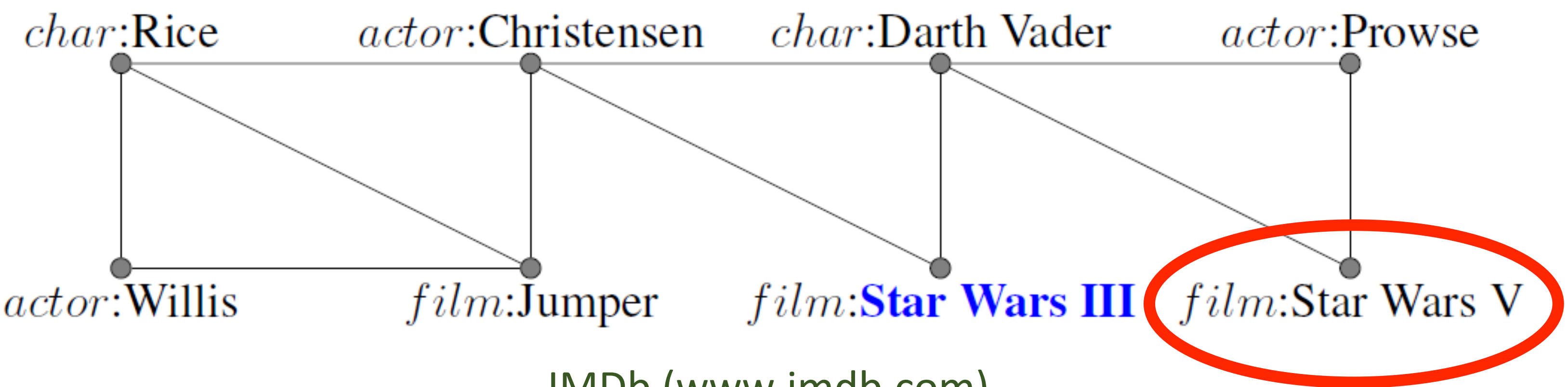
Towards Representation-Independent Graph Querying & Analytics

Arash Termehchy, Oregon State University

Searching for interesting relationships over graph data

Finding related or similar entities to an entity

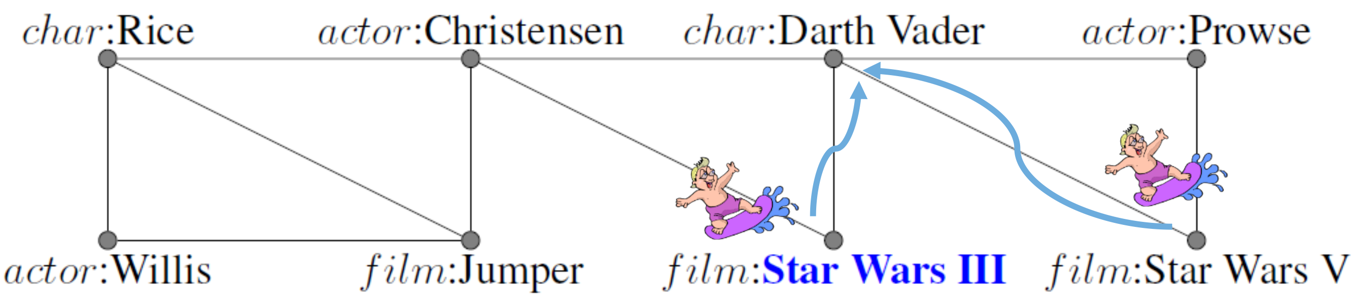
E.g., find similar movies to the movie "Star Wars III"



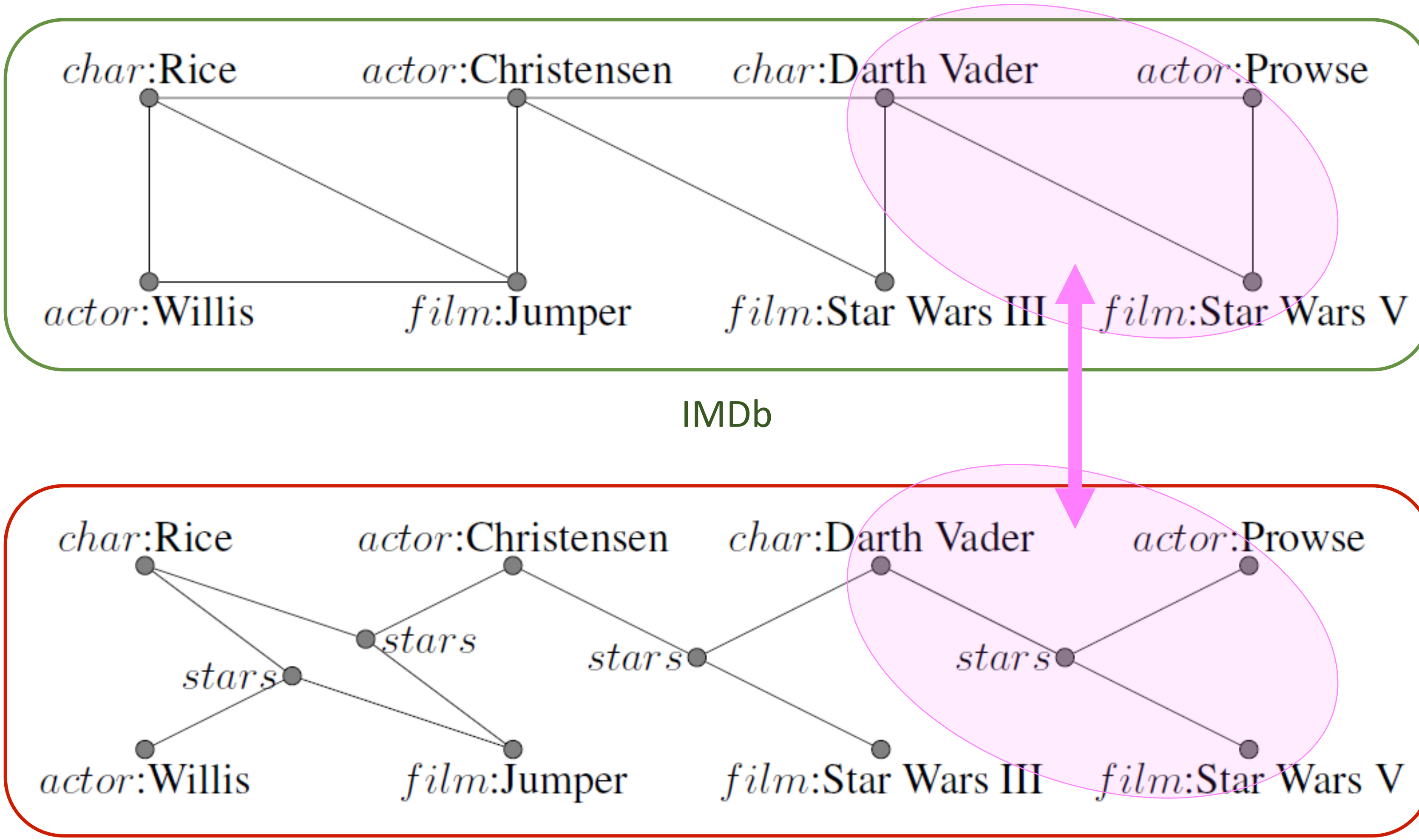
IMDb (www.imdb.com)

Algorithms use the graph structure to quantify similarity

- **SimRank**: two objects are similar, if they are referenced by similar objects.
  - how likely two random surfers will meet each other if they start from the two entities.



