Quantifying Causal Effects on Query Answering in Databases

Babak Salimi
University of Washington

February 2016

Collaborators: Leopoldo Bertossi (Carleton University), Dan Suciu (University of Washington), Guy Van den Broeck (UCLA)

Causality

- Causality appears at the foundations of many scientific disciplines
- Since inferring causal relationship is one of the central tasks of science, it is a topic that has been heavily debated in Philosophy, Statistics, AI, Law ...
- It turns out that we have similar questions in data management!
  - which tuple(s) in the instance caused that output to the query?
  - which tuple(s) in the instance caused the inconsistency?
  - which tuple(s) in the instance caused an undesired tuple(s) in a view?

Causality

- There are two lines of research in causal inference
  - Discovering actual causation: In this line of work the intention is to answer a particular instantiation of a causal relation e.g., Joes smoking caused his cancer
    - Many competing approaches, the one by Halpern and Pearl (HP) has been more influential in computer science

      (Pearl and Halpern 2005, Halpern 2015)
  - Discovering type causation or general causal claims: the objective is to discover general causal claims such as smoking causes cancer which refer to a class of events. There are two prominent approach:
    - Statistical Models of Causation (Neyman 1923, Rubin 1970)
    - Structural Equation Modeling (Wright 1921, Pearl 2000)

Causality in Databases

- Study of causality in data management started with HP-causality for query answers from relational databases

(Meliou et al., 2010a,2010b)

- A tuple t is a counterfactual cause to a query answer if excluding from the database falsifies the answer

- A tuple t is an actual cause for a query answer if there exists a contingency (obtained by removing a set of tuples) in which t is counterfactual

Causality in Databases

- Responsibility of t defined as $\rho_t = \frac{1}{1+n}$; n = the minimum number of changes (the interventions) in the model that make t a counterfactual cause (Meliou et al., 2010a)
- The more changes needed, the less responsibility of t
- An example due to (Meliou et al., 2010a) (reused here for comparison purposes)

Example

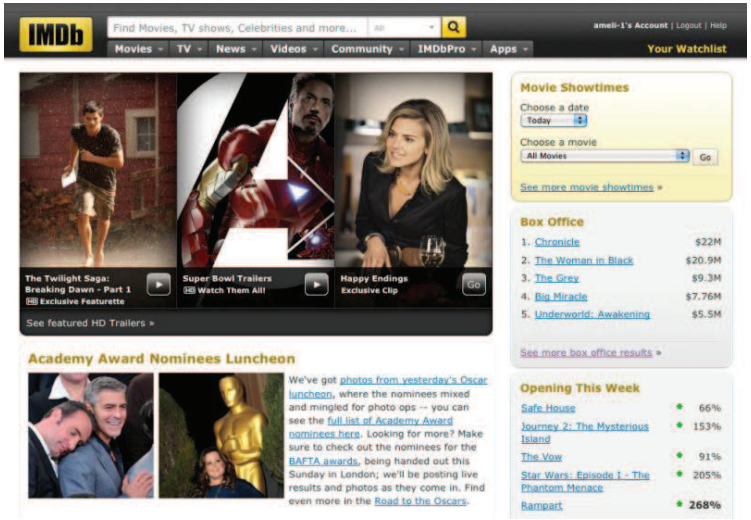
IMDB Database Schema

$Director(Did, Firstname, Lastname)$

$Movie(Mid, Name, Year, Rank)$

$Genre(Mid, Genre)$

$Movie\_Director(Did, Mid)$



Example

Query

IMDB Database Schema

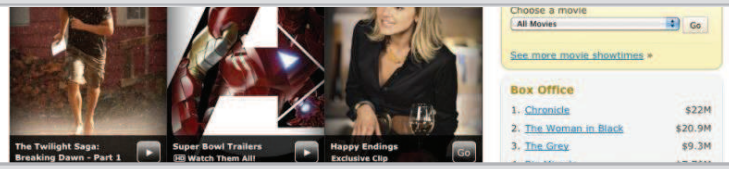
$Director(Did, Firstname, Lastname)$

$Movie(Mid, Name, Year, Rank)$

$Genre(Mid, Genre)$

$Movie\_Director(Did, Mid)$

"What genres does Tim Burton direct?"



$\exists\, Did\, \exists Mid\, \exists Firstname\, \exists Name\, \exists Year\, \exists Rank$

$(Director(Did, Firstname, \text{Burton}) \wedge Movie\ Director(Did, Mid,$

$Year, Rank) \wedge Movie(Mid, Name) \wedge Genre(Mid, Genre))$

Genre
...
Fantasy
History
Horror
Music
Musical
Mystery
Romance
...

?

