coverage-based method

Coverage(Genres) = 5

Entropy-based method

Entropy(Genres) = (2/3) log(3/2)+(1/3) log(3/1)
= 0.28

FILM	Director	Genres
Men in Black	Barry Sonnenfeld	{Action Film, Science Fiction}
Men in Black II	Barry Sonnenfeld	{Action Film, Science Fiction}
I, Robot	_	{Action Film}



Optimal Preview Discovery

Find the preview with highest score that satisfies

- Size constraint
  - Number of key attributes *K*
  - Number of non-key attributes *N*

  } Concise

- Distance between two preview tables *d*

} Tight: dist(Ti, Tj) ≤ d

} Diverse: dist(Ti, Tj) ≥ d

FILM	Performances	Genres	Directed By
FILM DIRECTOR	Films Directed		
FILM PRODUCER	Films Produced		

Tight

FILM FESTIVAL	Location	Focus
FILM COMPANY	Films	
FILM CHARACTER	Portrayed in Film	

Diverse

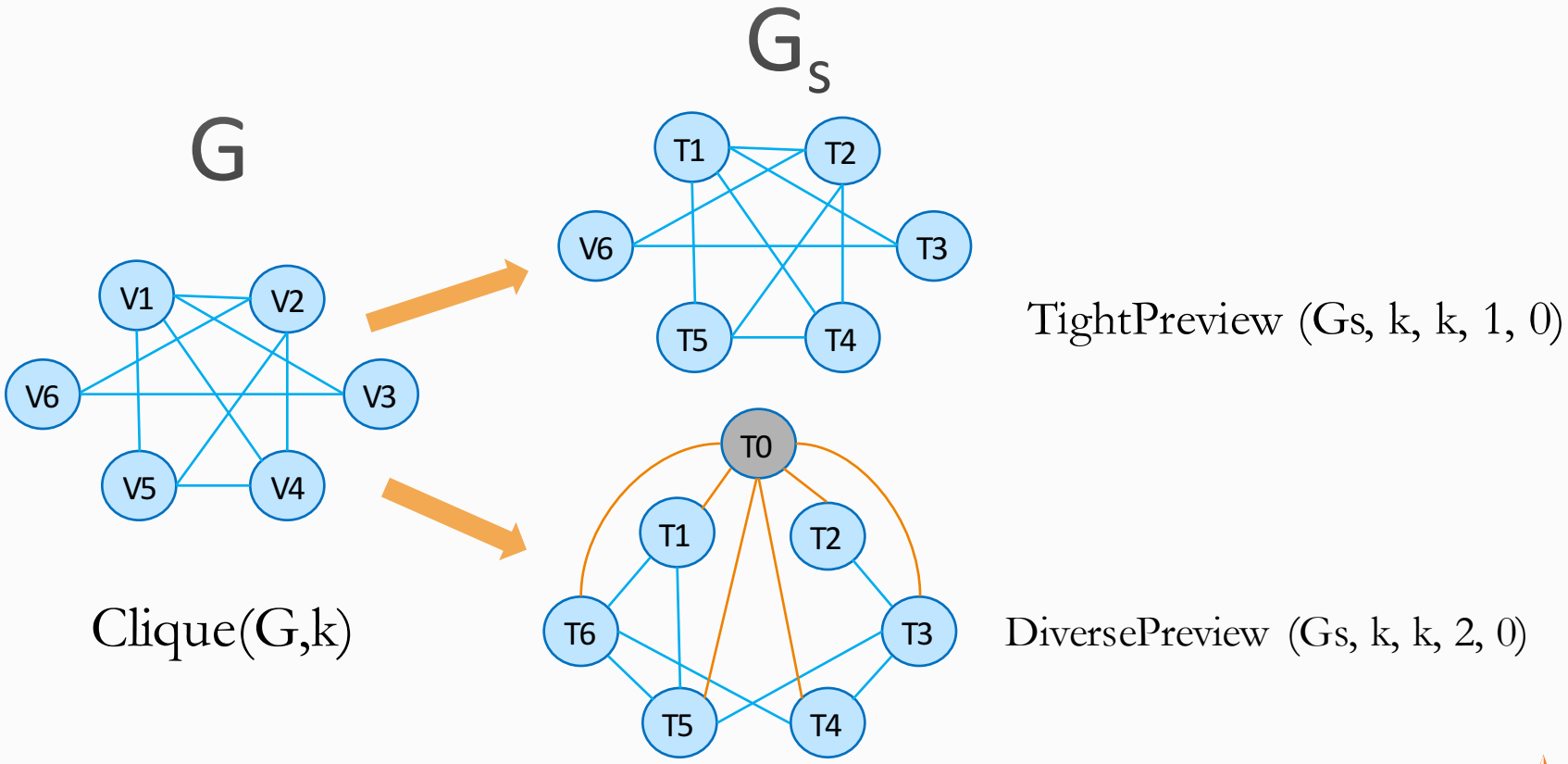
Concise Preview, Dynamic Programming Algorithm

