Toward Representation Independent Similarity Search Over Graphs

Yodsawalai Chodpathumwan, University of Illinois at Urbana-Champaign

Arash Termehchy, Oregon State University

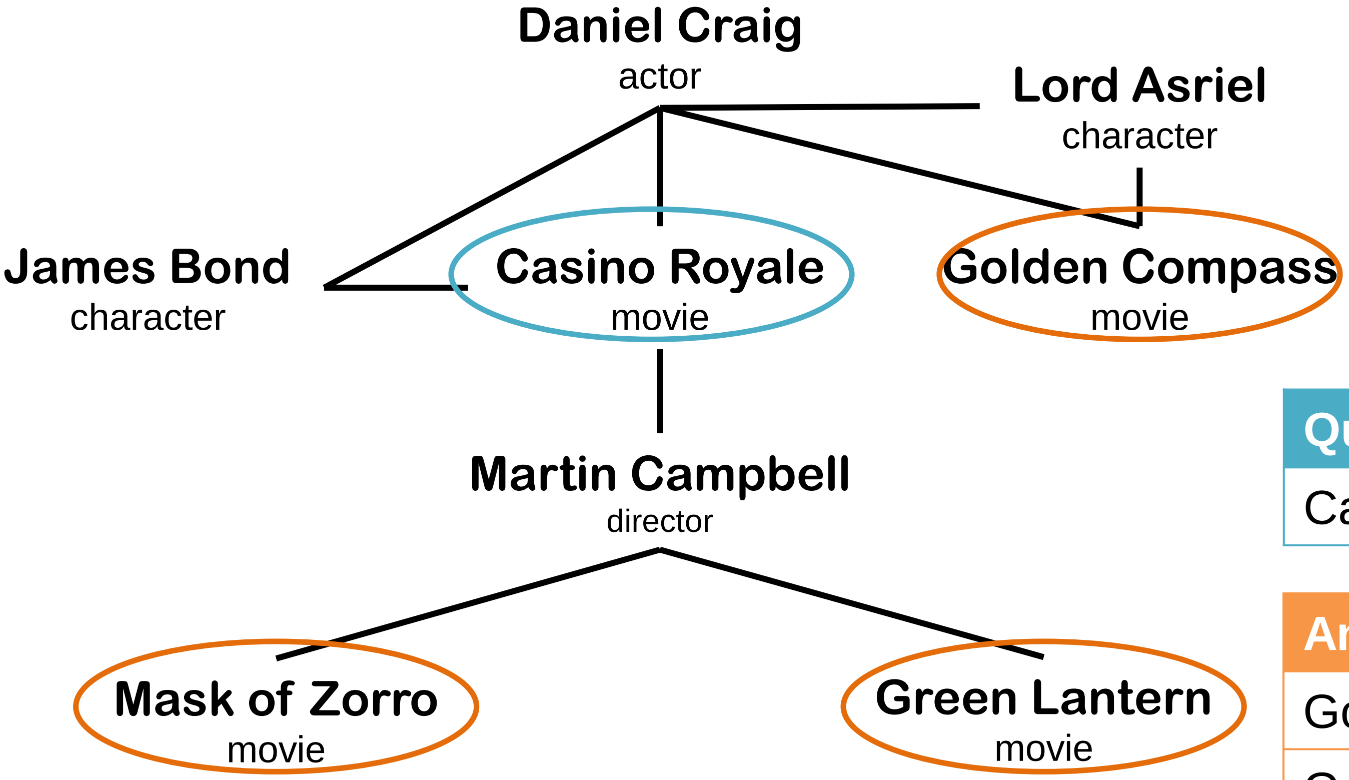
Yizhou Sun, Northeastern University

Amirhossein Aleyasin, University of Illinois at Urbana-Champaign

Jose Picado, Oregon State University

Similarity Search over Graph Databases

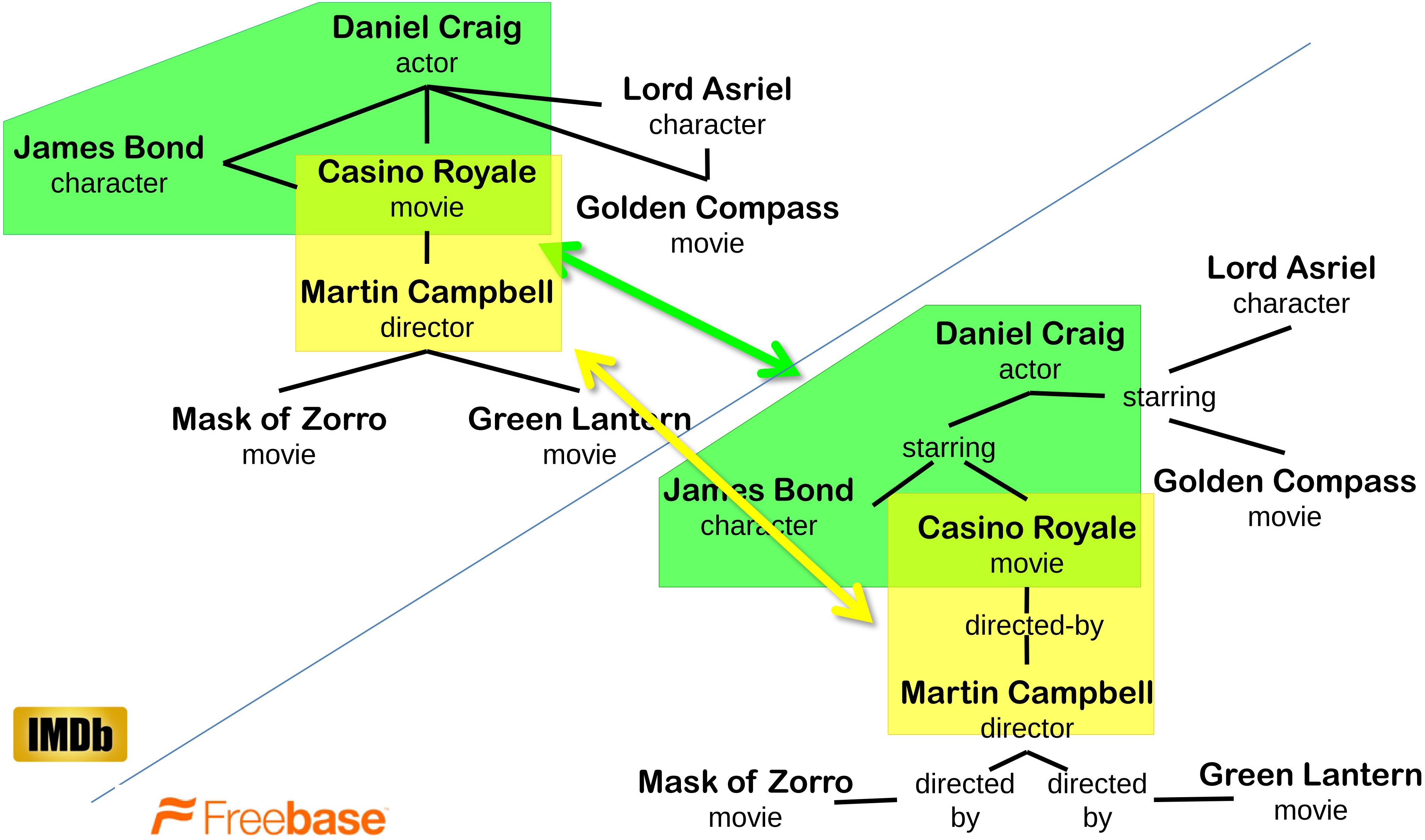
Which movies are similar to “Casino Royale” in IMDb?