Example

Query

IMDB Database Schema

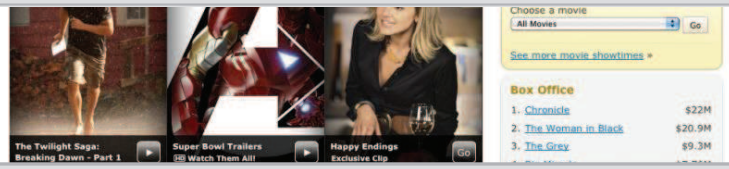
$Director(Did, Firstname, Lastname)$

$Movie(Mid, Name, Year, Rank)$

$Genre(Mid, Genre)$

$Movie\_Director(Did, Mid)$

"What genres does Tim Burton direct?"



$\exists\, Did\, \exists Mid\, \exists Firstname\, \exists Name\, \exists Year\, \exists Rank$
$(Director(Did, Firstname, \text{Burton}) \wedge Movie\ Director(Did, Mid,$
$Year, Rank) \wedge Movie(Mid, Name) \wedge Genre(Mid, Genre))$

Genre
...
Fantasy
History
Horror
Music
Musical
Mystery
Romance
...

?

I didn't know that Tim Burton directs Musicals!
Why are these items in the result of my query?

Example

Query

IMDB Database Schema

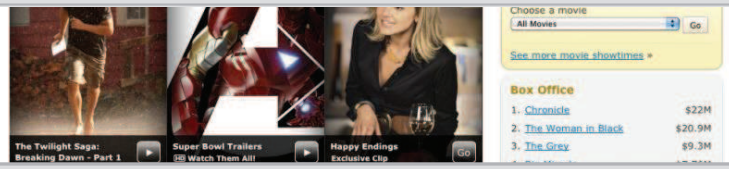
$Director(Did, Firstname, Lastname)$

$Movie(Mid, Name, Year, Rank)$

$Genre(Mid, Genre)$

$Movie\_Director(Did, Mid)$

"What genres does Tim Burton direct?"



$\exists\, Did\, \exists Mid\, \exists Firstname\, \exists Name\, \exists Year\, \exists Rank$
$(Director(Did, Firstname, \text{Burton}) \wedge Movie\ Director(Did, Mid,$
$Year, Rank) \wedge Movie(Mid, Name) \wedge Genre(Mid, Genre))$

Genre
...
Fantasy
History
Horror
Music
Musical
Mystery
Romance
...

?

What can databases do ?

Why-Provenance / Lineage:
Set of all minimal subsets of the database that entail the output tuple

(Buneman ICDT, 2001, Tannen EDBT 2010)

But

In this example, the lineage includes **137 tuples !!**

Example

Query

IMDB Database Schema

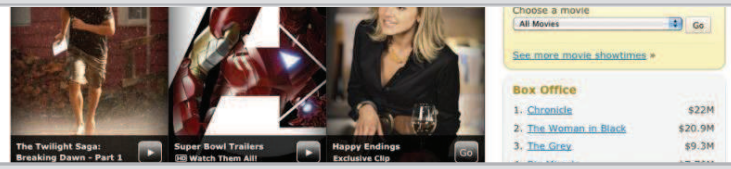
$Director(Did, Firstname, Lastname)$

$Movie(Mid, Name, Year, Rank)$

$Genre(Mid, Genre)$

$Movie\_Director(Did, Mid)$

"What genres does Tim Burton direct?"



$\exists\, Did\, \exists Mid\, \exists Firstname\, \exists Name\, \exists Year\, \exists Rank$
$(Director(Did, Firstname, \text{Burton}) \wedge Movie\ Director(Did, Mid,$
$Year, Rank) \wedge Movie(Mid, Name) \wedge Genre(Mid, Genre))$

