Logic and Algebra for Query Evaluation, Simons 2023

On the Accessibility and Privacy of Provenance-Based Explanations

Amir Gilad

The Hebrew University

Joint work with Daniel Deutch, Yuval Moskovitch, Nave Frost, Ariel Frankenthal

PRACTICAL ISSUES WITH RAW PROVENANCE

RAW PROVENANCE CAN BE...

TOO LONG AND COMPLEX

```
(oname,Duke)·(aname,Jun Y.)·(ptitle,iCheck...)·(cname,SIGMOD)·(pyear,14′)+
(oname,Duke)·(aname,Jun Y.)·(ptitle, Scalable...)·(cname,VLDB)·(pyear,06′)+
(oname,Duke)·(aname,Jun Y.)· (ptitle, Making...)·(cname,VLDB)·(pyear,07′)+
(oname,Duke)·(aname,Brett W.)·(ptitle,iCheck...)·(cname,SIGMOD)·(pyear,14′)+
(oname,Duke)·(aname,Jun Y.)·(ptitle,Cumulon...)·(cname,SIGMOD)·(pyear,14′)+
...
```

TOO REVEALING

PROVENANCE-BASED EXPLANATIONS → PRIVATE/PROPRIETARY QUERY

IN THIS TALK: WHEN RAW PROVENANCE IS NOT ENOUGH

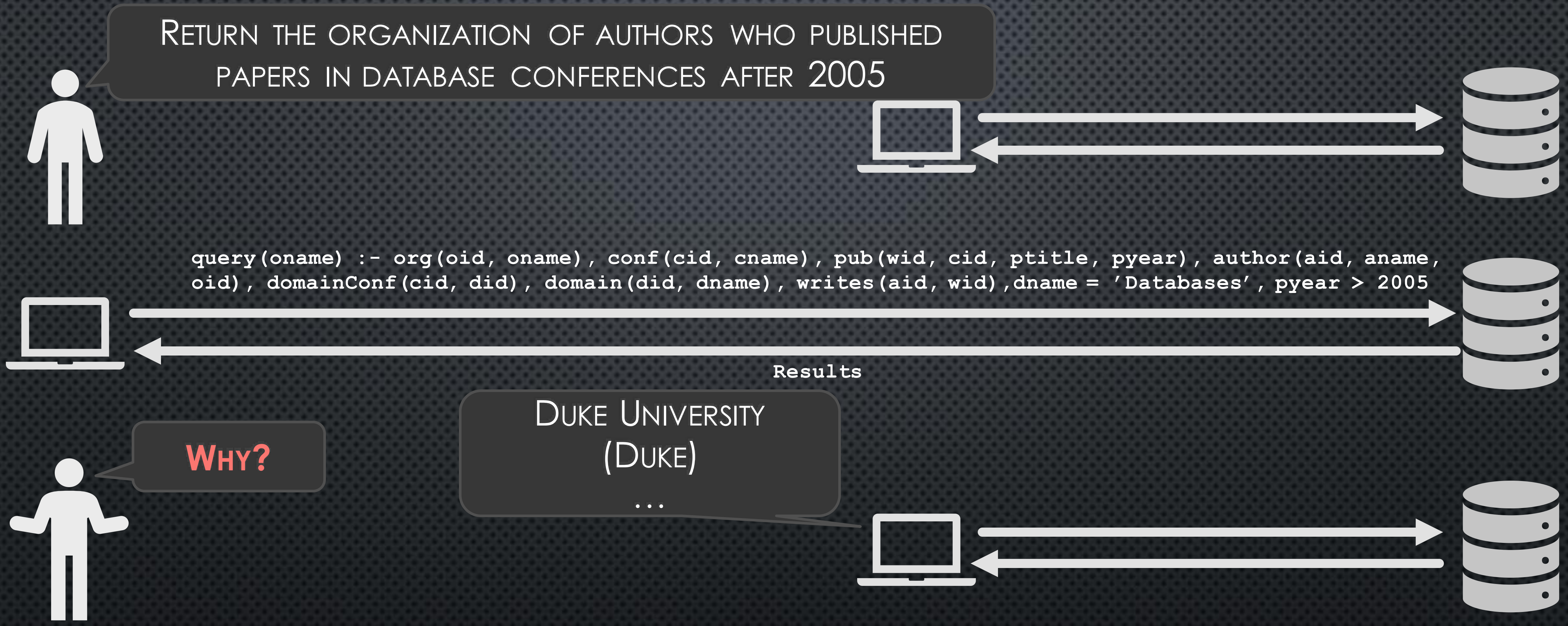
FACTORIZING AND SUMMARIZING PROVENANCE FOR NATURAL LANGUAGE EXPLANATIONS

VLDB 16', VLDB 17', SIGMOD REC. 18', VLDB J. 20'

ABSTRACTING PROVENANCE FOR QUERY PRIVACY

SIGMOD 21', ICDE 21'

NATURAL LANGUAGE INTERFACES AND EXPLANATIONS



Duke is the organization of 63 authors who published 170 papers in 31 conferences in 2006 - 2015

PROVENANCE MODEL

NL QUERY:

RETURN THE ORGANIZATION OF AUTHORS WHO PUBLISHED PAPERS IN DATABASE CONFERENCES AFTER 2005

QUERY:

```
query(oname) :- org(oid, oname), conf(cid, cname), pub(wid, cid, ptitle,
pyear), author(aid, aname, oid), domainConf(cid, did), domain(did,
dname), writes(aid, wid),dname = 'Databases', pyear > 2005
```

PROVENANCE MODEL

NL Query:

Return the organization of authors who published papers in database conferences after 2005

Query:

```
query(oname) :- org(oid, oname), conf(cid, cname), pub(wid, cid, ptitle,
pyear), author(aid, aname, oid), domainConf(cid, did), domain(did,
dname), writes(aid, wid),dname = 'Databases', pyear > 2005
```