Query
Casino Royale

Answers
Golden Compass
Green Lantern
Mask of Zorro

Same information is represented in many ways



Daniel Craig
actor

Lord Asriel
character

James Bond
character

Casino Royale
movie

Golden Compass
movie

Martin Campbell
director

Mask of Zorro
movie

Green Lantern
movie

Lord Asriel
character

Daniel Craig
actor

starring

starring

Golden Compass
movie

James Bond
character

Casino Royale
movie

directed-by

Martin Campbell
director

Mask of Zorro
movie

directed
by

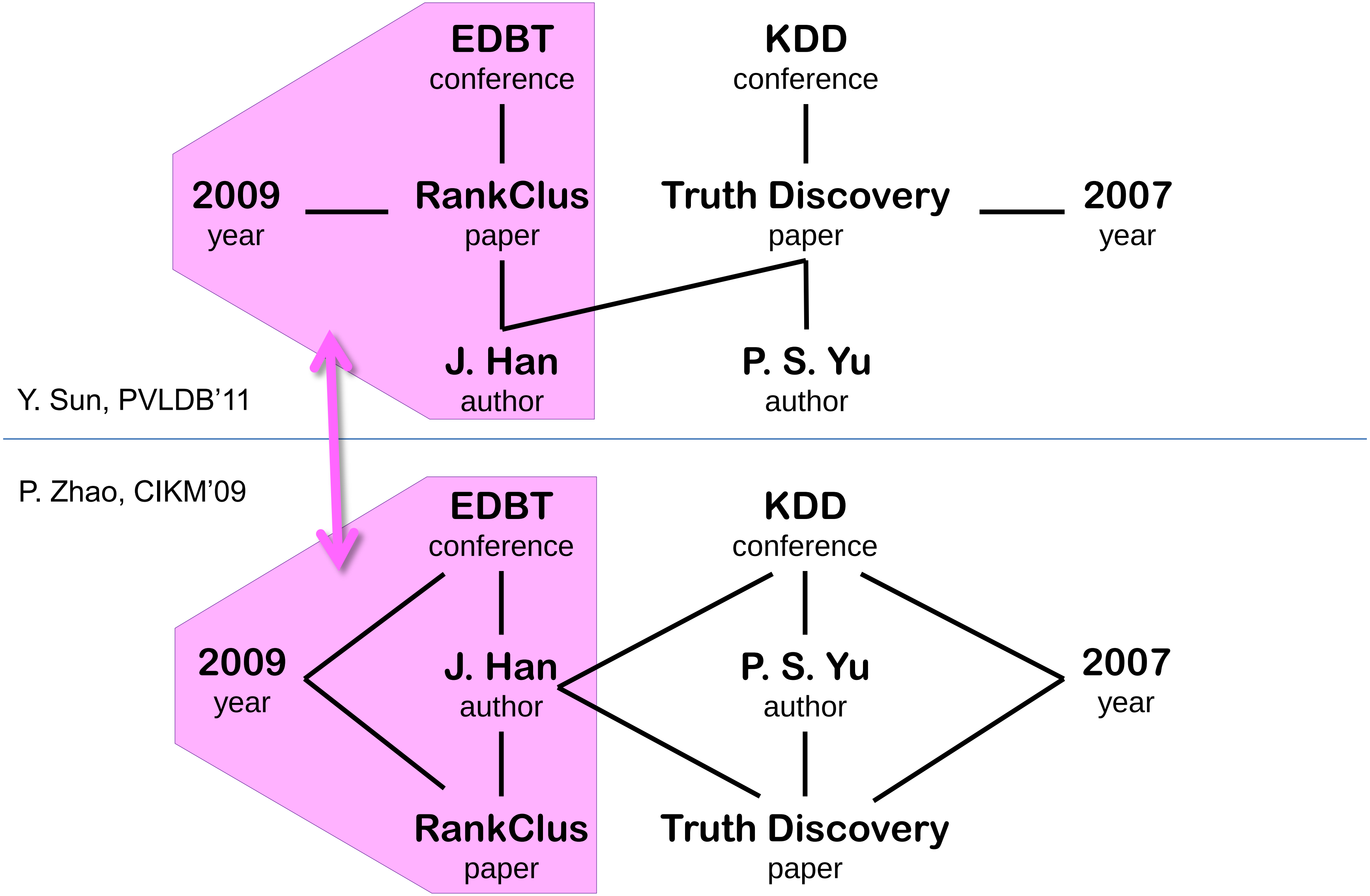
directed
by

Green Lantern
movie

IMDb

Freebase

Same information is represented in many ways



EDBT
conference

KDD
conference

2009
year

RankClus
paper

Truth Discovery
paper

2007
year

J. Han
author

P. S. Yu
author

Y. Sun, PVLDB'11

P. Zhao, CIKM'09

EDBT
conference

KDD
conference

2009
year

J. Han
author

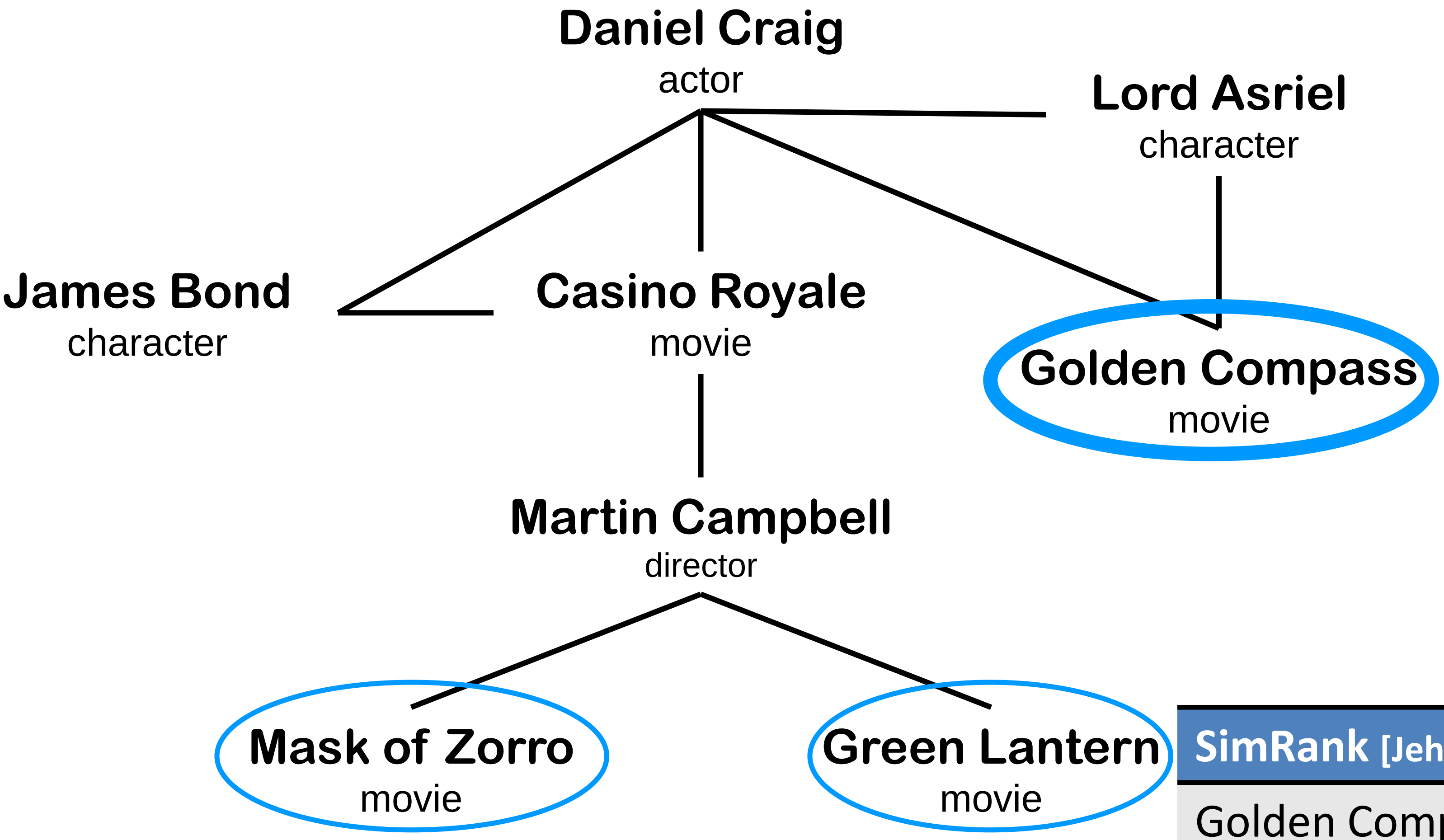
P. S. Yu
author

2007
year

RankClus
paper

Truth Discovery
paper

Which movies are similar to "*Casino Royale*" ?