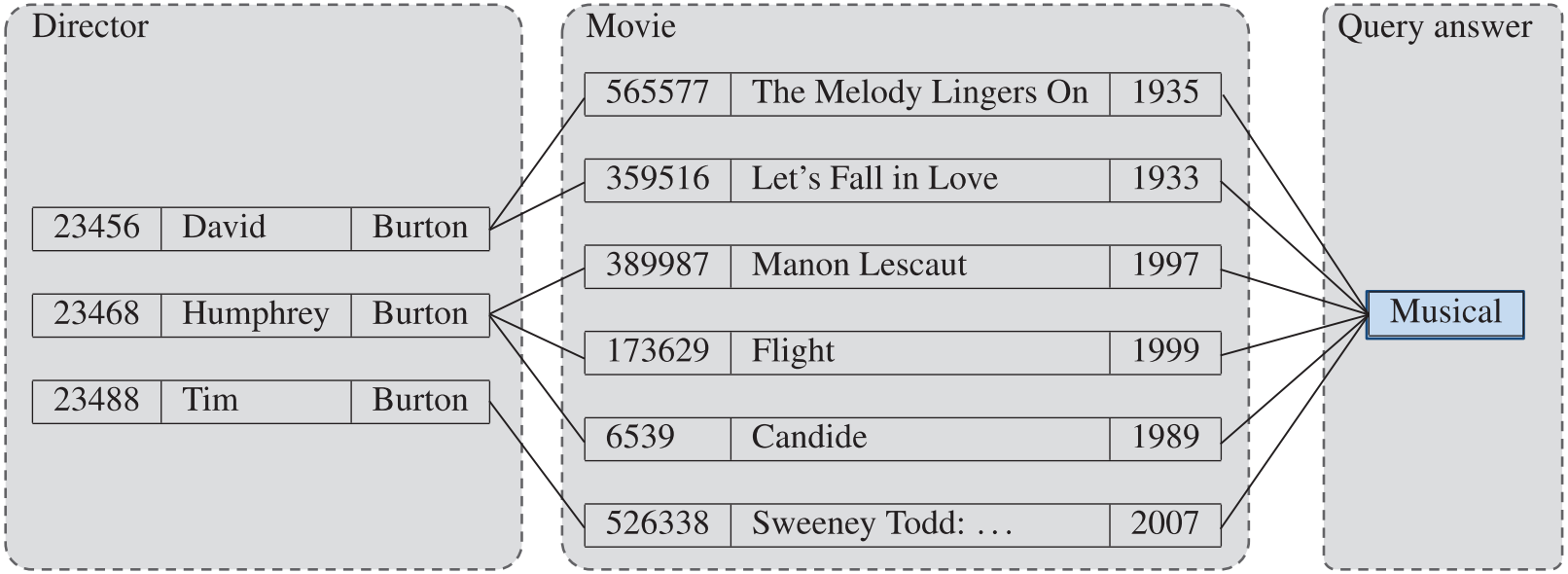
Genre
...
Fantasy
History
Horror
Music
Musical
Mystery
Romance
...

?

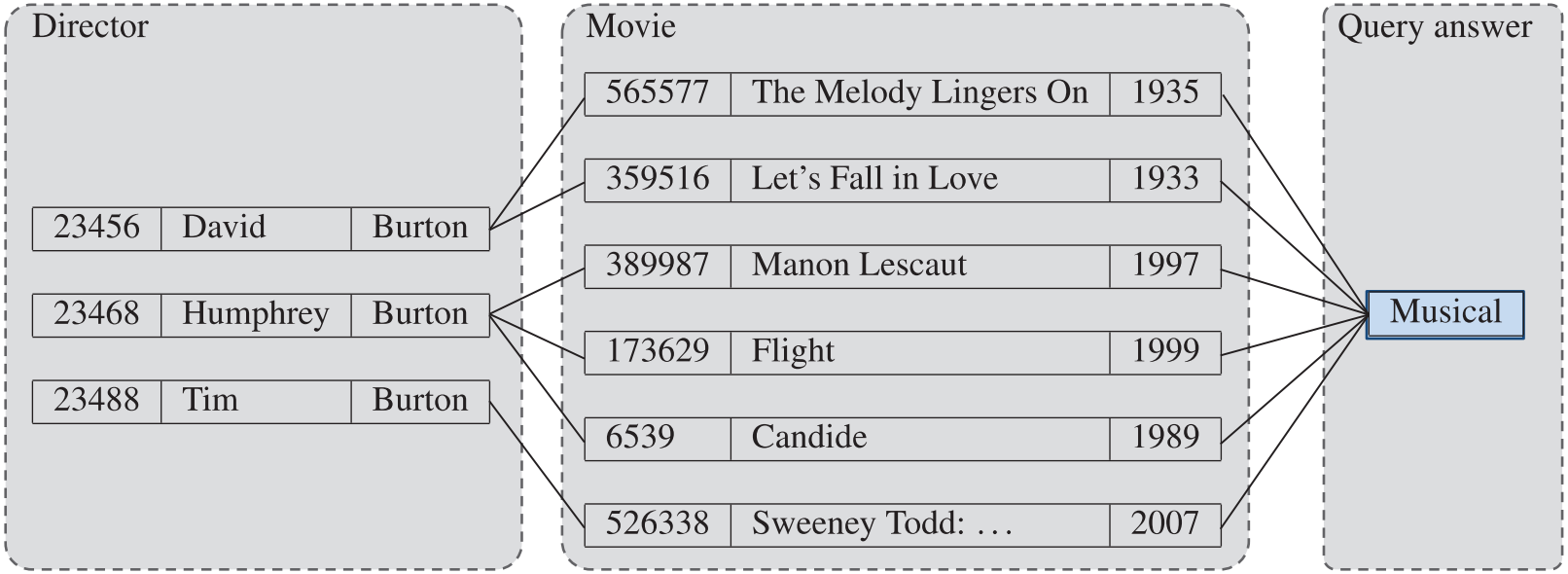
When the size of the **lineage** is large it becomes **impractical** to manually examine it ...