Provenance of the result Duke:

```
(oname,Duke)·(aname,Jun Y.)·(ptitle,iCheck...)·(cname,SIGMOD)·(pyear,14')+
(oname,Duke)·(aname,Jun Y.)·(ptitle, Scalable...)·(cname,VLDB)·(pyear,06')+
(oname,Duke)·(aname,Jun Y.)· (ptitle, Making...)·(cname,VLDB)·(pyear,07')+
(oname,Duke)·(aname,Brett W.)·(ptitle,iCheck...)·(cname,SIGMOD)·(pyear,14')+
(oname,Duke)·(aname,Jun Y.)·(ptitle,Cumulon...)·(cname,SIGMOD)·(pyear,14')+
...
```

PROVENANCE MODEL

NL QUERY:

RETURN THE ORGANIZATION OF AUTHORS WHO PUBLISHED PAPERS IN DATABASE CONFERENCES AFTER 2005

QUERY:

```
query(Duke) :- org(oid, Duke), conf(cid, cname), pub(wid, cid, iCheck...,
2014), author(aid, Jun Y., oid), domainConf(cid, did), domain(did,
SIGMOD), writes(aid, wid),dname = 'Databases', 2014 > 2005
```

PROVENANCE OF THE RESULT DUKE:

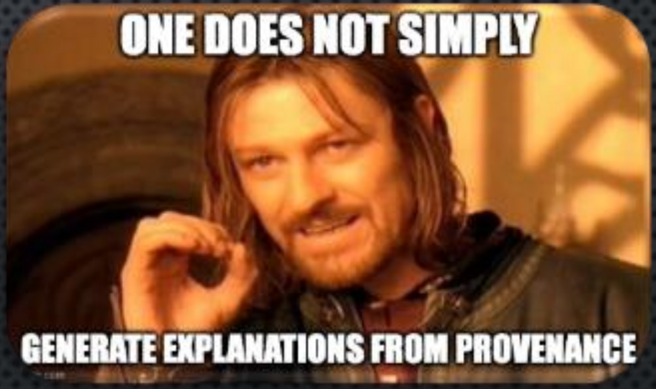
```
(oname,Duke)·(aname,Jun Y.)·(ptitle,iCheck...)·(cname,SIGMOD)·(pyear,14')+
(oname,Duke)·(aname,Jun Y.)·(ptitle, Scalable...)·(cname,VLDB)·(pyear,06')+
(oname,Duke)·(aname,Jun Y.)· (ptitle, Making...)·(cname,VLDB)·(pyear,07')+
(oname,Duke)·(aname,Brett W.)·(ptitle,iCheck...)·(cname,SIGMOD)·(pyear,14')+
(oname,Duke)·(aname,Jun Y.)·(ptitle,Cumulon...)·(cname,SIGMOD)·(pyear,14')+
…
```

Solution Overview

How do we convert provenance to a Natural Language explanation?

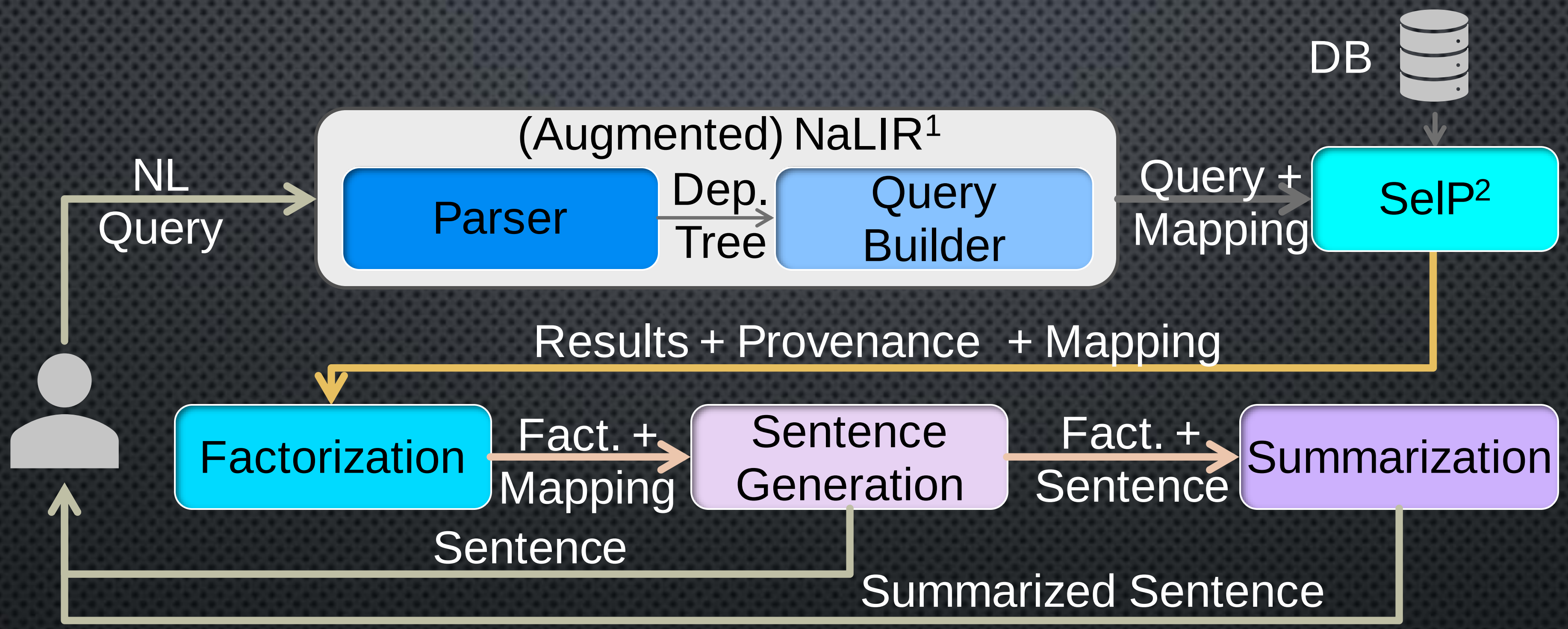
Challenges:

1. The formal provenance is far from an NL sentence
2. The provenance can be very long and convoluted



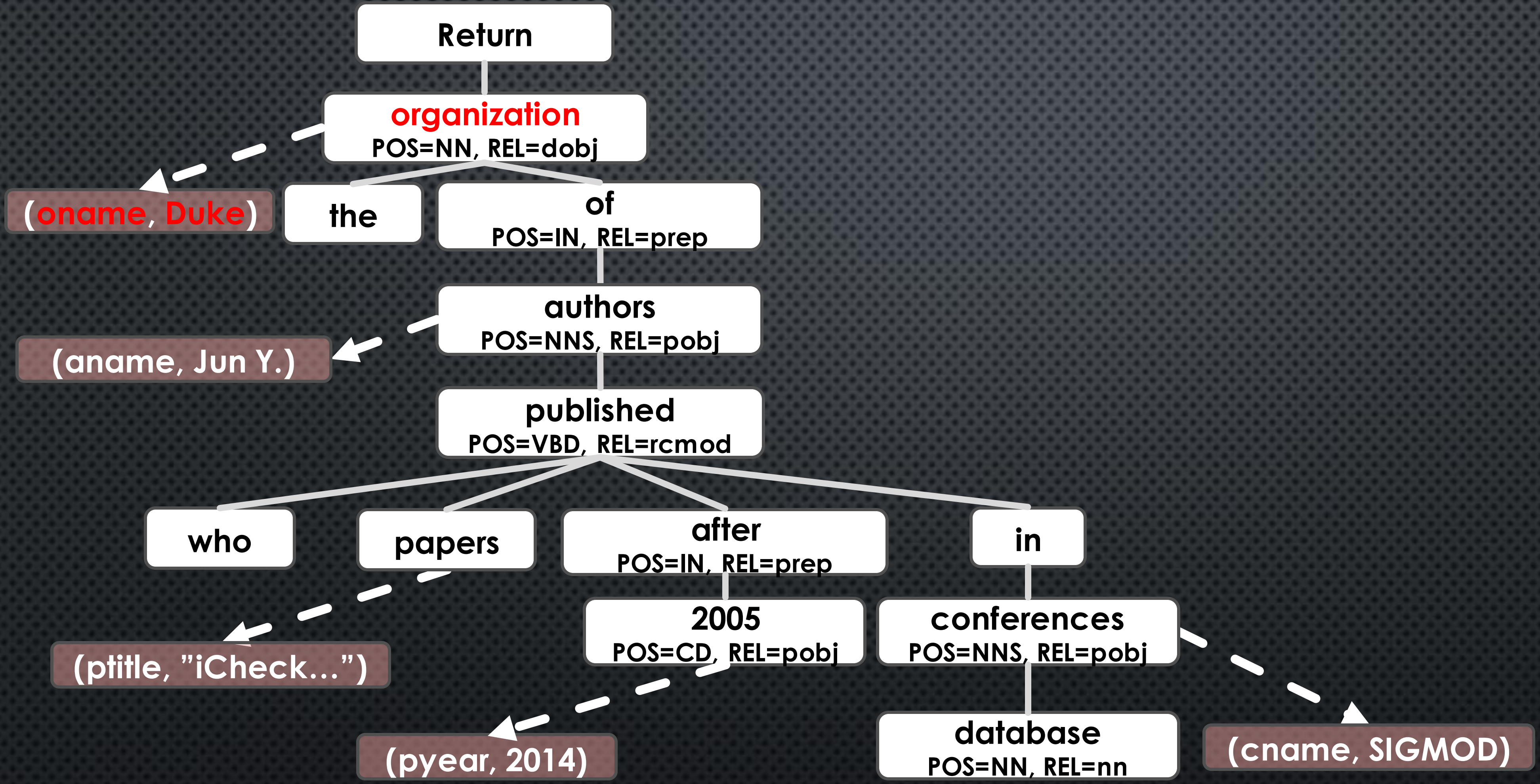
Use the structure of the input question!