Daniel Craig
actor

Lord Asriel
character

James Bond
character

Casino Royale
movie

Golden Compass
movie

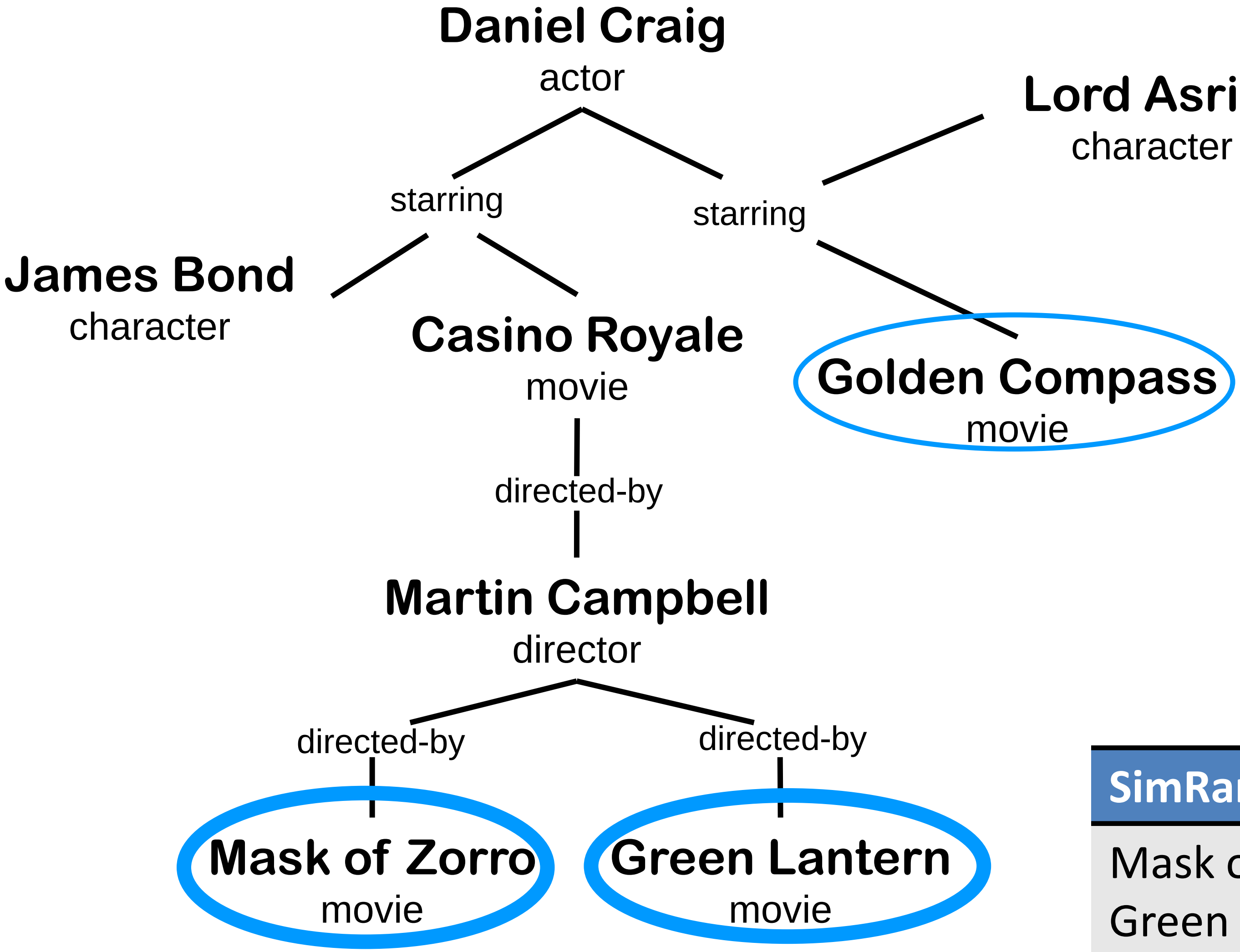
Martin Campbell
director

Mask of Zorro
movie

Green Lantern
movie

SimRank [Jeh, KDD'02]
Golden Compass
Mask of Zorro, Green Lantern

Which movies are similar to “*Casino Royale*” ?