We assume all K key attributes are ordered arbitrarily.

optimal concise preview (k, n, X) is the best of:

- optimal concise preview (k, n, X-1)
- optimal concise preview (k-1, n-1, X-1) ∪ X-th Key-attribute with 1 non-key attribute
- optimal concise preview (k-1, n-2, X-1) ∪ X-th Key-attribute with 2 non-key attributes
- … …
- optimal concise preview (k-1, k-1, X-1) ∪ X-th Key-attribute with (n-k+1) non-key attributes

Tight/Diverse Preview, NP-hardness



Tight/Diverse Preview, Apriori Property Algorithm

1. Construct 2-cliques by enumerating all key attribute pairs

2. for i = 3 to k

   generate i-cliques from (i-1)-cliques based on Apriori property

3. find the k-clique with highest score, **return** as optimal preview

Experiments

Dataset: Freebase

Accuracy of scoring measures

- Compared with Freebase ground truth:
  - Key attributes: Precision-at-k (p@k), Average Precision, Discounted Cumulative Gain (nDCG)
  - Non-key attributes: Mean Reciprocal Rank (MRR)
- Compared with crowd ranking:
  - Pearson Correlation Coefficient (PCC)

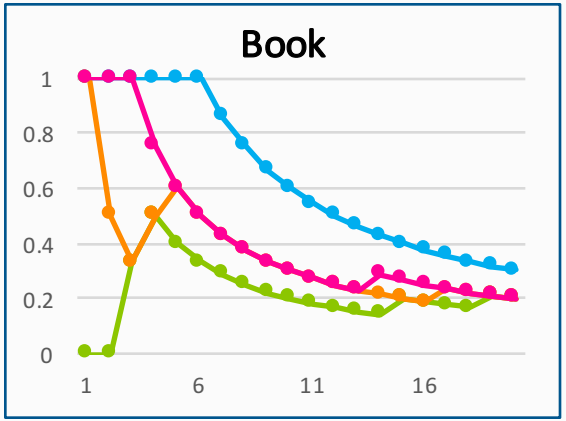
Efficiency of algorithms

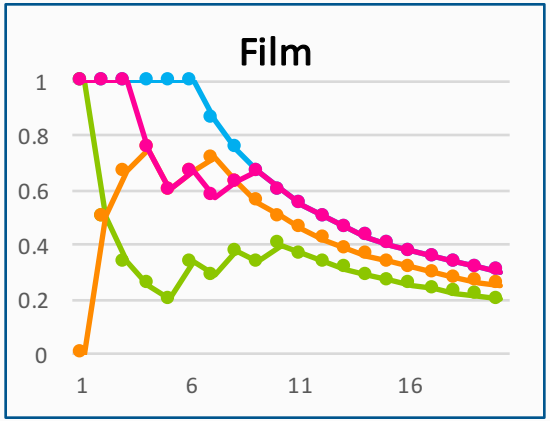
- Execution time

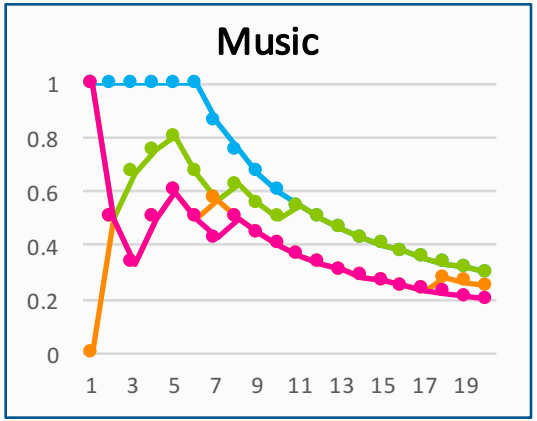
User study

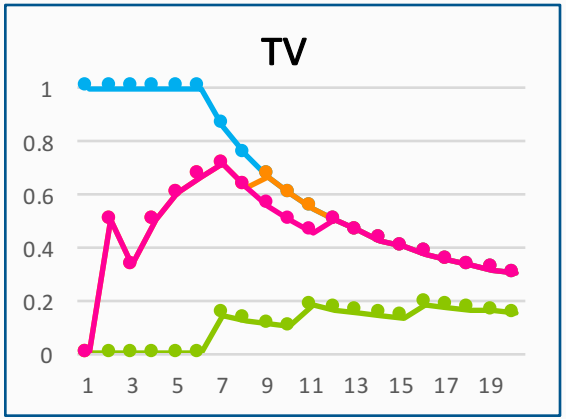
- Existence test questions
- User experience questions

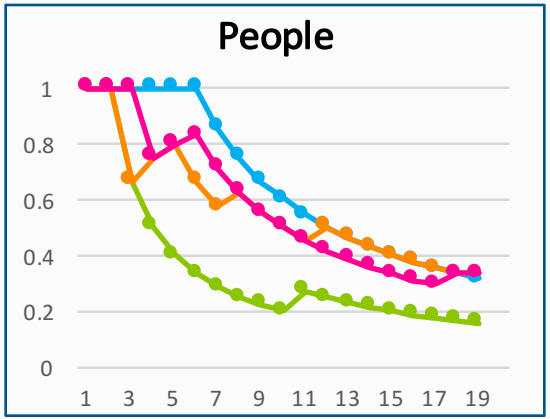
Key Attribute Scoring (p@k)

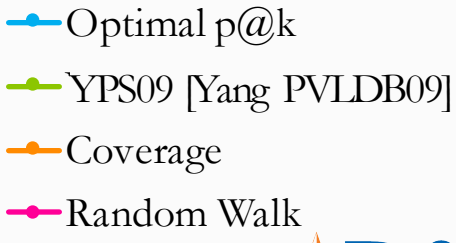












MRR of Non-key Attributes (MRR)

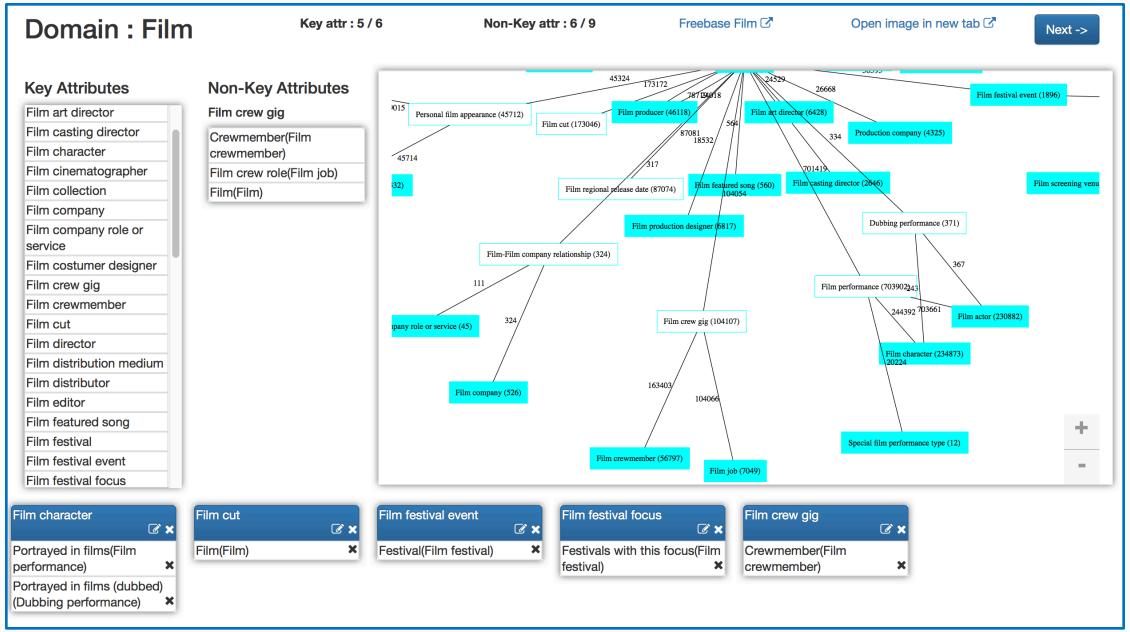
Domain	Coverage	Entropy
books	0.8	0.786
film	0.2	0.25
music	0.528	0.589
TV	0.622	0.379
people	0.708	0.606