Causality???

Example


Director
23456 David Burton
23468 Humphrey Burton
23488 Tim Burton
Movie
565577 The Melody Lingers On 1935
359516 Let's Fall in Love 1933
389987 Manon Lescaut 1997
173629 Flight 1999
6539 Candide 1989
526338 Sweeney Todd: ... 2007
Query answer
Musical

Example

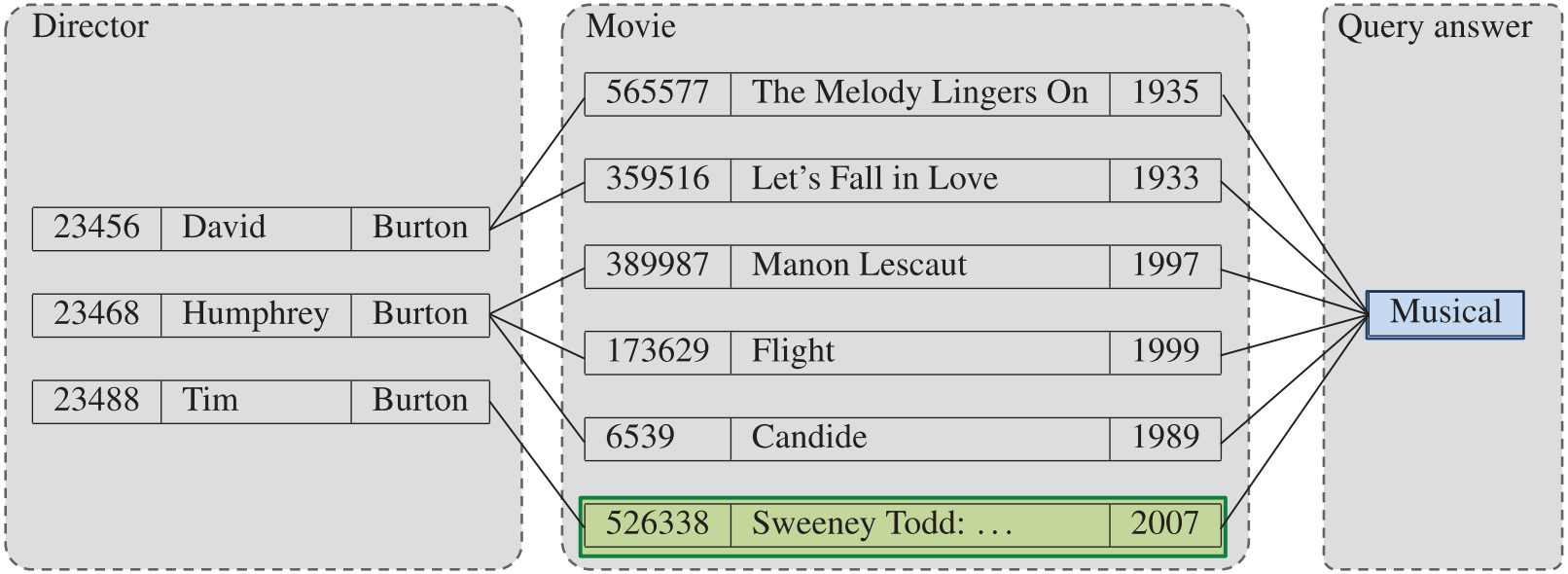


The lineage contains **uninteresting** tuples, e.g. tuples from "Gender" and "Movie_Director" are **not** of interest