FRAMEWORK

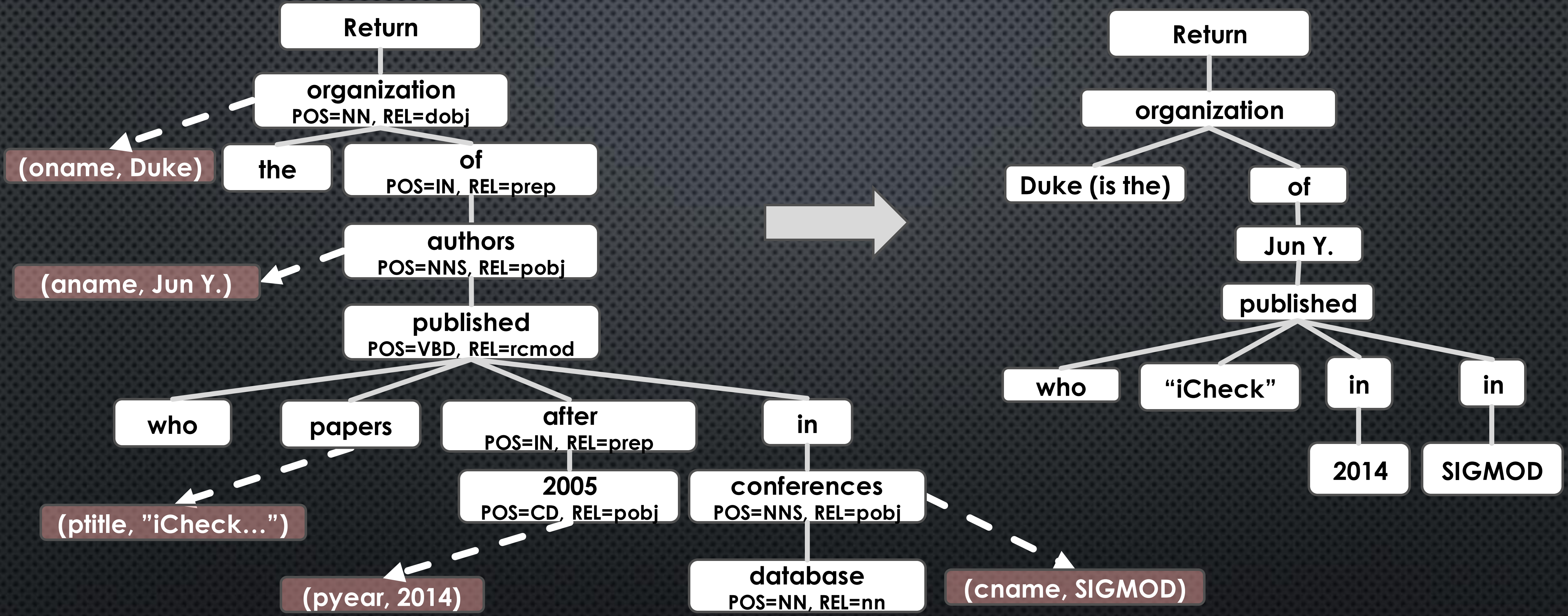


1. Li, F., Jagadish, H. V., "Constructing an Interactive Natural Language Interface for Relational Databases". In: Proc. VLDB Endow. (2014), pp. 73–84
2. Deutch, D., G., Moskovitch, Y., "Efficient provenance tracking for datalog using top-k queries". In: VLDB J. 27.2 (2018), pp. 245–269