Compare Rankings by Our Approach and the Crowd (PCC)

	Key Attribute			Non-key Attribute	
Domain	YPS09	Coverage	Random Walk	Coverage	Entropy
books	0.4	0.55	0.43	0.43	0.43
film	-0.01	0.48	0.25	0.35	0.35
music	0.37	0.33	0.46	0.42	0.41
TV	0.37	0.69	0.65	0.47	0.47
people	0.36	0.31	0.29	0.43	0.43

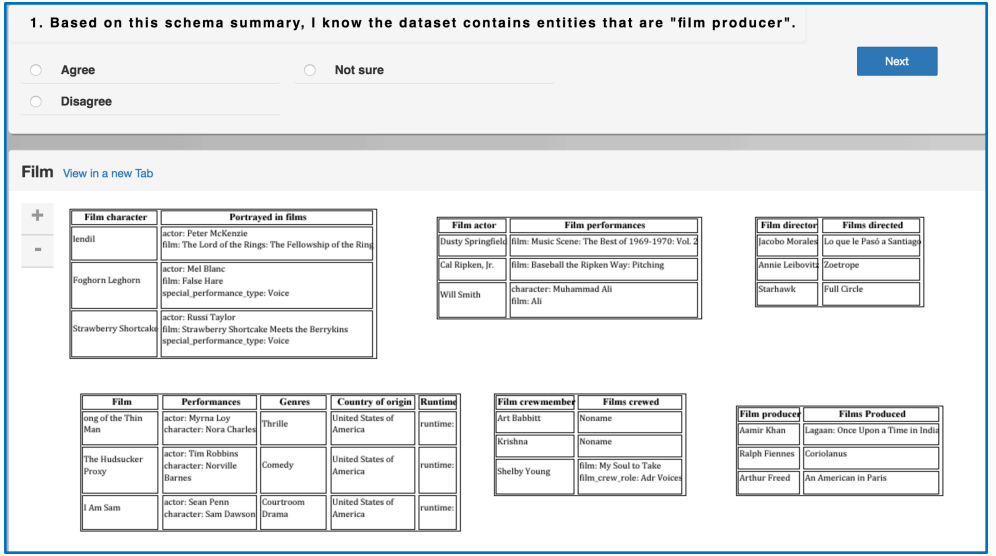
User Study, Hand-crafted Preview Tables

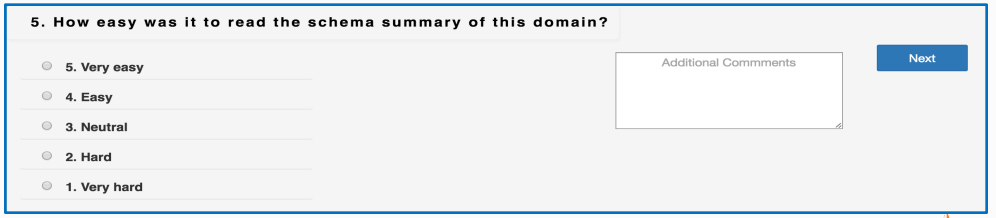
- Domains:
  - film
  - books
  - music
  - TV
  - people
- Hand-crafted preview tables
- 10 PhD students in Database research group
- Individually and as a group
- $20 gift card



User Study, Approaches Compared

- Domains:
  film, books, music, TV, people
- Approaches:
  - Schema graph
  - Concise preview
  - Tight preview
  - Diverse preview
  - Freebase ground truth
  - YPS09
  - Hand-crafted preview tables
- 4 existence questions
- 4 experience questions
- 84 Master's and PhD students in database area
- $15 gift card