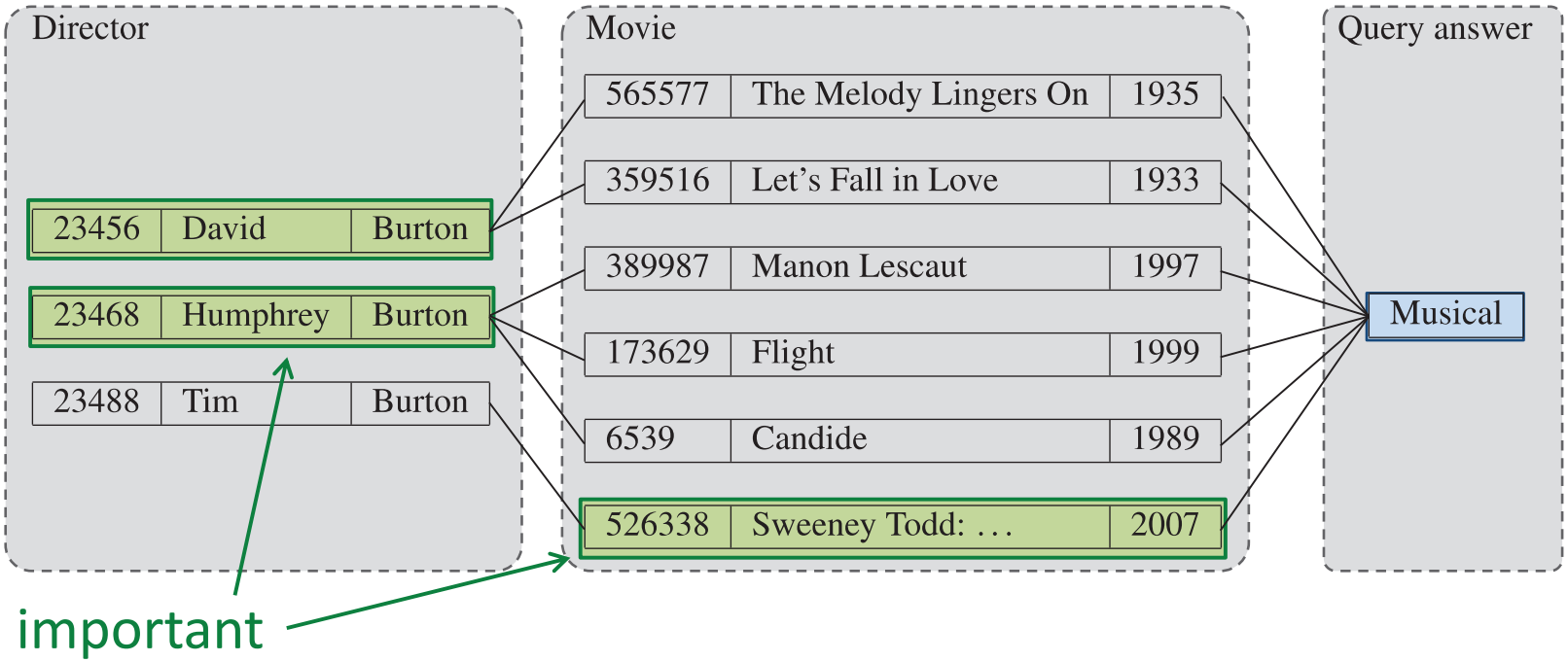
Goal: Partition the database : Capture users' **Preferences**