Same Information – Various Representations

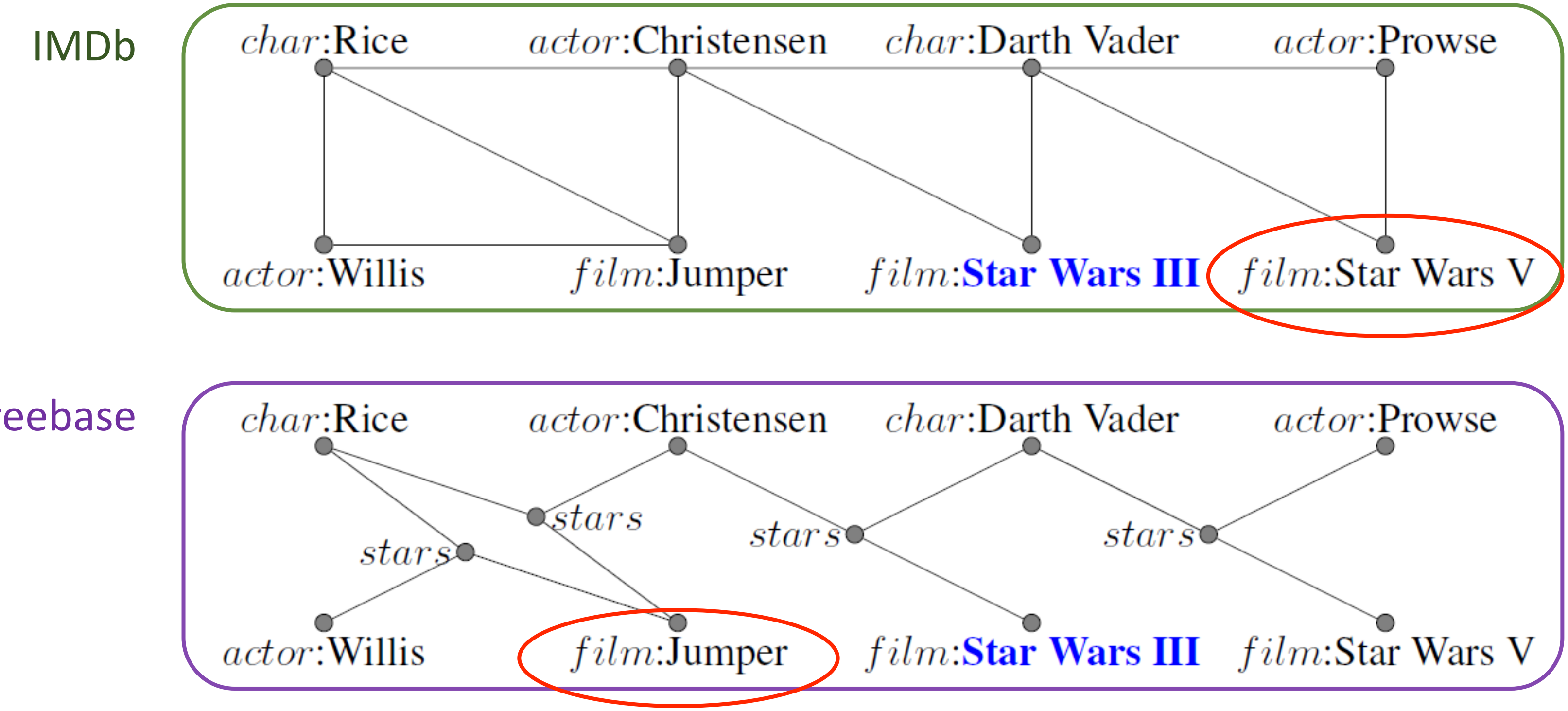


Freebase (www.freebase.com)

Other examples: blank nodes, redundancy, …

Same Information – Various Representations – Different Answers

- Use SimRank to find similar movie to Star Wars III



Algorithms are effective only over databases that follow certain representations.

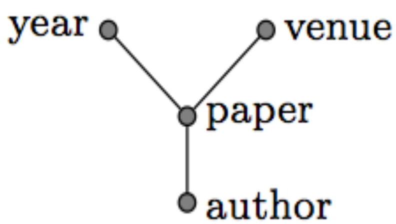
Current solution: Data Conversion & Wrangling

- Manually convert data to the desired representation for the algorithm.

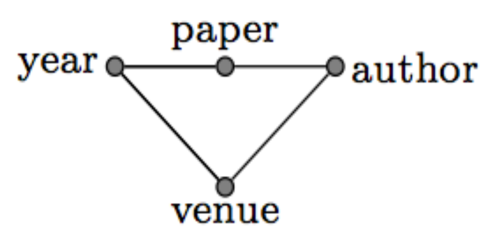
  - Hard and time consuming
  - Algorithms do not provide any definition of *desired* representations. Thus, users have to apply trial and error.

Each researcher uses her own representation

- It is hard to compare different algorithms because they are evaluated over different representations.
  - E.g. research papers use different representations for DBLP data



Y. Sun et al., **PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks**, PVLDB'11



P. Zhao et al., **P-rank: a comprehensive structural similarity measure over information networks**, CIKM'09

Our approach: representation independence

- We do **NOT** want to convert / wrangle the data!

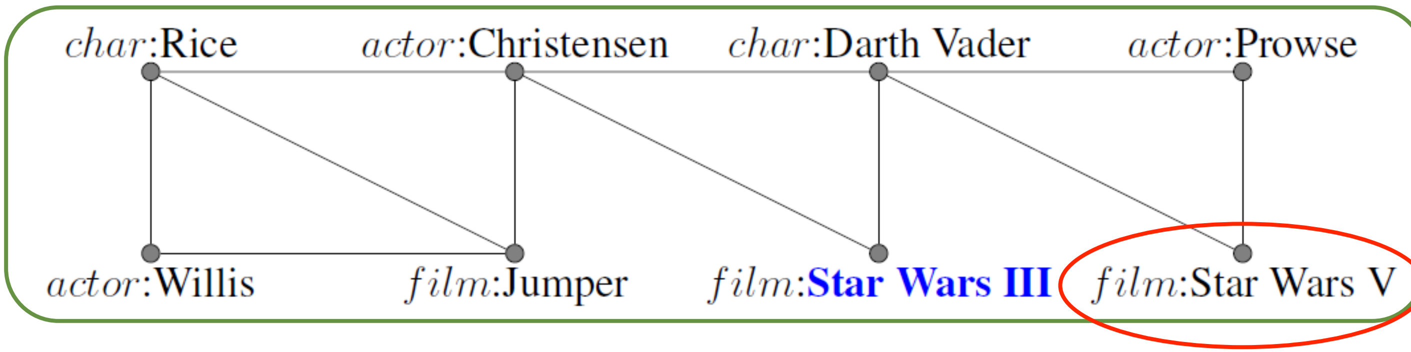
- Develop algorithms that return the same results for the same query over databases with the same information.

Let’s precisely define representation independent algorithm.

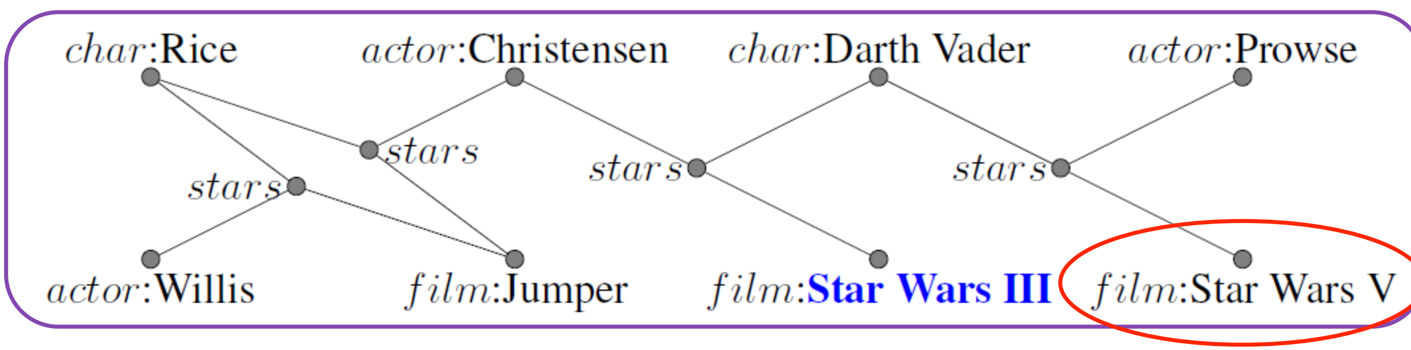
Representation independent algorithm

- An algorithm is ***representation independent*** if it returns the same answers over databases with the same information.

IMDb



Freebase



When do databases represent same information?