SimRank
Mask of Zorro, Green Lantern
Golden Compass

Freebase

Generality of Similarity Search over Graphs

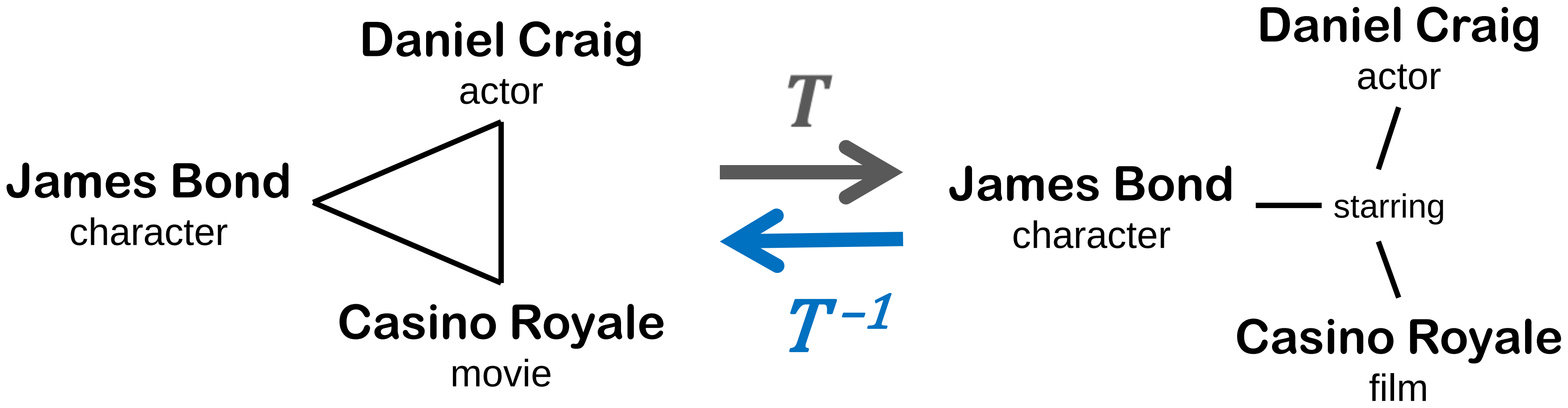
- A similarity search algorithm A is ***general***

iff A returns the same ranked lists of answers over equivalent databases, for all queries.

Invertible Transformation

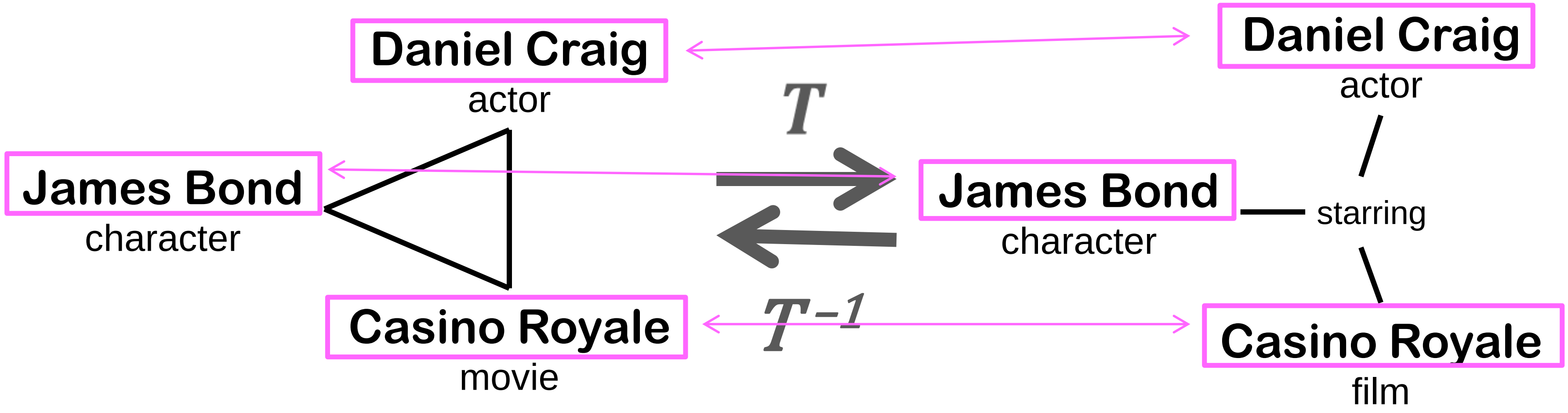
Given a graph database D and a transformation T over D.

- **Invertible** transformation
  - There exists T^{-1} such that $T^{-1}(T(D)) = D$