These **partitions** are well-defined in the context of **causality**

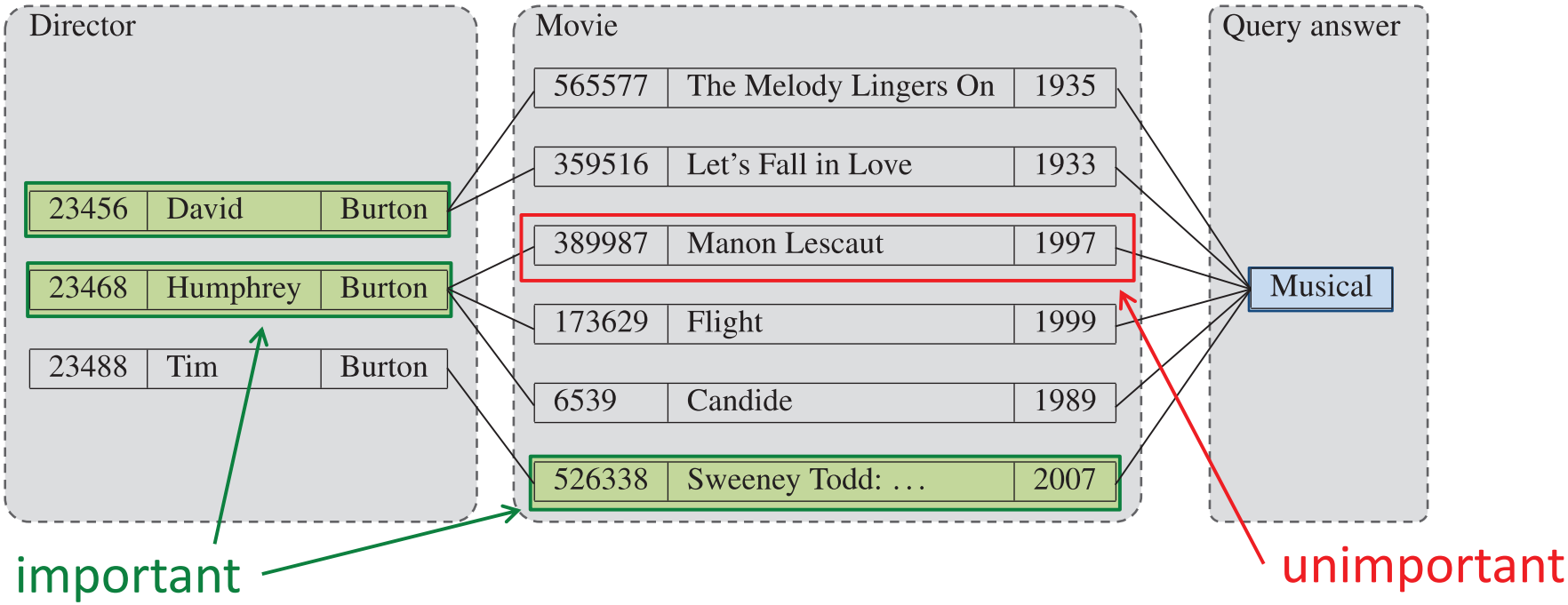
Example


Director
23456 David Burton
23468 Humphrey Burton
23488 Tim Burton
Movie
565577 The Melody Lingers On 1935
359516 Let's Fall in Love 1933
389987 Manon Lescaut 1997
173629 Flight 1999
6539 Candide 1989
526338 Sweeney Todd: ... 2007
Query answer
Musical

Example

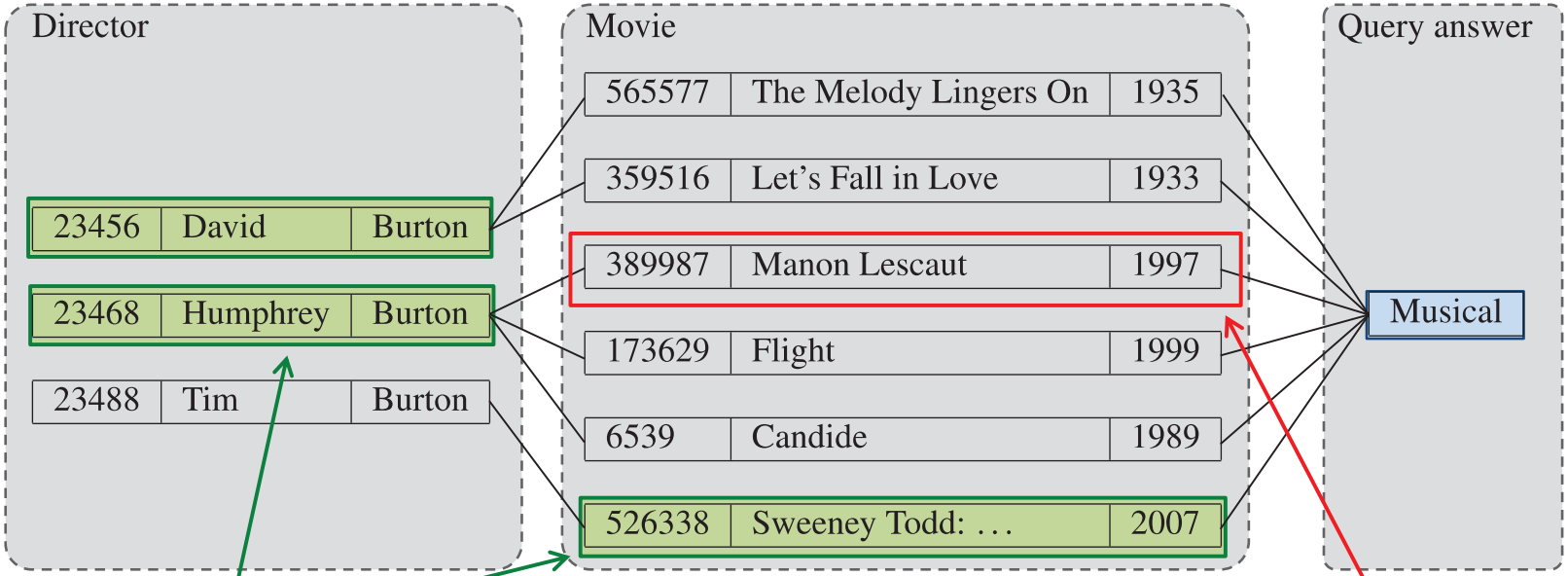

Director
23456 David Burton
23468 Humphrey Burton
23488 Tim Burton
Movie
565577 The Melody Lingers On 1935
359516 Let's Fall in Love 1933
389987 Manon Lescaut 1997
173629 Flight 1999
6539 Candide 1989
526338 Sweeney Todd: ... 2007
Query answer
Musical
important

Example



Captured by the notion of **causal responsibility** (degree of causation)

Example



important

unimportant

Goal:

Rank tuples in order of importance

Ranking Provenance

Answer tuple	ρ_t
Movie(526338, "Sweeney Todd", 2007)	0.33
Director(23456, David, Burton)	0.33
Director(23468, Humphrey, Burton)	0.33
Director(23488, Tim, Burton)	0.33
Movie(359516, "Let's Fall in Love", 1933)	0.25
Movie(565577, "The Melody Lingers On", 1935)	0.25
Movie(6539, "Candide", 1989)	0.20
Movie(173629, "Flight", 1999)	0.20
Movie(389987, "Manon Lescaut", 1997)	0.20