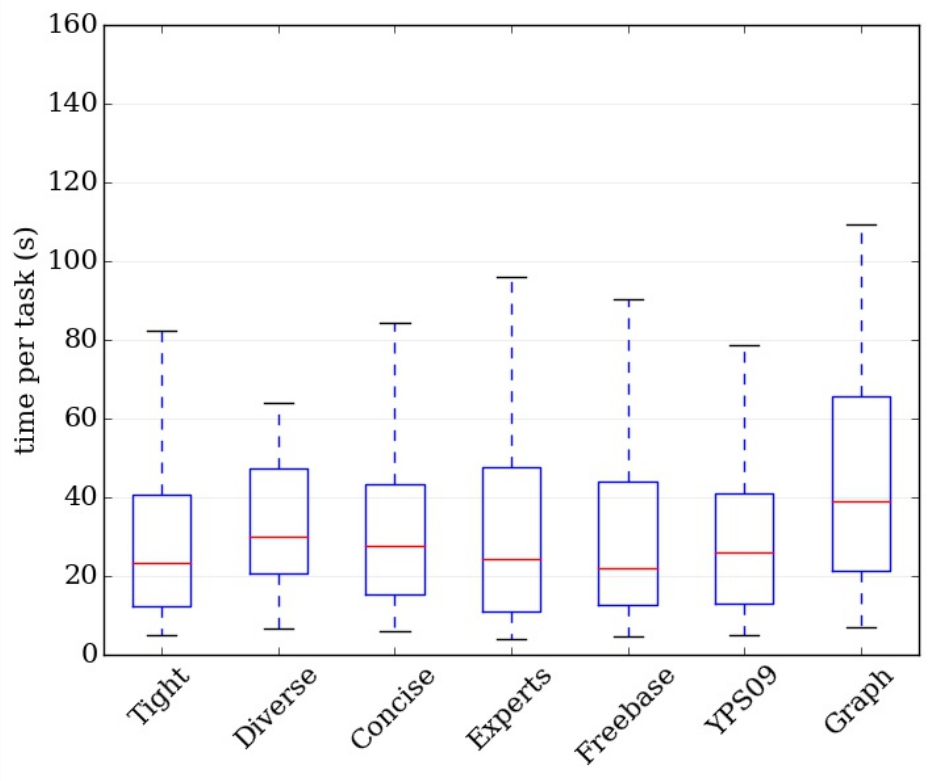


User Study: Existence Test Questions

	Tight	Diverse	Freebase	Experts	YPS09	Schema Graph
Concise	z=1.59 p=0.0559	z=−2.28 p=0.0113	z=0.49 p=0.3121	z=−0.13 p=0.4483	z=0.36 p=0.3594	z=−0.43 p=0.3336
Tight		z=−3.48 p=0.0003	z=−1.12 p=0.1314	z=−1.69 p=0.0455	z=−1.282 p=0.0999	z=−1.93 p=0.0268
Diverse			z=2.57 p=0.0051	z=2.10 p=0.0179	z=2.60 p=0.0047	z=1.70 p=0.0446
Freebase				z=−0.61 p=0.2709	z=−0.15 p=0.4404	z=−0.87 p=0.1922
Experts					z=0.49 p=0.3121	z=−0.29 p=0.3859
YPS09						z=−0.77 p=0.2206

Pairwise comparisons of conversion rates, domain="music", α=0.1

User Study: Existence Test Questions



Time taken on existence tests, domain="music"

User Study: User Experience Questions

Questions	most favorable → least favorable						
How easy was it to read the schema summary?	Freebase	Diverse	Graph	Experts	YPS09	Concise	Tight
How much understanding of the data can you gain from it?	Graph	Freebase	YPS09	Diverse	Concise	Tight	Experts
How helpful was it in assisting you to understand the data?	Graph	Freebase	YPS09	Diverse	Experts	Concise	Tight
Is it missing important information?	YPS09	Concise	Experts	Graph	Tight	Freebase	Diverse

Systems sorted by average user experience scores across five domains

Acknowledgment

Disclaimer: This material is based upon work partially supported by the National Science Foundation Grants 1018865, 1408928 and the National Natural Science Foundation of China Grant 61370019. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

Thank You! Questions?

Generating Preview Tables for Entity Graphs

Ning Yan, Sona Hasani, Abolfazl Asudeh, Chengkai Li