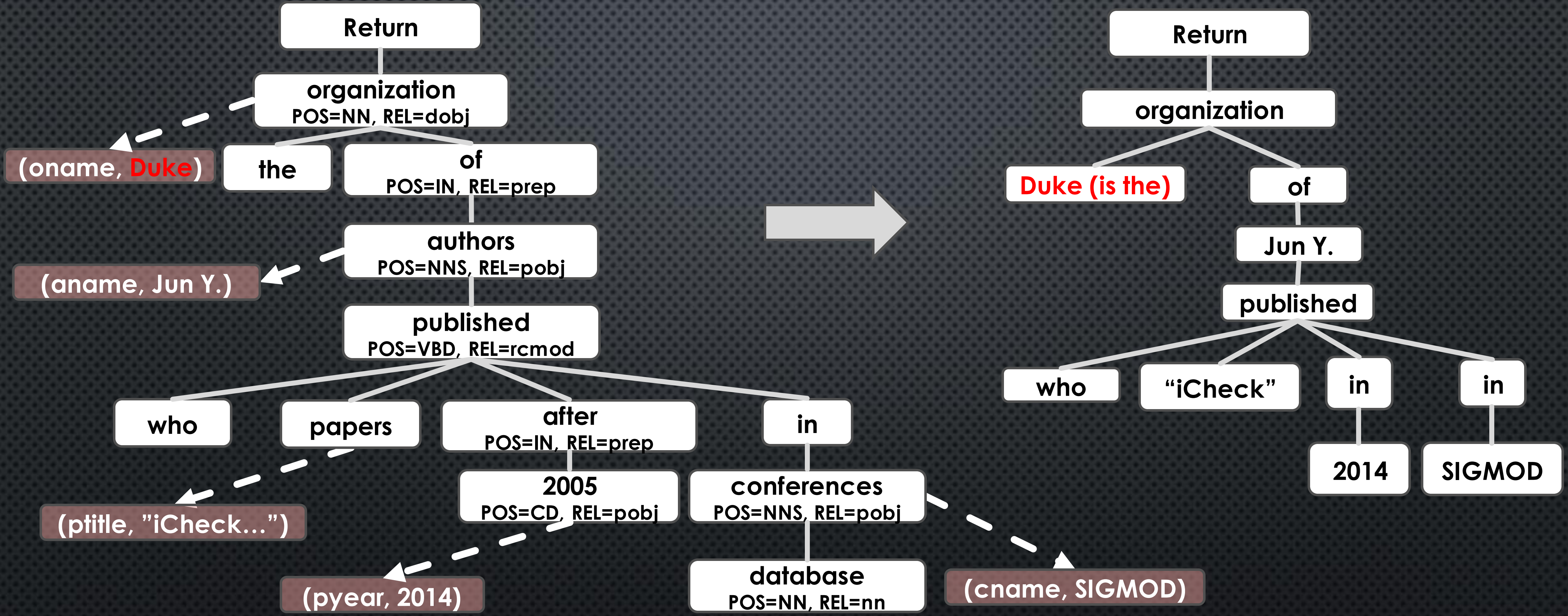
FROM MAPPINGS TO AN EXPLANATION



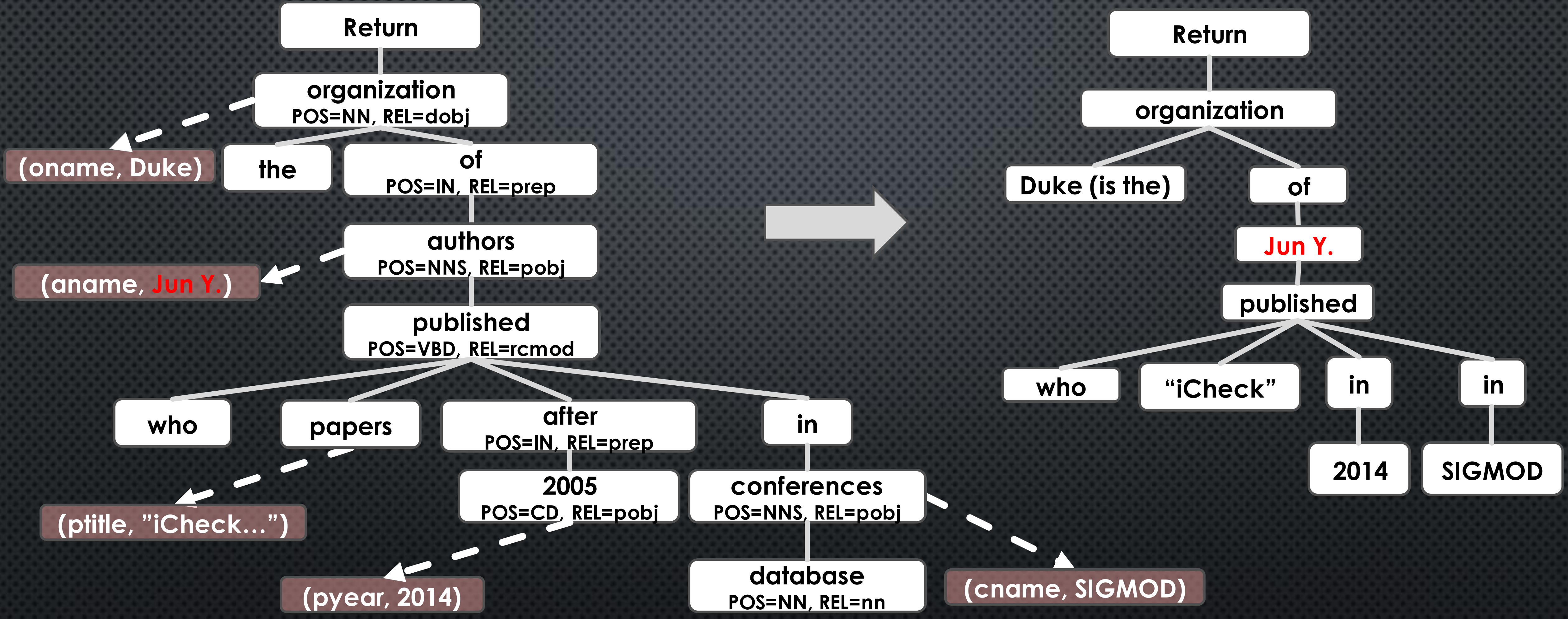
FROM MAPPINGS TO AN EXPLANATION



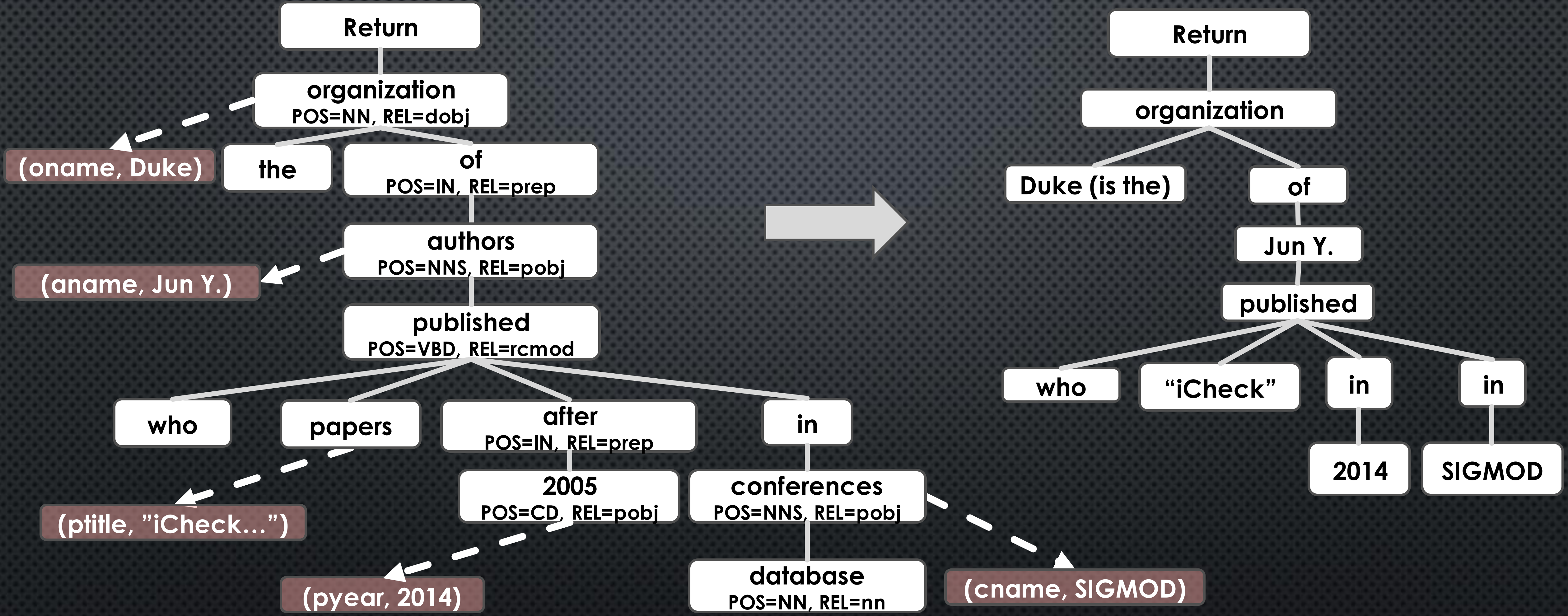
FROM MAPPINGS TO AN EXPLANATION



FROM MAPPINGS TO AN EXPLANATION



FROM MAPPINGS TO AN EXPLANATION



DUKE IS THE ORGANIZATION OF JUN Y. WHO PUBLISHED 'ICHECK…' IN SIGMOD IN 2014

PROVENANCE FACTORIZATION

Idea: Use algebraic factorization to take out common values that appear in multiple assignments

```
[Duke]·[Jun Y.]·[iCheck...]·[SIGMOD]·[2014]+
[Duke]·[Jun Y.]·[Scalable...]·[VLDB]·[2006]+
[Duke]·[Jun Y.]·[Making..]·[VLDB]·[2007]+
[Duke]·[Brett W.]·[iCheck...]·[SIGMOD]·[2014]+
[Duke]·[Jun Y.]·[Cumulon...]·[SIGMOD]·[2014]
```