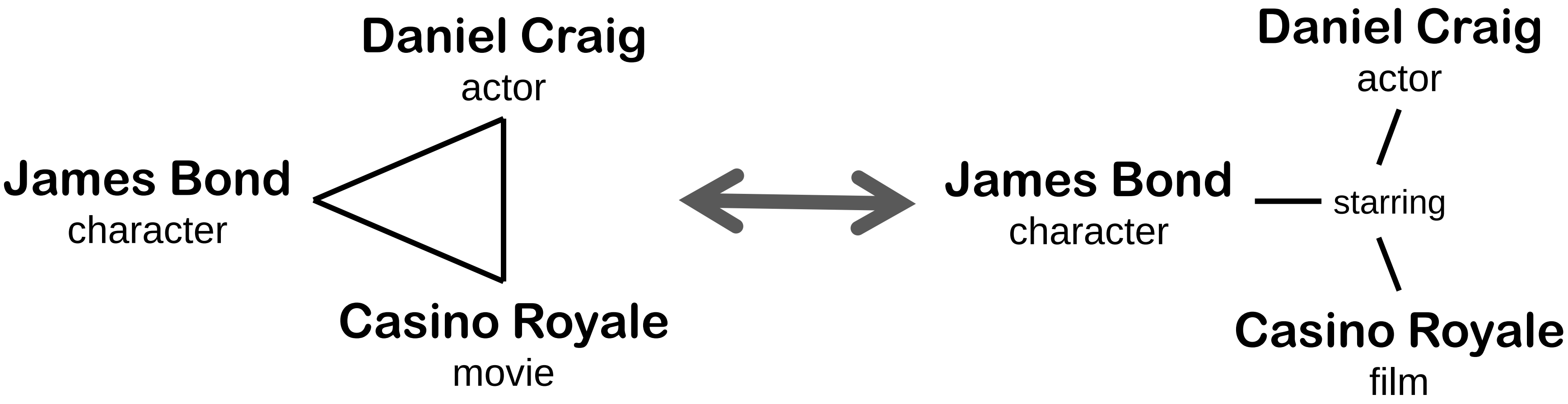
Information Preserving Transformation

- T is **Information preserving** transformation iff
  - T is invertible, and
  - T preserves value of each node

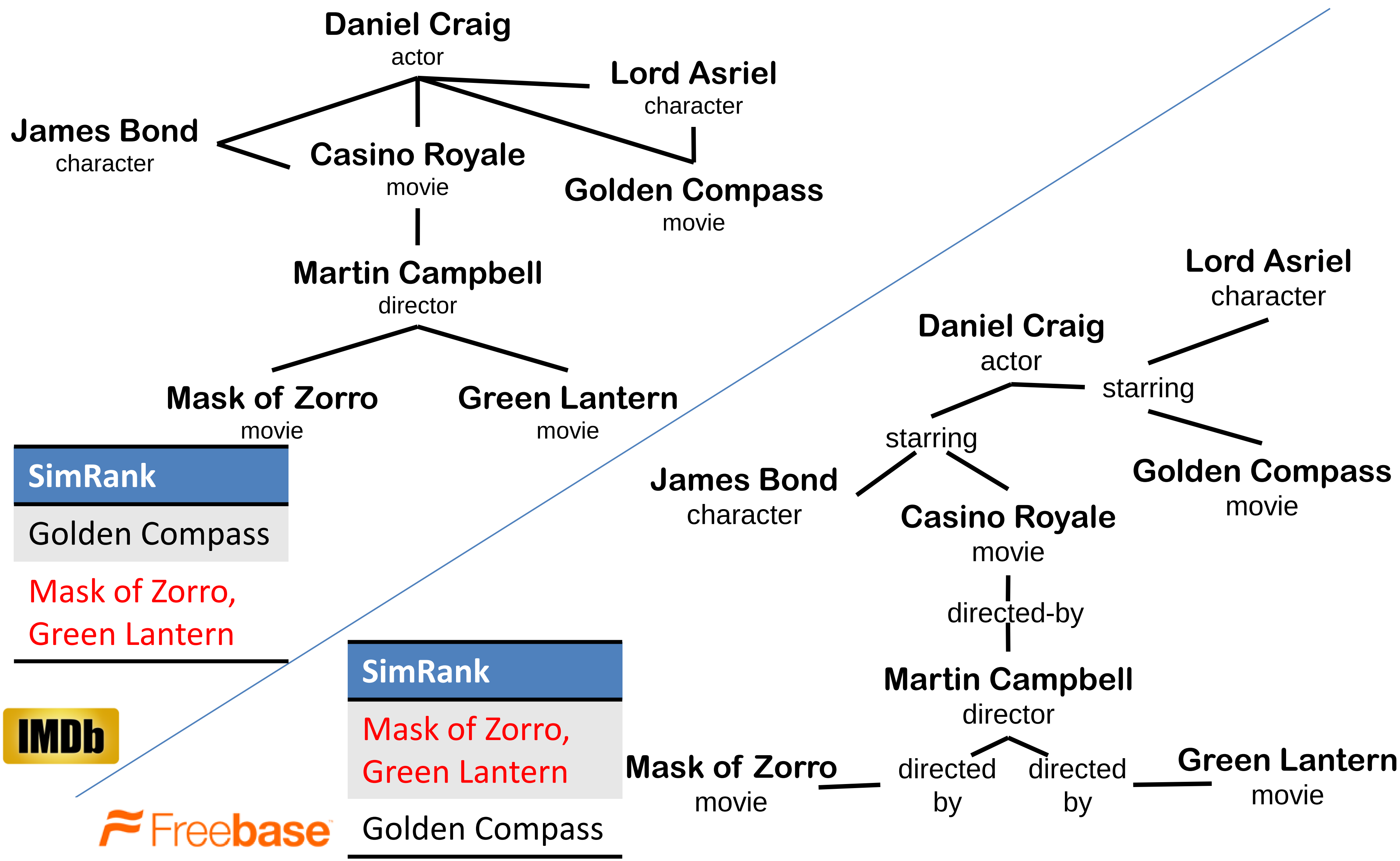


Information Equivalent Graphs

Databases D_1 and D_2 are **information equivalent** iff there exists an information preserving transformation from D_1 to D_2 and vice versa.