Database Transformation

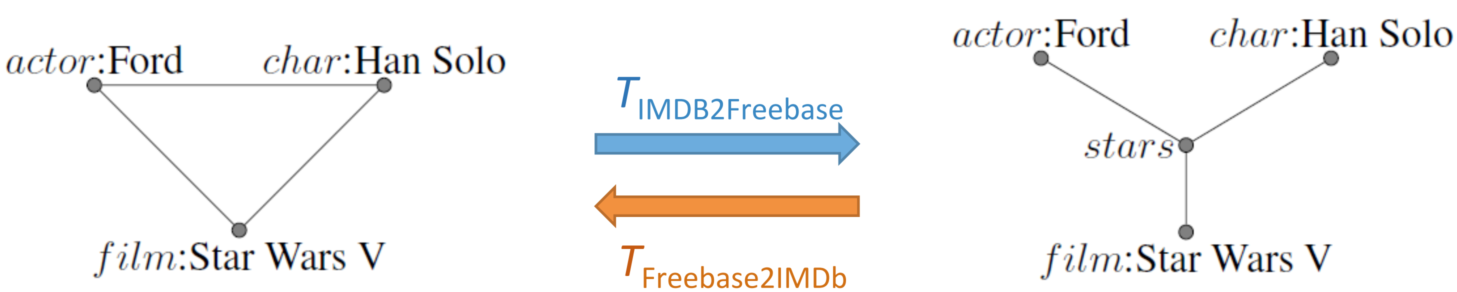
A transformation is a function that maps a database to another one.

IMDb

Freebase

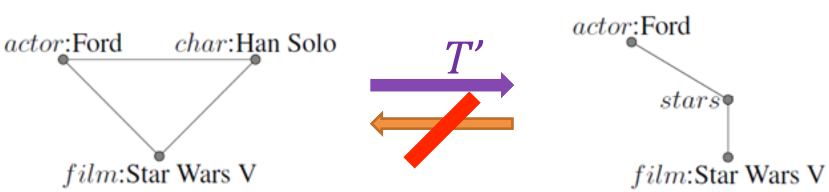
Invertible Transformation

A transformation T is **invertible** if one can reconstruct D from $T(D)$.

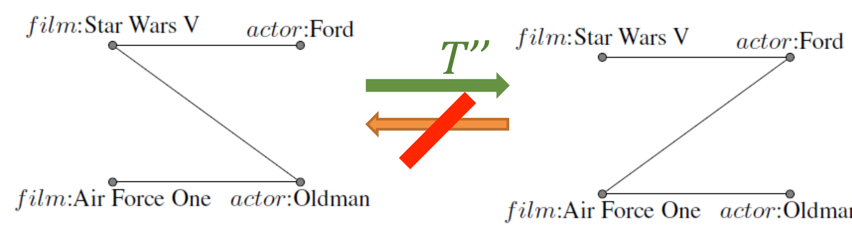


IMDb

Freebase



T' is not invertible. Cannot recover "*char*: Han Solo".



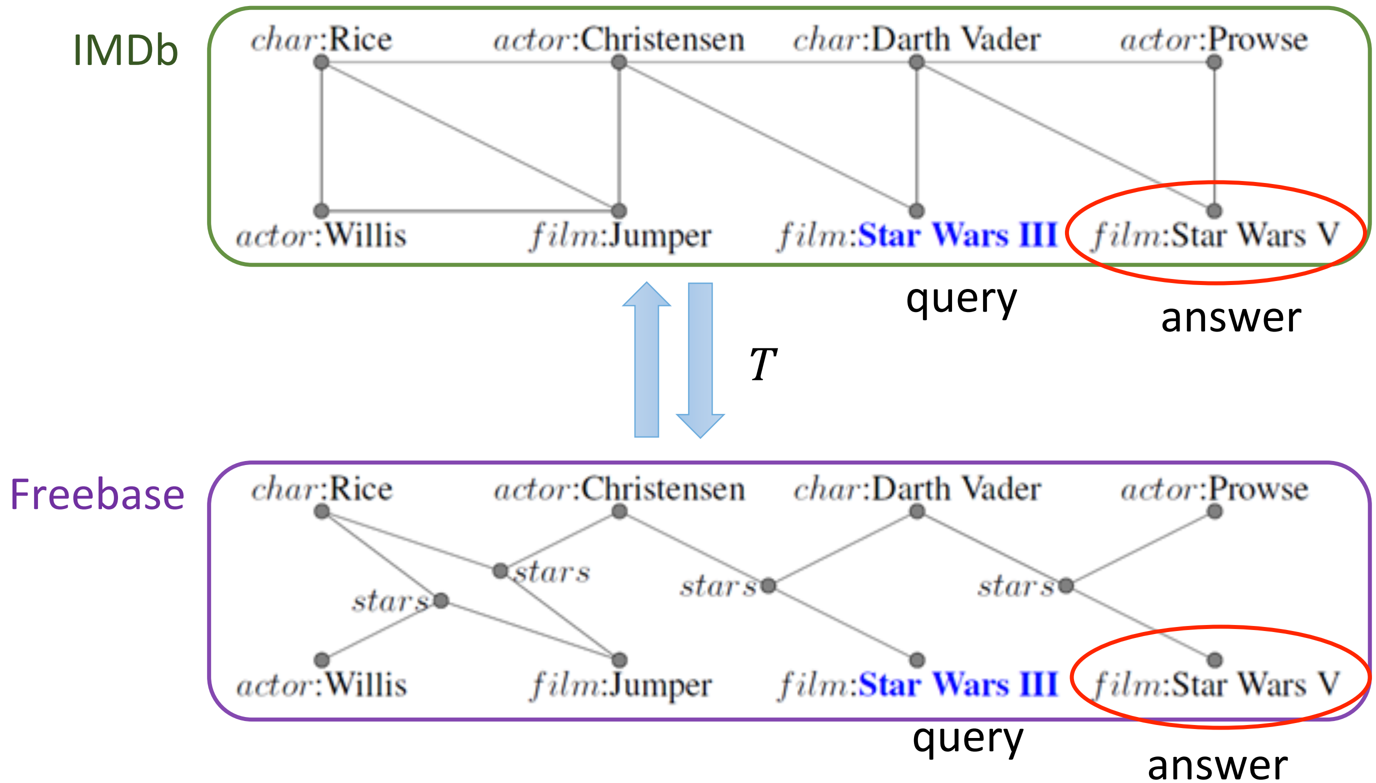
T'' is not invertible. Cannot recover relationship.

Invertible transformation preserves information.

> D_1 and D_2 have the same information
> if there is an invertible transformation between them.

Representation independent algorithm

- Given an invertible transformation T, an algorithm is representation independent under T if it returns the same answers for all queries over a database D and $T(D)$.



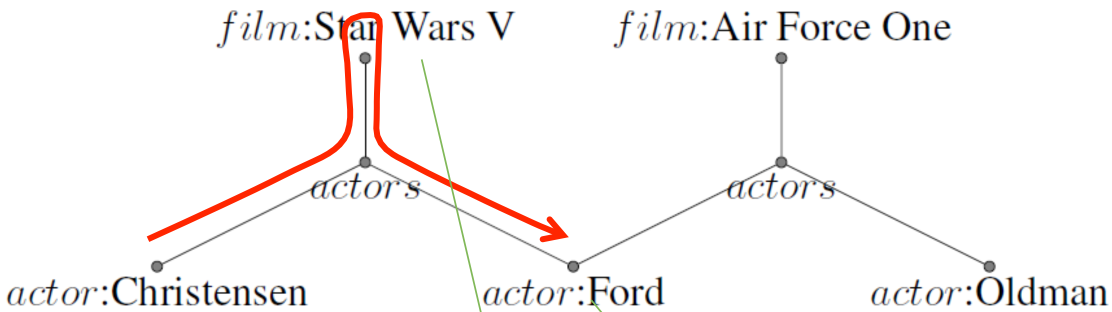
- Larger set of transformations $\Rightarrow$ more representation independent .