Intuition: we want a factorization that follows the structure of the NL query to be able to generate a sentence

Shortest Fa

```
[Duke]·
 ([SIGMOD]·[2014]·
 ([iCheck...]·
 ([Jun Y.] + [Brett W.]))
  + [Jun Y.]·[Cumulon...])
  + [VLDB]·[Jun Y.]·
 ([2006]·[Scalable...])
  + [2007]·[Making...])
```

```
[Duke]·
([Jun Y.]·
  ([VLDB]·
  ([2006]·[Scalable...]
   + [2007]·[Making...]))
   + [SIGMOD]·[2014]·
  ([iCheck...] + [Cumulon...]))
+ [Brett W.]·[iCheck...]·[SIGMOD]·[2014])
```

T-COMPATIBILITY

NL QUERY:

RETURN THE ORGANIZATION OF AUTHORS WHO PUBLISHED PAPERS IN DATABASE CONFERENCES AFTER 2005

SHORTEST FACTORIZATION:

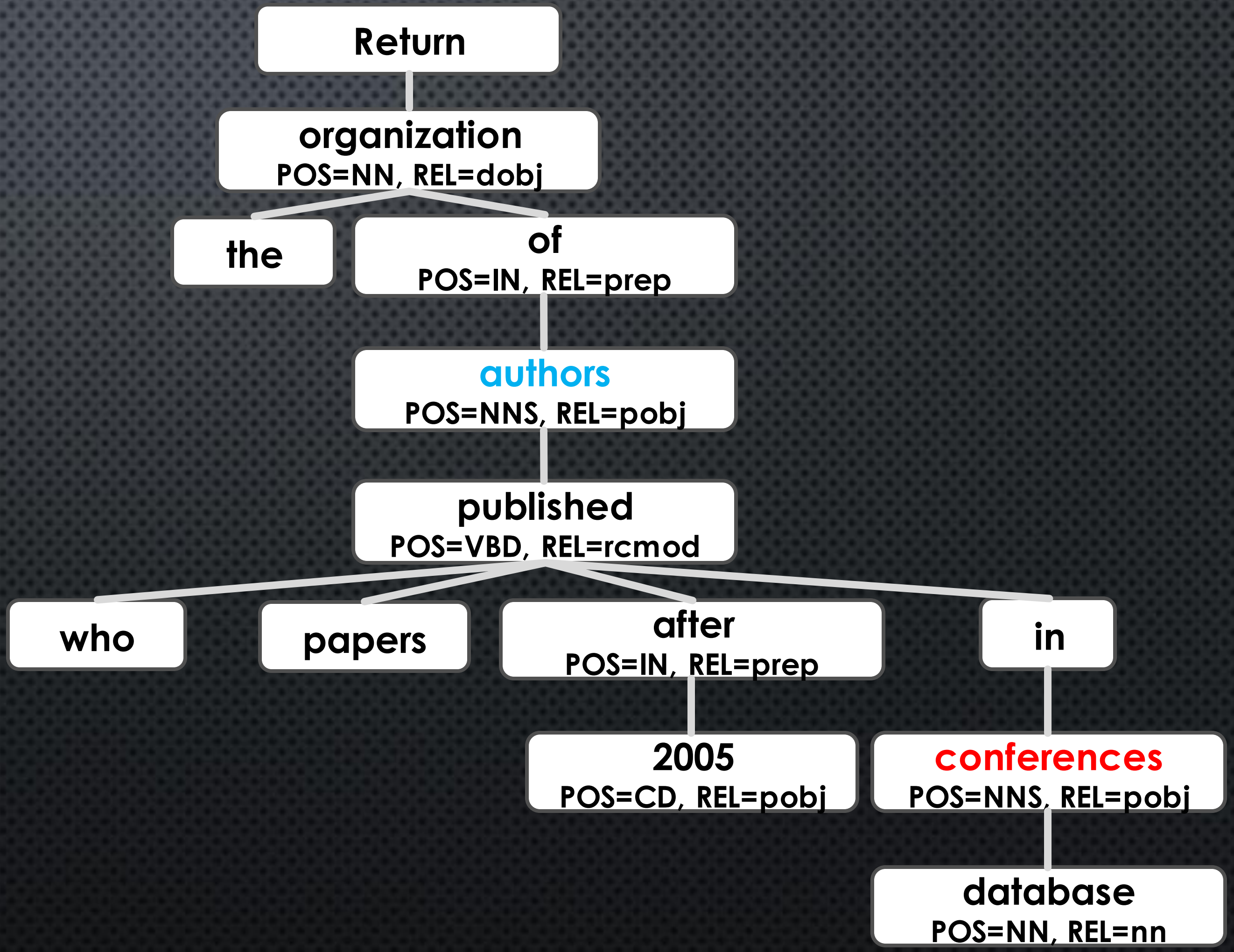
```
[Duke]·
 ([SIGMOD]·[2014]·
 ([iCheck...]·
 ([Jun Y.] + [Brett W.]))
  + [Jun Y.]·[Cumulon...])
  + [VLDB]·[Jun Y.]·
 ([2006]·[Scalable...])
  + [2007]·[Making...])
```