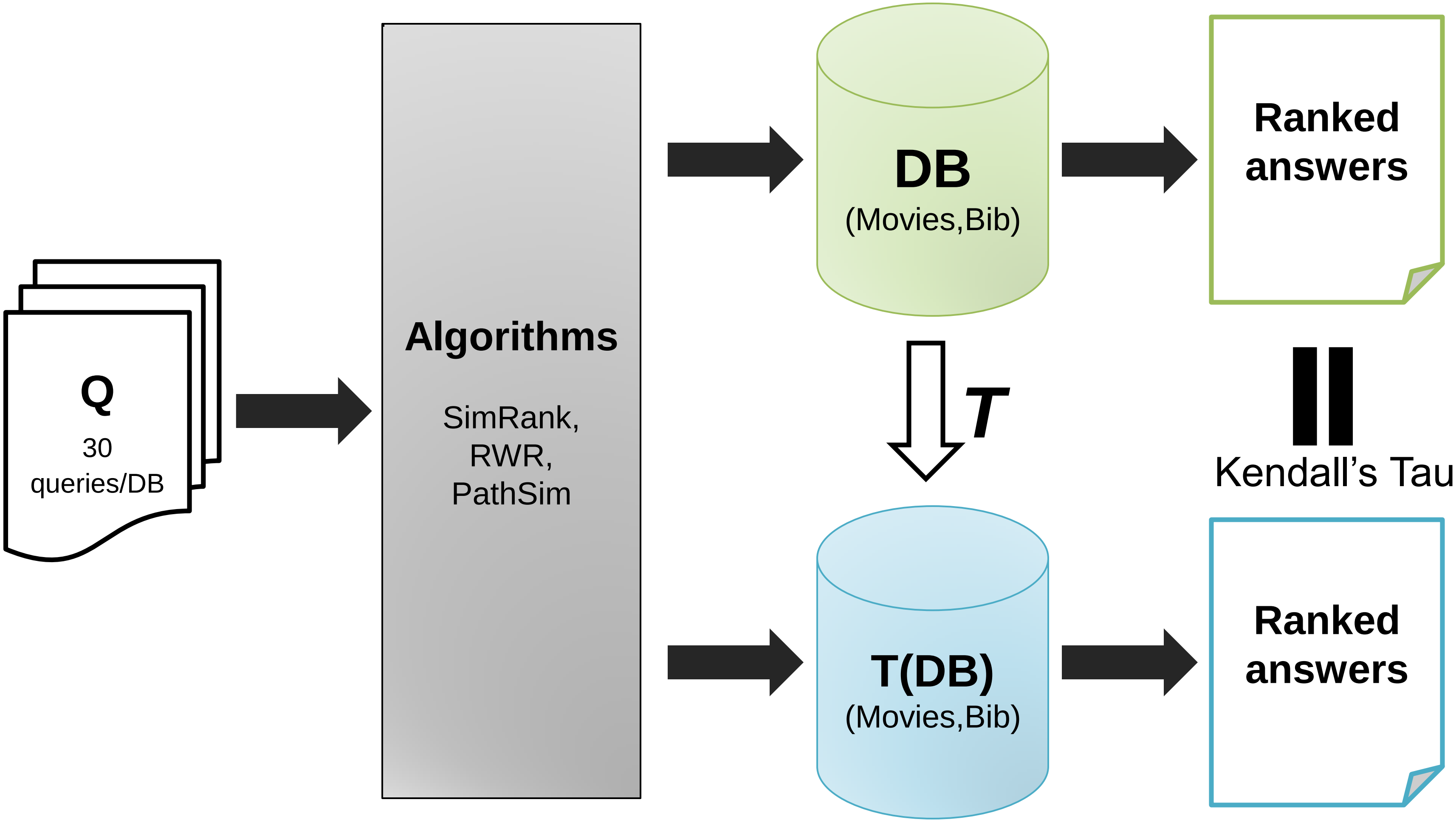
SimRank is not general



SimRank
Golden Compass
Mask of Zorro, Green Lantern

SimRank
Mask of Zorro, Green Lantern
Golden Compass

Evaluation on Generality of Similarity Search Algorithms



Q

30 queries/DB

Algorithms

SimRank, RWR, PathSim

DB (Movies,Bib)

T

T(DB) (Movies,Bib)

Ranked answers

=

Kendall's Tau

Ranked answers

Average Ranking Differences

- Top 10 answers

Algorithms	Movie DBs	Bib DBs
RWR [Tong, KDD'06]	0.375	0.773
SimRank [Jeh, KDD'02]	0.418	0.626
PathSim [Sun, VLDB'11]	0.375	0.953

- Top 50 answers

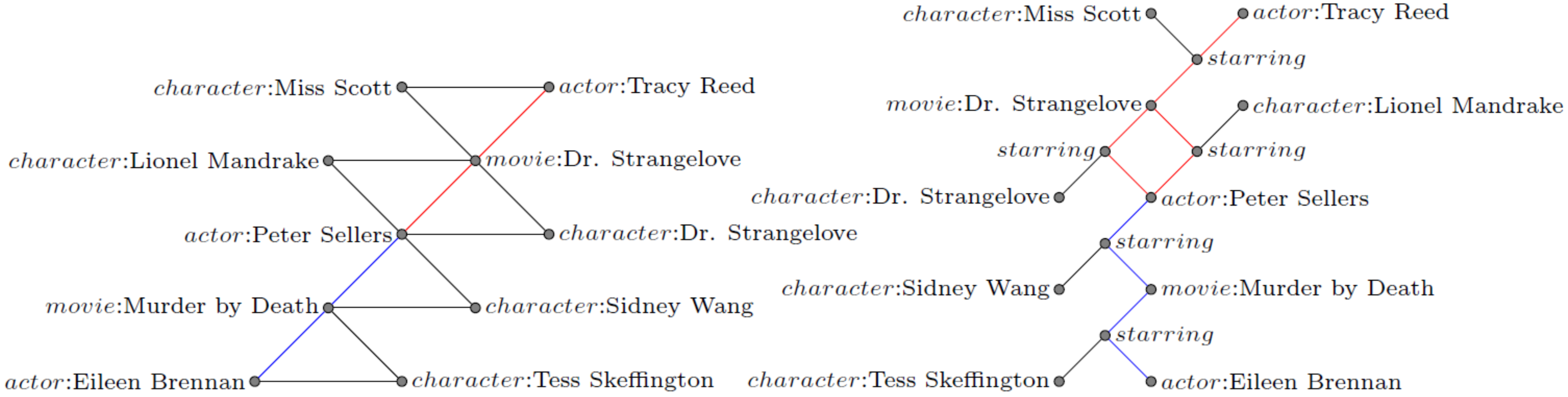
Algorithms	Movie DBs	Bib DBs
RWR [Tong, KDD'06]	0.204	0.718
SimRank [Jeh, KDD'02]	0.264	0.673
PathSim [Sun, VLDB'11]	0.110	0.688