Our plan for finding representation independent algorithm

- Representation independent similarity search over two types of transformations.
  - Relationship-reorganizing transformation
  - Entity-rearranging transformation

- Extend current algorithms
  - They are effective over certain representations
  - People have already adapted and used these existing methods

Representing relationships between entities in graphs



- **Walk**: a sequence of consecutive nodes and edges

  it represents a relationship between entities

  [actor: Christensen, actors, film: Star Wars V, actors, actor: Ford]

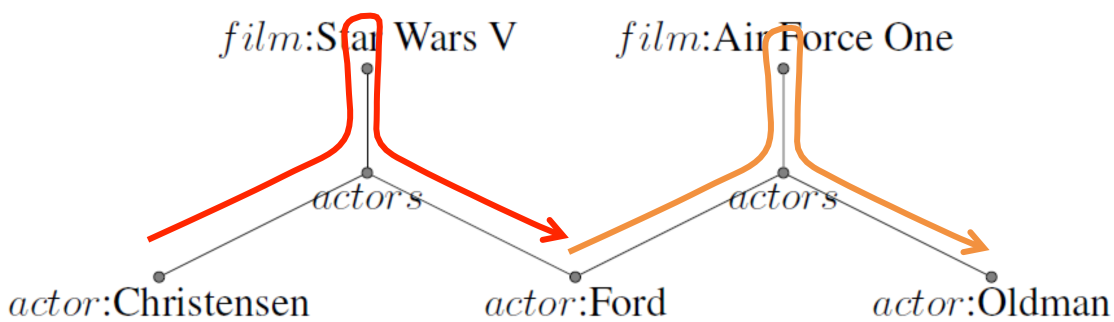
- **Value** of a walk: tuple of nodes with values in the walk

  [actor: Christensen, film: Star Wars V, actor: Ford]

Representing types of relationships in graphs

- **Meta-walk** : a sequence of labels of nodes in walks

  Meta-walk represents type of relationships between entities



[actor: Christensen, actors, film: Star Wars V, actors, actor: Ford]

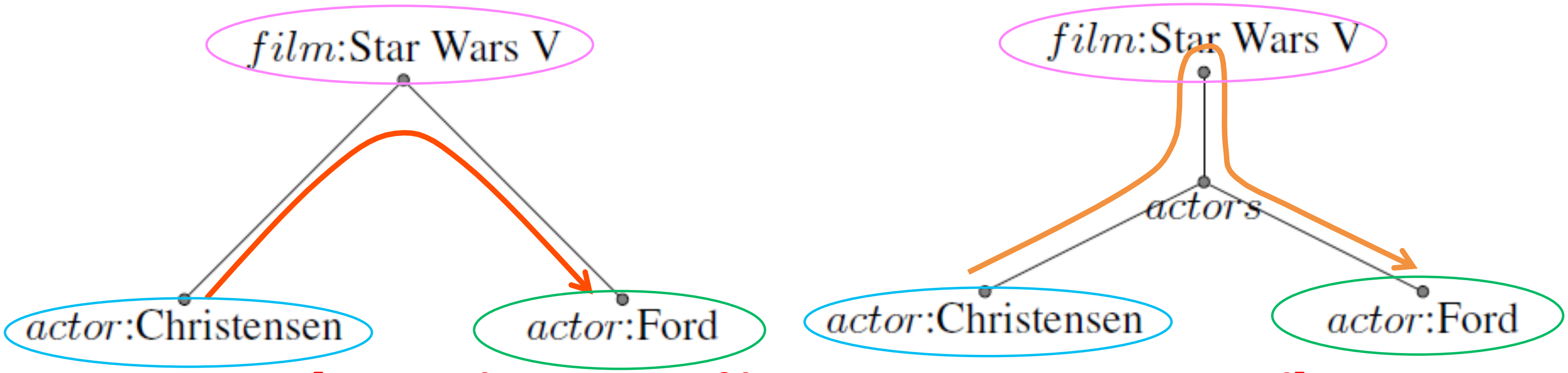
[actor: Ford, actors, film: Air Force One, actors, actor: Oldman]

are walks of a meta-walk

[actor, actors, film, actors, actor]

Equivalent relationships

- **Content-equivalent**
  - Two walks are content-equivalent if their values are equal.



[actor: Christensen, film: Star Wars V, actor: Ford]

is content equivalent to

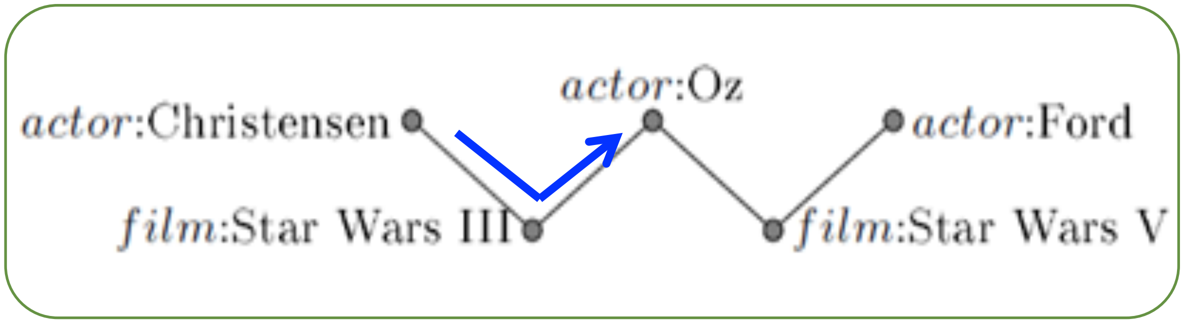
[actor: Christensen, actors, film: Star Wars V, actors, actor: Ford]

  - Content-equivalent walks represents same relationship between set of entities

- Notion of content equivalent extends naturally for meta-walks
- Two content equivalent meta-walks represent same type of relationship.

Relationship-constrained similarity search methods

- Measure similarity between entities over a given type of relationship, i.e., meta-walk.



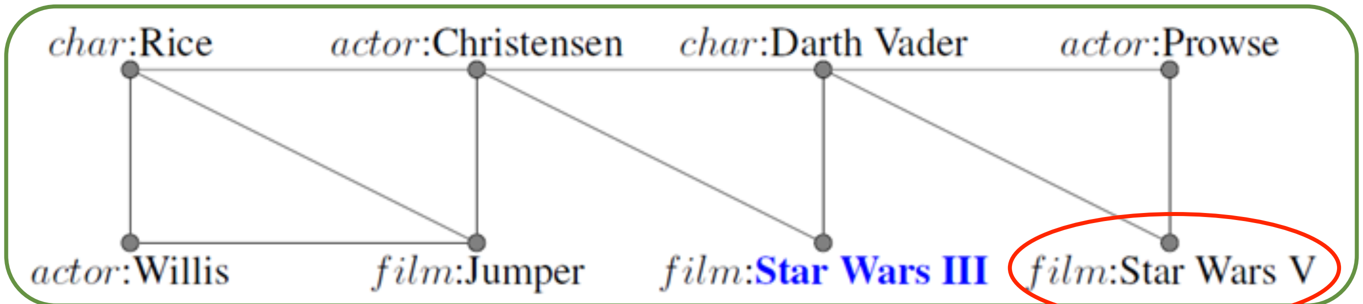
IMDb

- E.g. find similar actors based on their common movies
  - Meta-walk: [actor, film, actor]
- Different ways of computing similarity within a meta-walk
  - Random walk, enumerating # walks.
- Current methods use paths (meta-paths) to represent relationships.
  - We use walks (meta-walks) for reason which we will later explain.

Relationship-reorganizing transformation

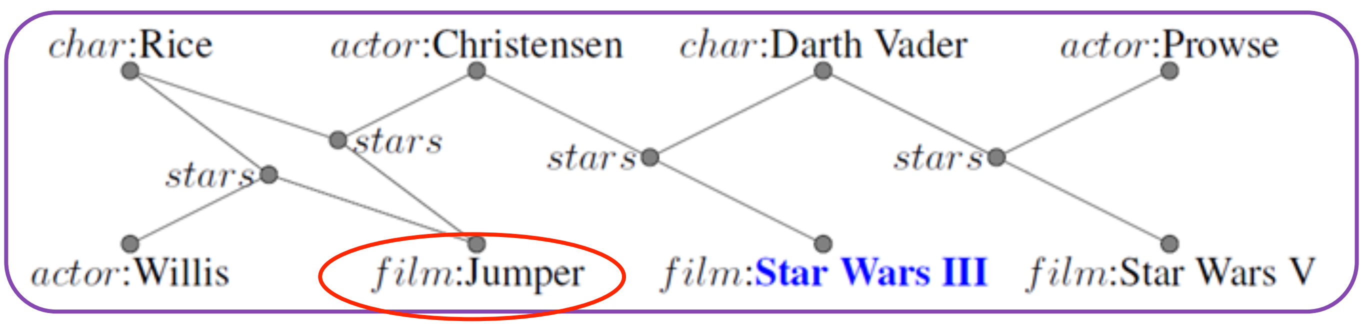
Databases contain the same set of entities and relationships, but relationship are represented in different forms.