Causal Responsibility

- It has been shown that there is a close connection between causal responsibility and other notions in databases, e.g. some forms of delete-propagation, database repairs and diagnosis problems

  (Salimi and Bertossi, NMR 2014, ICDT 2015, UAI 2015, FLAIRS 2015)

- Therefore, causal responsibility is an important notion, which captures and unifies many problems in data management

- However, we argue that causal responsibility only partially fulfill its original intention
  - i.e., to provide a metric for the causal contribution of a tuple to a query answer

An Issue with Causal Responsibility

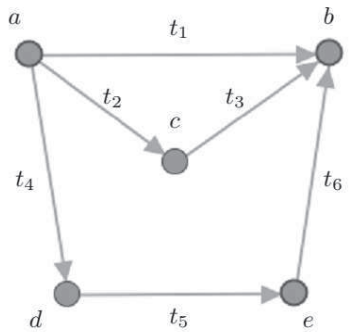
Database D (single binary relation E):

E	X	Y
t_1	a	b
t_2	a	c
t_3	c	b
t_4	a	d
t_5	d	e
t_6	e	b

Boolean Datalog query:

$$\begin{aligned} \Pi: ans &\leftarrow P(a,b) \\ P(x,y) &\leftarrow E(x,y) \\ P(x,y) &\leftarrow P(x,z), E(z,y) \end{aligned}$$

Graph G associated to D



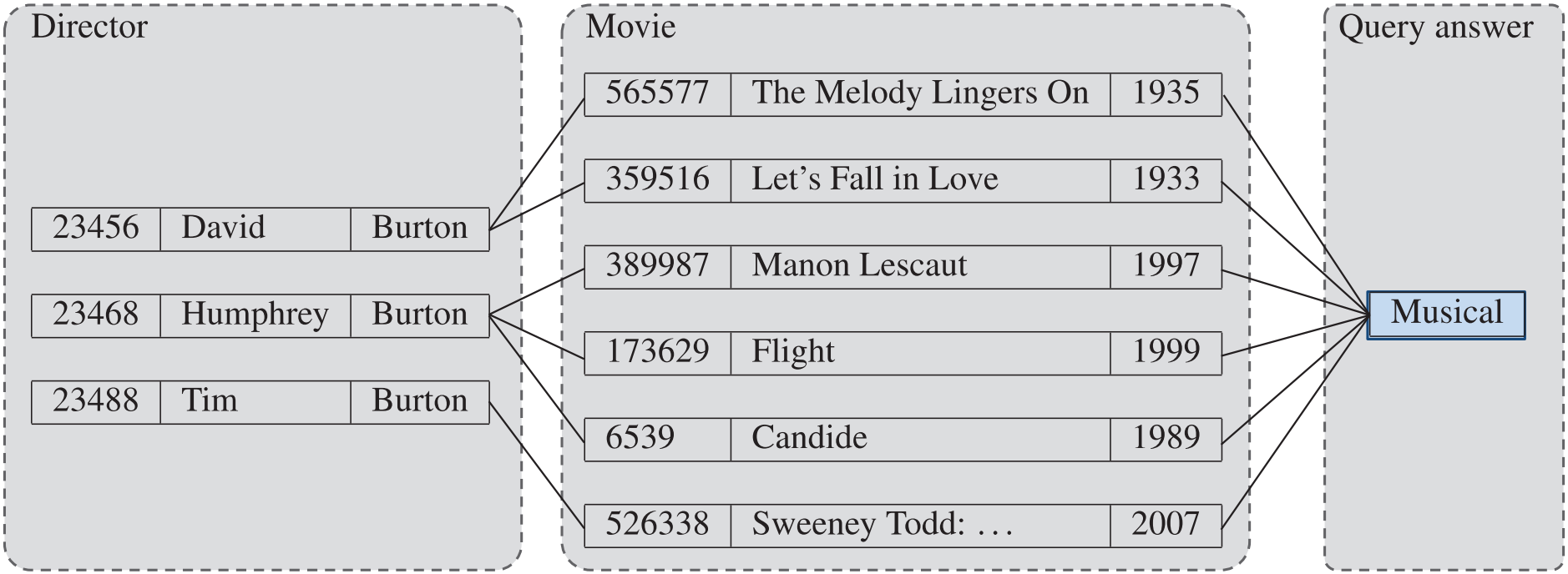
Π : Is there any path between a and b in G?

$D \cup \Pi \models ans$ **YES!**

All tuples have the **same** causal responsibility $\frac{1}{3}$

Counterintuitive!