AS A SENTENCE:

```
Duke is the organization of authors who published in
 SIGMOD 2014
 'iCheck...' which was published by
 Jun Y. and Brett W.
 and Jun Y. published 'Cumulon...'
 and Jun Y. published in VLDB
 'Scalable...' in 2014
 and 'Making...' in 2007.
```

T-COMPATIBILITY

```
[Duke]·
 ([SIGMOD]·[2014]·
 ([iCheck...]·
 ([Jun Y.] + [Brett W.]))
  + [Jun Y.]·[Cumulon...])
  + [VLDB]·[Jun Y.]·
 ([2006]·[Scalable...])
  + [2007]·[Making...])
```



T-Compatibility

NL Query:

Return the organization of authors who published papers in database conferences after 2005

Longer Factorization:

```
[Duke]·
([Jun Y.]·
  ([VLDB]·
  ([2006]·[Scalable...]
   + [2007]·[Making...]))
   + [SIGMOD]·[2014]·
  ([iCheck...] + [Cumulon...]))
+ [Brett W.]·[iCheck...]·[SIGMOD]·[2014])
```

As a Sentence:

```
Duke is the organization of
 Jun Y. who published
 in VLDB
 'Scalable...' in 2006 and
 'Making...' in 2007
 and in SIGMOD in 2014
 'iCheck...' and 'Cumulon...'
 and Brett W. who published
 'iCheck...' in SIGMOD in 2014.
```

Finding T-Compatible Factorizations

<u>Algorithm:</u>

- traverse the dependency tree level-by-level
- For every level with mapped words, factorize their corresponding values in the provenance
- Prioritize which values to take out at each level by frequency

<u>Guarantee (informal)</u>: The algorithm generates a T-compatible factorization, ensuring that the factorization can be used to generate an NL explanation.

SUMMARIZATION

Two levels of summarization:

```
[Duke]·
([Jun Y.]·
 ([VLDB]·
 ([2006]·[Scalable...]
  + [2007]·[Making...]))
  + [SIGMOD]·[2014]·
 ([iCheck...] + [Cumulon...]))
+ [Brett W.]·[iCheck...]·[SIGMOD]·[2014])
```

A (outer bracket: all lines from "([Jun Y.]·" to "+ [Brett W.]·[iCheck...]·[SIGMOD]·[2014])") B (inner bracket: lines from "([VLDB]·" to "([iCheck...] + [Cumulon...]))")

Shorter Summarized Explanation Based on A:

Duke is the organization of 2 authors who published 4 papers in 2 conferences in 2006 - 2014

More Detailed Summarized Explanation Based on B:

Duke is the organization of Jun Y. who published 4 papers in 2 conferences in 2006 - 2014 and Brett W. who published 'iCheck...' in SIGMOD in 2014

SAMPLE USE-CASES

REPRESENTATIVE USE-CASES FROM THE USER STUDY:

- Q: RETURN THE AUTHORS WHO PUBLISHED PAPERS IN VLDB BEFORE 2016 AND AFTER 2007
  - A: JUN Y. PUBLISHED 9 PAPERS IN VLDB IN 2008 – 2015
- Q: RETURN THE AUTHORS WHO PUBLISHED PAPERS IN DATABASE CONFERENCES
  - A: JUN Y. PUBLISHED 64 PAPERS IN 18 CONFERENCES
- Q: RETURN THE ORGANIZATION OF AUTHORS WHO PUBLISHED PAPERS IN DATABASE CONFERENCES AFTER 2005
  - A: DUKE IS THE ORGANIZATION OF 63 AUTHORS WHO PUBLISHED 170 PAPERS IN 31 CONFERENCES IN 2006 - 2015

IN THIS TALK: WHEN RAW PROVENANCE IS NOT ENOUGH

FACTORIZING AND SUMMARIZING PROVENANCE FOR NATURAL LANGUAGE EXPLANATIONS

VLDB 16', VLDB 17', SIGMOD REC. 18', VLDB J. 20'

ABSTRACTING PROVENANCE FOR QUERY PRIVACY

SIGMOD 21', ICDE 21'

EXPLANATIONS FOR QUERY RESULTS

WHY WAS I SHOWN THIS AD?

YOUR HOBBY IS DANCE ACCORDING TO FACEBOOK AND IT WAS PUBLISHED ON REDDIT THAT YOU ARE INTERESTED IN MUSIC

YOUR HOBBY IS DANCE ACCORDING TO LINKEDIN AND YOU ARE INTERESTED IN MUSIC ACCORDING TO FACEBOOK

Provenance-Based Explanations Reveal the Query

	PID	Interest	Source
i_1	1	Music	Reddit
i_2	2	Music	Facebook
i_3	3	Music	LinkedIn
i			Reddit
i			Facebook
i			Reddit

	PID	Hobby	Source
h_1	1	Dance	Facebook
h_2	2	Dance	LinkedIn
h_3	4	Dance	Facebook
h_4	1	Trips	Facebook
h_5	2	Trips	LinkedIn
h_6	3	Trips	Reddit

	PD	Name	Age
p_1	1	James T	27
p_2	2	Brenda P	31

The general proprietary criterion for showing the ad

Your hobby is dance according to Facebook and it was published on Reddit that you are interested in music

Your hobby is dance according to LinkedIn and you are interested in music according to Facebook

Deutch, G., "Reverse-Engineering Conjunctive Queries from Provenance Examples". In EDBT 2019, pp. 277-288

PRIVACY-PRESERVING EXPLANATIONS

	PID	Interest	Source
i_1	1	Music	Reddit
i_2	2	Music	Facebook
i_3	3	Music	LinkedIn
i_4	1	Parties	Reddit
i_5	2	Parties	Facebook
i_6	4	Movies	Reddit

	PID	Hobby	Source
h_1	1	Dance	Facebook
h_2	2	Dance	LinkedIn
h_3	4	Dance	Facebook
h_4	1	Trips	Facebook
h_5	2	Trips	LinkedIn
h_6	3	Trips	Reddit

	PD	Name	Age
p_1	1	James T	27
p_2	2	Brenda P	31

?