IMDb



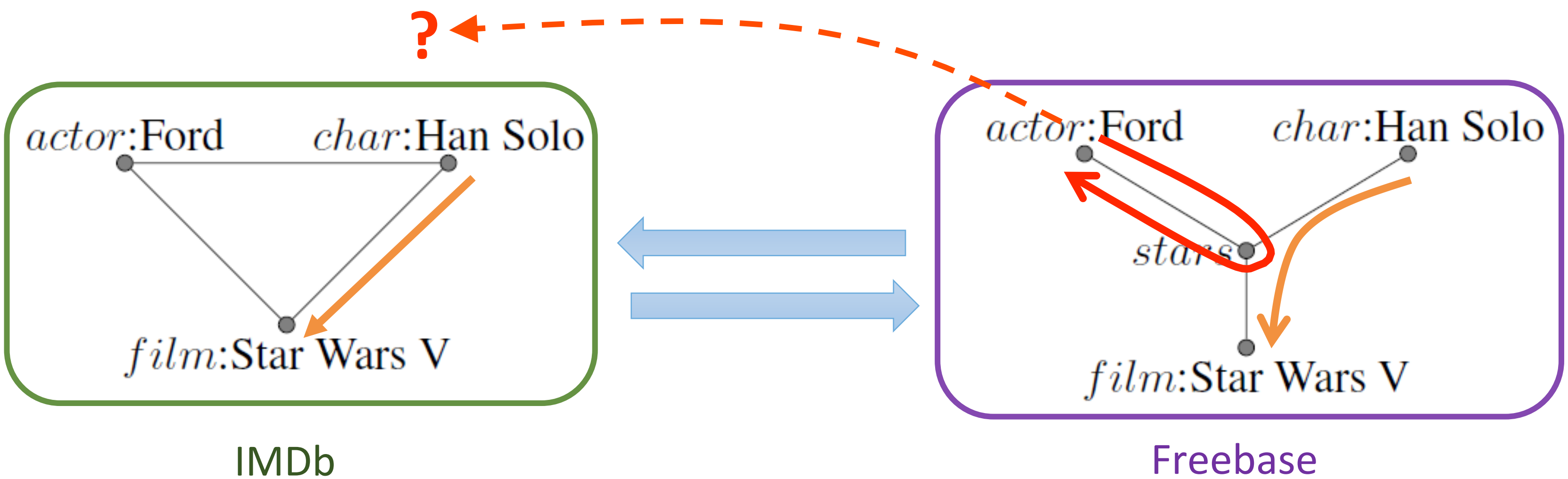
Current similarity algorithms are not representation independent under this type transformation.

Freebase



Why current algorithms fail?

- Relationship reorganization introduces/removes walks



- A walk *with* consecutive forward and backward traverses from an entity to a node without value is called **non-informative** walk.

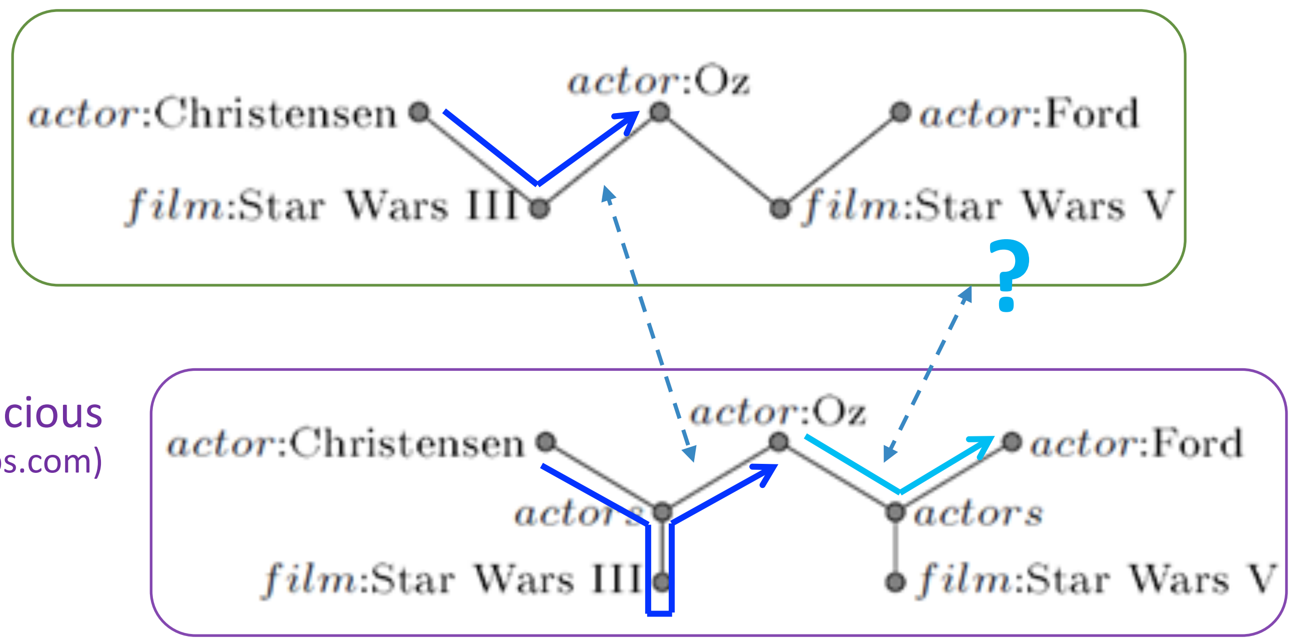
Solution: Ignore non-informative walks

Why current algorithms fail?

- Relationship reorganization introduces/removes meta-walks

IMDb

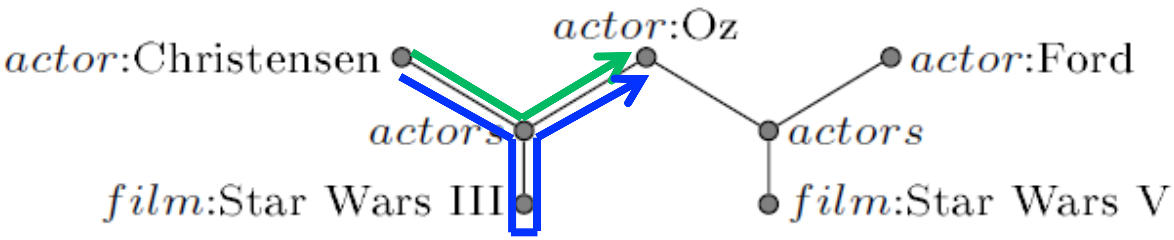
Movielicious
(www.netwalkapps.com)



There is no content equivalent meta-walk to [actor, actors, actor] in IMDb.

Solution: **use inclusion between meta-walks**

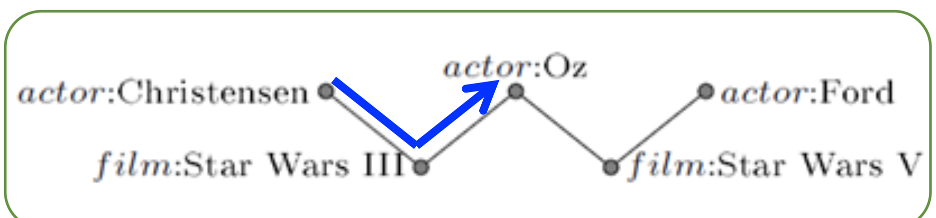
Movielicious



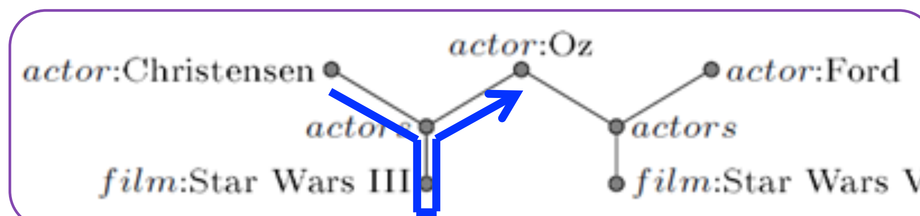
Observation: every walk of [actor, actors, actor] is included in exactly one walk of [actor, actors, film, actors, actor].