Conclusion

- Introduced the problem of representation independent similarity search over graph data.
- Proposed a formal framework to measure generality
- Performed empirical study and showed that some of the well-known similarity search algorithms are not general.

Unused Slides

When PathSim fails?



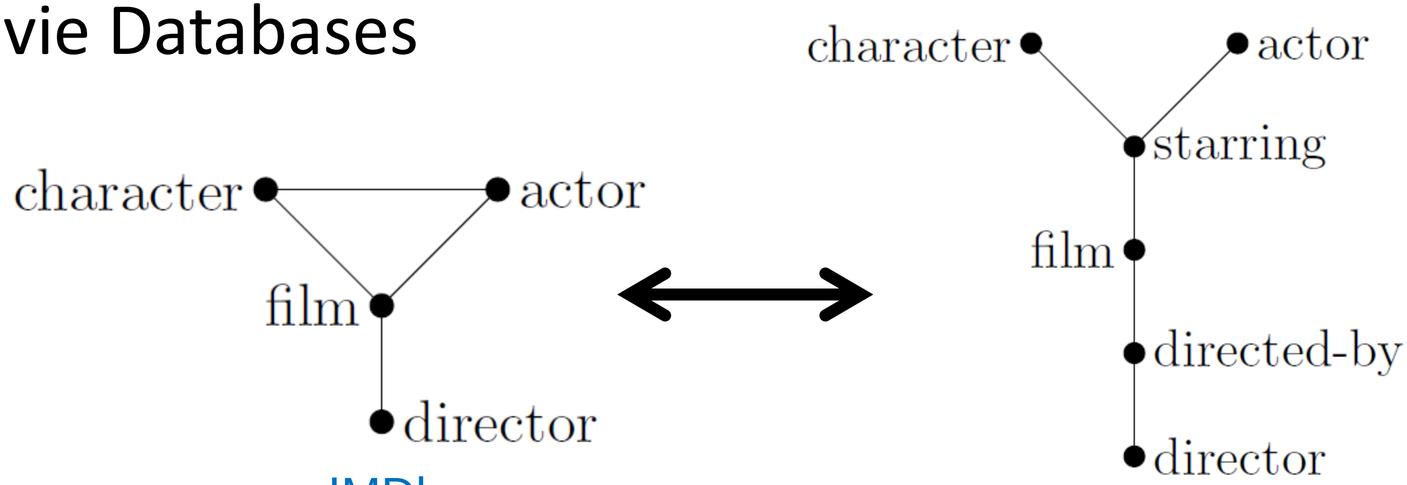
(a) IMDb

(b) Freebase

Figure 5: Fragments of IMDb (*imdb.com*) and Freebase (*freebase.com*) databases.

Transformations

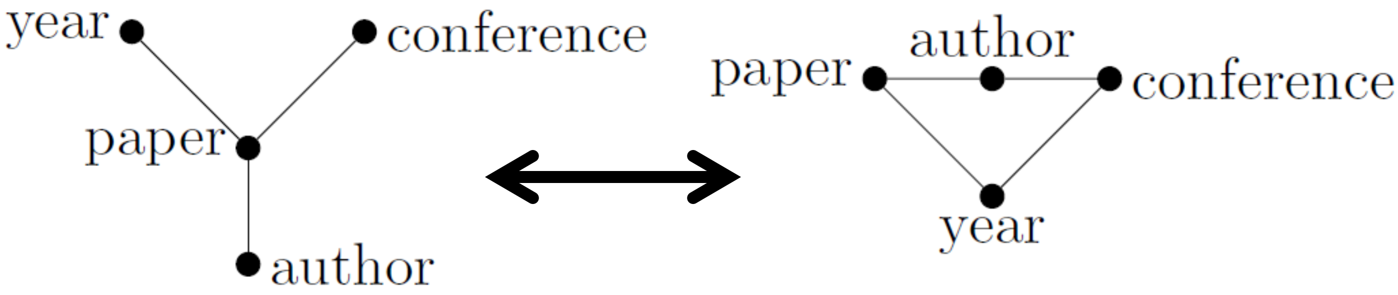
Movie Databases



IMDb

Freebase

Bibliography Databases



[Y. Sun, PVLDB’11]

[P. Zhao, CIKM’09]