A Problem with Causal Responsibility



- The query was about all genders associated with "Burton"
- Who makes more contribution to the category musical?
- The answer is "Humphrey Burton" (movies should come after the directors in the ranking)

From Responsibility to Causal Effect

- Lets backtrack to the question we start with and see if we can come up with a more meaningful metric
- QA-Causality: "What would be (or how would change) the answer to a query $\mathcal{Q}$ if the tuple t is deleted/inserted (*intervention*) from/into the database D?"
- It turns out that the lineage expression can be used as a basis to do this sort of analysis as statistician approach it
- Example: Instance $D = \{R(a,b), R(a,c), R(c,b), S(c)\}$

  Query $\mathcal{Q}: \exists x(R(x,y) \wedge S(y))$

  Lineage: $\Phi_{\mathcal{Q}} = (X_{R(a,b)} \wedge X_{S(b)}) \vee (X_{R(c,b)} \wedge X_{S(b)})$

Causal Effect

- Now, we can use Pearl's notation for *interventions*, i.e. expressions of the form $do(X_t = x)$, where $x \in \{0, 1\}$
- Interventions are assignments (or changes) of truth values to (some of) the variables in the lineage
- If we consider that assignments can be randomly and uniformly chosen, we obtain a probability space
- Furthermore, we obtain properly defined conditional probabilities of the form $P(\mathcal{Q} = y \mid do(\bar{X} = \bar{x}))$

Causal Effect

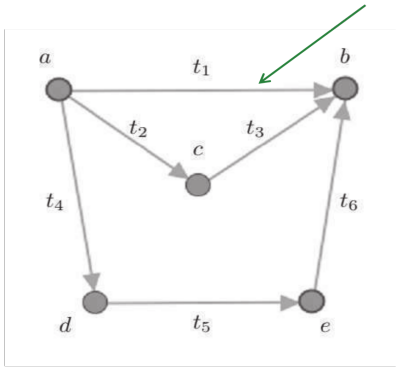
- Causal Effects of t:

$$\mathcal{E}^{D}_{t,\mathcal{Q}} := E(\mathcal{Q} \mid do(X_t = 1)) - E(\mathcal{Q} \mid do(X_t = 0))$$

- t is an actual cause for $\mathcal{Q}$ iff $\mathcal{E}^{D}_{t,\mathcal{Q}} > 0$

Causal Effect: Examples

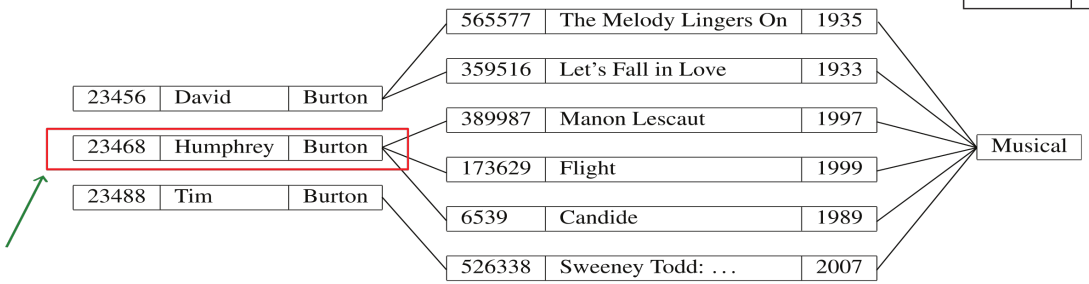
Path example:



All tuples have the **same** causal responsibility $\frac{1}{3}$

C	a, b
0.65	t_1
0.21	t_2
0.21	t_3
0.09	t_4
0.09	t_5
0.09	t_6

IMDB example:



ρ	Musical
0.33	Movie(526338, "Sweeney Todd", 2007)
0.33	Director(23456, David, Burton)
0.33	**Director(23468, Humphrey, Burton)**
0.25	Director(23488, Tim, Burton)
0.25	Movie(359516, "Lets Fall in Love", 1933)
0.25	Movie(565577, "he Melody Lingers On", 1935)
0.20	Movie(6539, "Candide", 1989)
0.20	Movie(173629, "Flight", 1999)
0.20	Movie(389987, "Manon Lescaut", 1997)

C	Musical
0.28	**Director(23468, Humphrey, Burton)**
0.22	Director(23456, David, Burton)
0.12	Director(23488, Tim, Burton)
0.12	Movie(526338, "Sweeney Todd", 2007)
0.07	Movie(359516, "Let's Fall in Love", 1933)
0.07	Movie(565577, "he Melody Lingers On", 1935)
0.04	Movie(6539, "Candide", 1989)
0.04	Movie(173629, "Flight", 1999)
0.04	Movie(389987, "Manon Lescaut", 1997)

Conclusions and Related Work

- Causal effect can deal with non-monotone queries and queries with aggregates
- Related to Pearson correlation coefficient
- Causal effect would capture the notion of network vulnerability
- It is related to the notion of influence in the context of Fourier transformation of Boolean functions
- Causal effect can rank predicate-based explanations for query answers! The scoring function as being used in this context is an extension of the notion of causal responsibility
- We can adopt results from probabilistic databases to analyse and compute causal effect!