A meta-walk is **maximal** if it is not included in any other meta-walk.

There is a bijection between maximal meta-walks in a database and its relationship-reorganizing transformation such that these meta-walks are content-equivalent.



IMDb



Movielicious

Robust-PathSim (R-PathSim)

Extends PathSim algorithm so that it recognizes and uses only informative walks of maximal meta-walks to computing similarity score between entities.

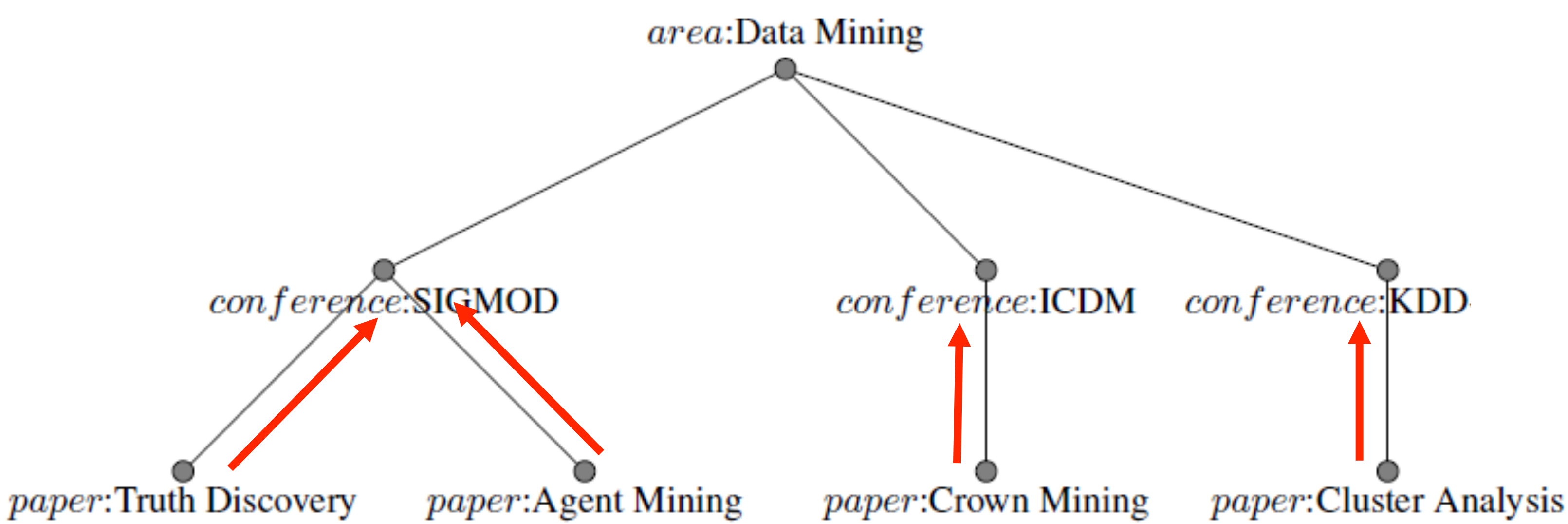
Theorem

R-PathSim is representation independent under relationship-reorganizing transformation.

Entity-Rearranging Transformation

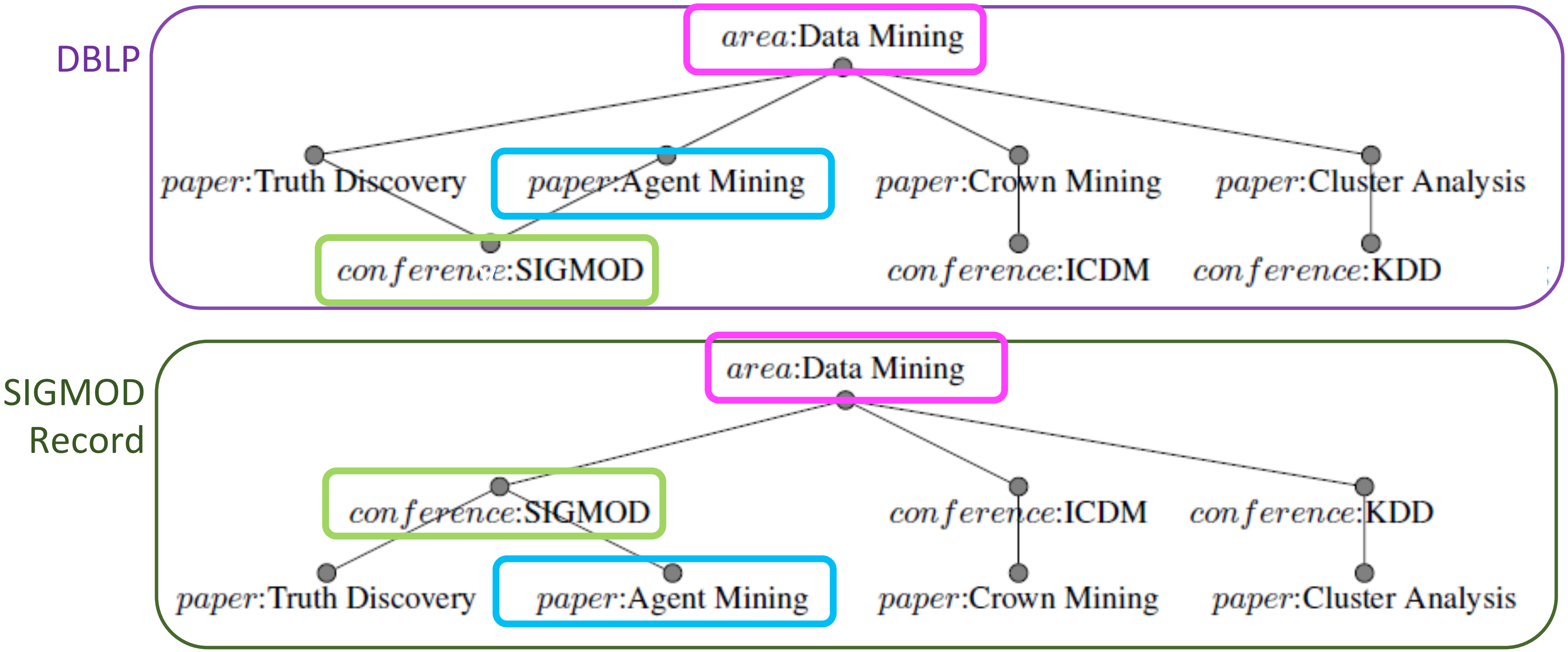
There is a functional dependency from entity type a to entity type b ($a \rightarrow b$) if every entity of a is connected to only one entity of b.

Functional dependencies: paper → conference



Entity-Rearranging Transformation

Given some functional dependencies, entity-rearranging transformation connects set of entities in different orders.

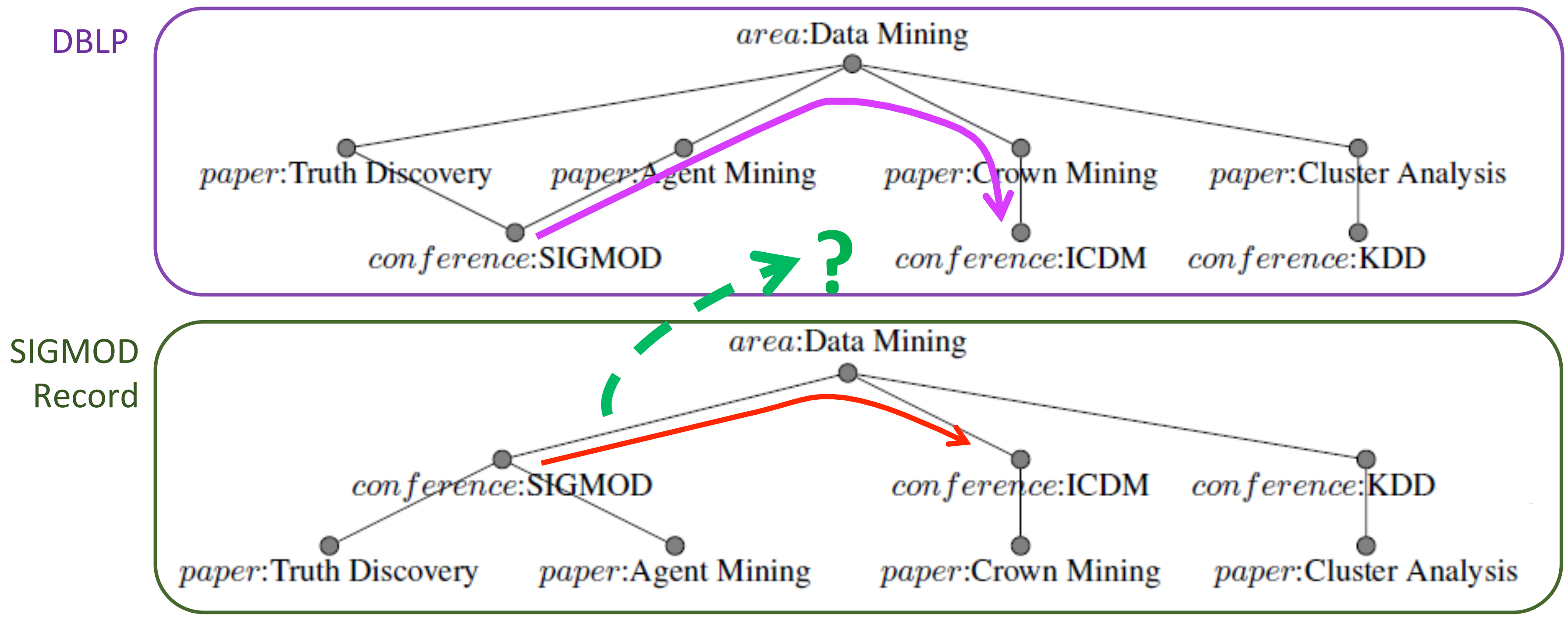


paper → conference, conference → area

Current similarity algorithms are not representation independent under this type transformation.

Why current algorithms fail?

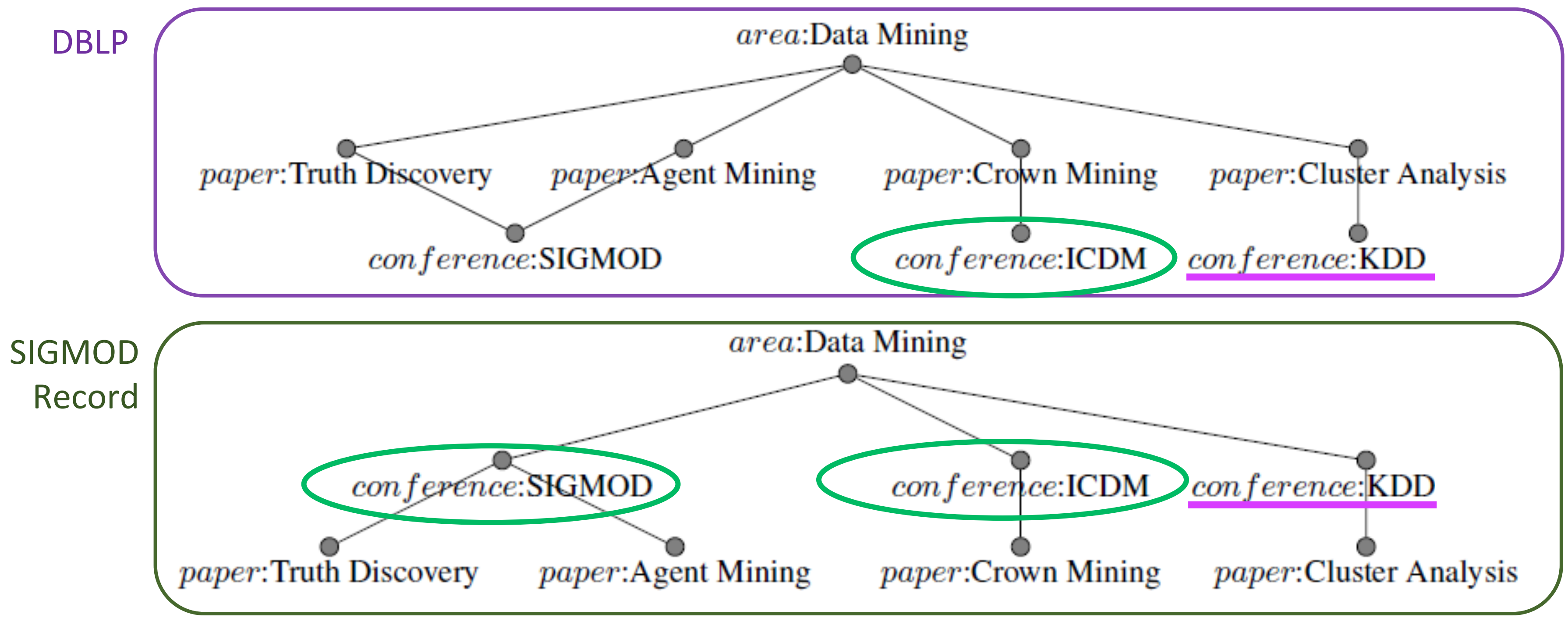
- Type of relationships in the transformed database may not remain in form of meta-walks.



- Which meta-walk in DBLP represents the same relationship as [conference, area, conference] in SIGMOD Record?
- Potential candidate is [conference, paper, area, paper, conference]

Why current algorithms fail?

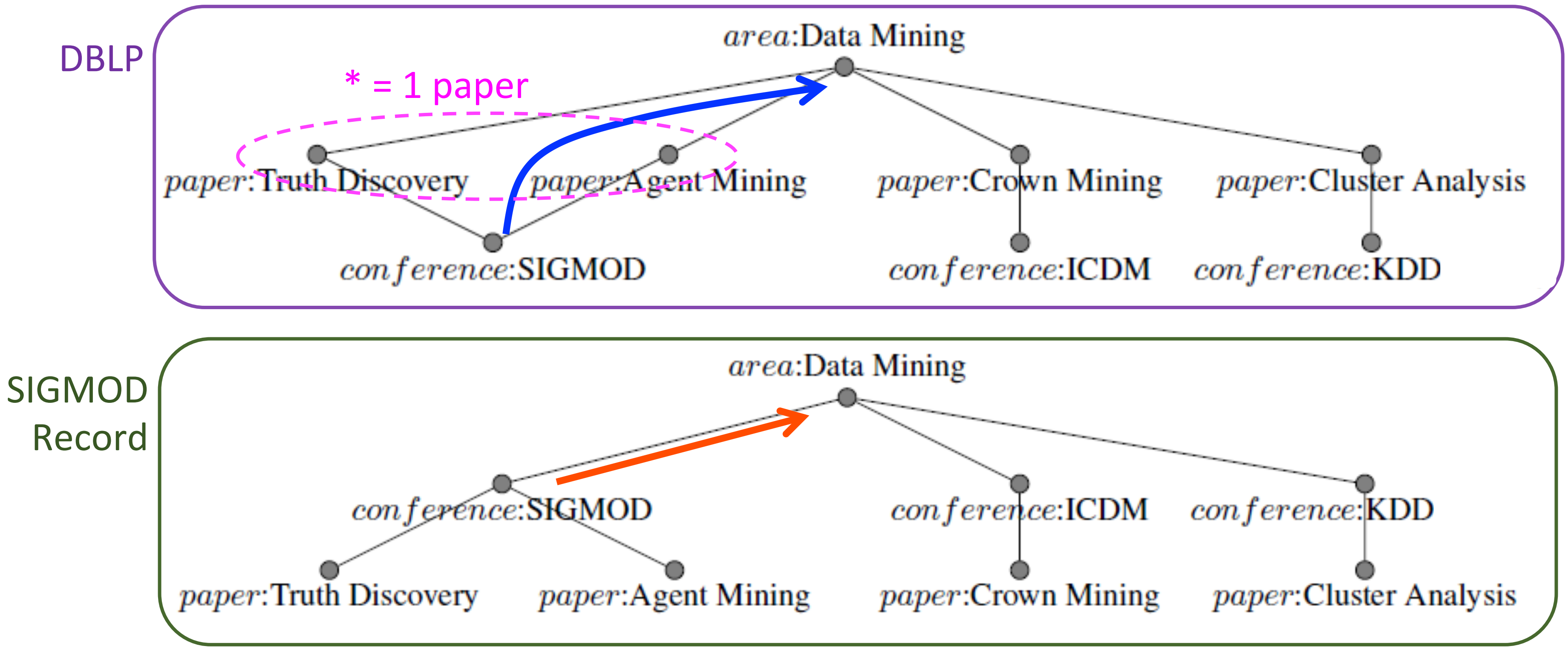
- But, [conference, area, conference] in SIGMOD Record and [conference, paper, area, paper, conference] in DBLP does not have the same meaning



- Find similar conference to KDD using PathSim.
- Number of papers in conferences influences the ranking

Solution: consider other representation of relationship beyond meta-walk

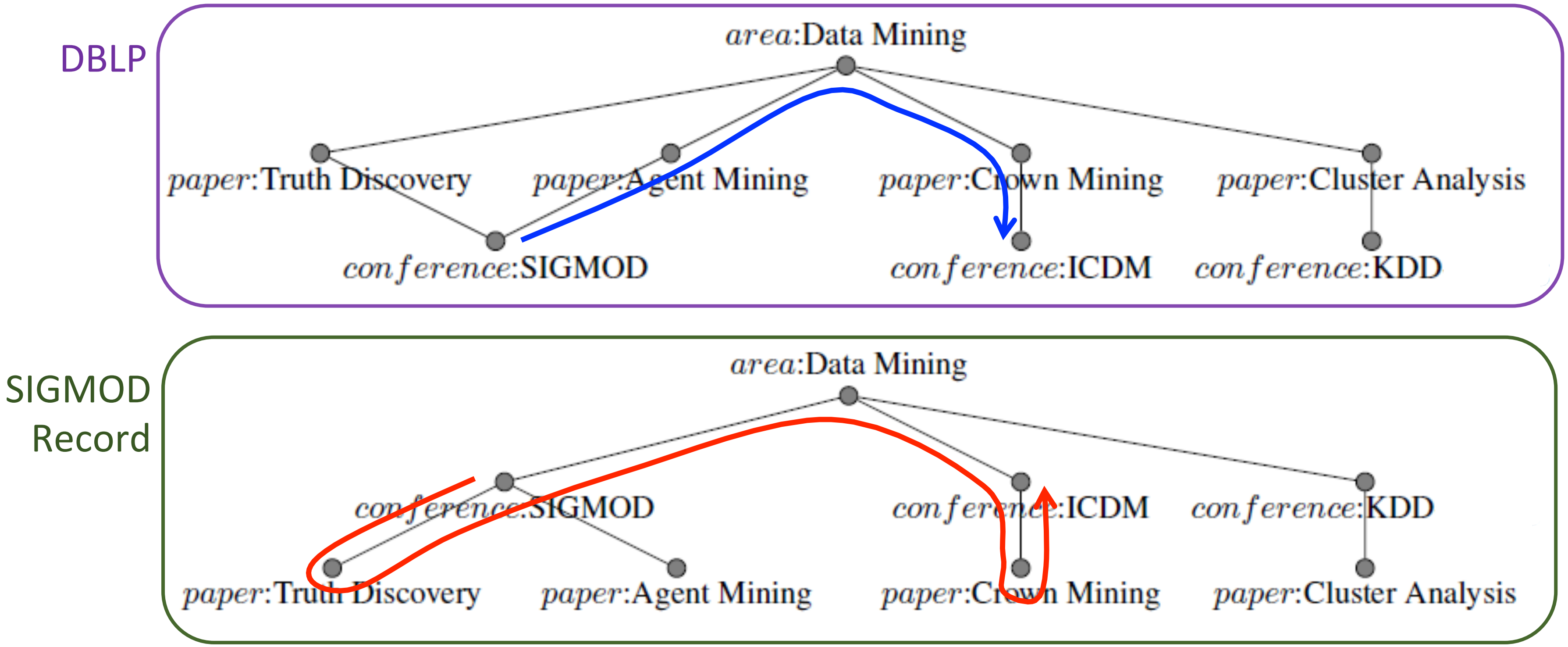
Meta-walk with *-label



[conference, area, conference] = [conference, *, area, *, conference]

Use meta-walk instead of meta-path to represent relationships

Which meta-walk in SIGMOD Rec. should be mapped to [conference, paper, area, paper, conference] in DBLP?



[conference, paper, area, paper, conference]

= [conference, ***paper, conference***, area, ***conference, paper***, conference]

This is why we use meta-walks instead of meta-paths.

Too many types of meta-walks.

- People who are not familiar with the database may not be able to express their desired meta-walk.
- *Solution:* Compute (weighted) average of similarity scores over all maximal meta-walks between entities.

- However, the set of all maximal meta-walks can be very large.
  - It may take a long time to compute score for all of them.
- *Solution:* pruning techniques to find a small subset of meta-walks to compute the similarity score efficiently.

<u>Theorem</u> R-PathSim is representation independent under relationship-reorganizing transformation and entity rearranging transformation.

Empirical results: Average Ranking Differences

Use Kendall's tau to measure ranking difference. (0 = no difference, 1 = reverse ranking)

No ranking difference for R-PathSim.

Movies DB Representations
IMDb, **MVL**: Movielicious, **ASM**: Assignment from evc-cit.info/cit0441x

Bibliographic DB Representations
DBLP, **SNAP**: Stanford Network Analysis Project

		Relationship reorganizing			
		IMDb2MVL	IMDb2ASM	IMDb2Freebase	DBLP2SNAP
Top 3	RWR	0.473	0.505	0.170	0.141
	SimRank	0.411	0.458	0.333	0.634
	PathSim	0	0	0	0.564
Top 5	RWR	0.444	0.459	0.158	0.134
	SimRank	0.365	0.392	0.337	0.578
	PathSim	0	0	0	0.522
Top 10	RWR	0.404	0.415	0.155	0.126
	SimRank	0.343	0.348	0.322	0.493
	PathSim	0	0	0	0.495

Empirical results: Average Ranking Differences

No ranking difference for R-PathSim.

DB about courses
WSU: WSU Course Dataset, **Alchemy**: Alchemy UW-CSE database

		Entity rearranging	
		DBLP to SIGMOD Record	WSU to Alchemy
Top 3	RWR	0.482	0.300
	SimRank	0.481	0.440
	PathSim	0.641	0.320
Top 5	RWR	0.447	0.259
	SimRank	0.455	0.387
	PathSim	0.608	0.310
Top 10	RWR	0.412	0.253
	SimRank	0.410	0.341
	PathSim	0.590	0.247

(0 = no difference, 1 = opposite ranking)

Effectiveness of R-PathSim

- Use the Microsoft Academic Search dataset.
- Randomly sample 50 conferences based on degrees in the dataset.
- For ground truth, given a conference, we manually group all other conferences in 3 categories: similar, quite-similar, least-similar.
- We measures the statistical significance of our results using the paired-t-test at a significant level of 0.05

	nDCG @ 5	nDCG @ 10
PathSim	0.625	0.564
R-PathSim	0.658	**0.630**

Efficiency of R-PathSim

- Datasets
  - Movielicious: 2.4M nodes, 7.5M edges
  - DBLP: 1.2M nodes, 2.7M edges
  - DBLP+: 1.9M nodes, 3.3M edges
- Hardware configuration: Linux server with 64GB RAM, 2 quad core CPU.
- Average query processing time per meta-walk in second

	Size of meta-walk	Movielicious	DBLP	DBLP+
PathSim	5	0.036	0.030	0.046
	7	0.068	0.347	0.227
R-PathSim	5	0.036	0.035	0.046
	7	0.068	0.343	0.233

- Average query processing time for aggregated R-PathSim

	Size of meta-walk	Movielicious	DBLP	DBLP+
PathSim	5	0.036	0.091	0.092
	7	0.136	1.041	0.681
R-PathSim	5	0.036	0.140	0.184
	7	0.136	1.714	1.165

Conclusion & future work

- Graph exploration algorithms are representation dependent and therefore hard-to-use.
  - scale algorithms to work on various representations.
  - scale for the second **V** in Big Data: **V**ariety.
- We've developed representation independent algorithms for some frequent representational shifts.
- To do:
  - benchmark for varieties of representations.
- More information:
  - **RIDE: R**epresentation **I**ndependent **D**ata **E**xploration
    http://eecs.oregonstate.edu/~termehca
  - VLDB'15 and VLDB'16 demos.