SOME INFORMATION FROM FACEBOOK AND IT WAS PUBLISHED ON REDDIT THAT YOU ARE INTERESTED IN MUSIC

SOME INFORMATION FROM LINKEDIN AND YOU ARE INTERESTED IN MUSIC ACCORDING TO FACEBOOK

SPJU QUERIES

	PID	Interest	Source
i_1	1	Music	Reddit
i_2	2	Music	Facebook
i_3	3	Music	LinkedIn
i_4	1	Parties	Reddit
i_5	2	Parties	Facebook
i_6	4	Movies	Reddit

	PID	Hobby	Source
h_1	1	Dance	Facebook
h_2	2	Dance	LinkedIn
h_3	4	Dance	Facebook
h_4	1	Trips	Facebook
h_5	2	Trips	LinkedIn
h_6	3	Trips	Reddit

	PD	Name	Age
p_1	1	James T	27
p_2	2	Brenda P	31

Q(id):-Person(id,name,age), Hobbies(id,'Dance',src1), Interests(id,'Music',src2)

RETURN THE ID OF A PERSON WHOSE HOBBY IS `DANCE' AND WHOSE INTEREST IS `MUSIC'

PROVENANCE MODEL

	PID	Interest	Source
i_1	1	Music	Reddit
i_2	2	Music	Facebook
i_3	3	Music	LinkedIn
i_4	1	Parties	Reddit
i_5	2	Parties	Facebook
i_6	4	Movies	Reddit

	PID	Hobby	Source
h_1	1	Dance	Facebook
h_2	2	Dance	LinkedIn
h_3	4	Dance	Facebook
h_4	1	Trips	Facebook
h_5	2	Trips	LinkedIn
h_6	3	Trips	Reddit

	PD	Name	Age
p_1	1	James T	27
p_2	2	Brenda P	31

Q(1):-Person(1,James T,27), Hobbies(1,'Dance',Facebook),
Interests(1,'Music',Reddit)

Output: 1

Provenance: $p_1 \cdot i_1 \cdot h_1$

Green, Karvounarakis, Tannen, "Provenance Semirings". PODS: pp. 31-40, 2007

PROVENANCE EXAMPLE FOR SPJU QUERY RESULTS

	PID	Interest	Source
i_1	1	Music	Reddit
i_2	2	Music	Facebook
i_3	3	Music	LinkedIn
i_4	1	Parties	Reddit
i_5	2	Parties	Facebook
i_6	4	Movies	Reddit

	PID	Hobby	Source
h_1	1	Dance	Facebook
h_2	2	Dance	LinkedIn
h_3	4	Dance	Facebook
h_4	1	Trips	Facebook
h_5	2	Trips	LinkedIn
h_6	3	Trips	Reddit

	PD	Name	Age
p_1	1	James T	27
p_2	2	Brenda P	31

PROVENANCE EXAMPLE WITH TWO TUPLES

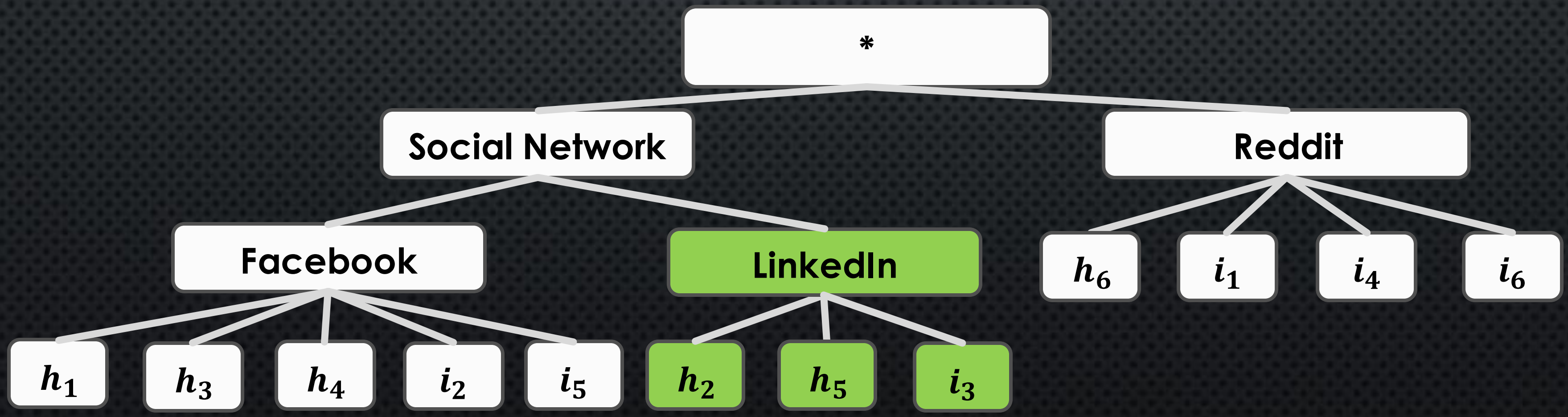
Output	Provenance
1	$p_1 \cdot i_1 \cdot h_1$
2	$p_2 \cdot i_2 \cdot h_2$

PROVENANCE ABSTRACTION

	PID	Interest	Source
i_1	1	Music	Reddit
i_2	2	Music	Facebook
i_3	3	Music	LinkedIn
i_4	1	Parties	Reddit
i_5	2	Parties	Facebook
i_6	4	Movies	Reddit

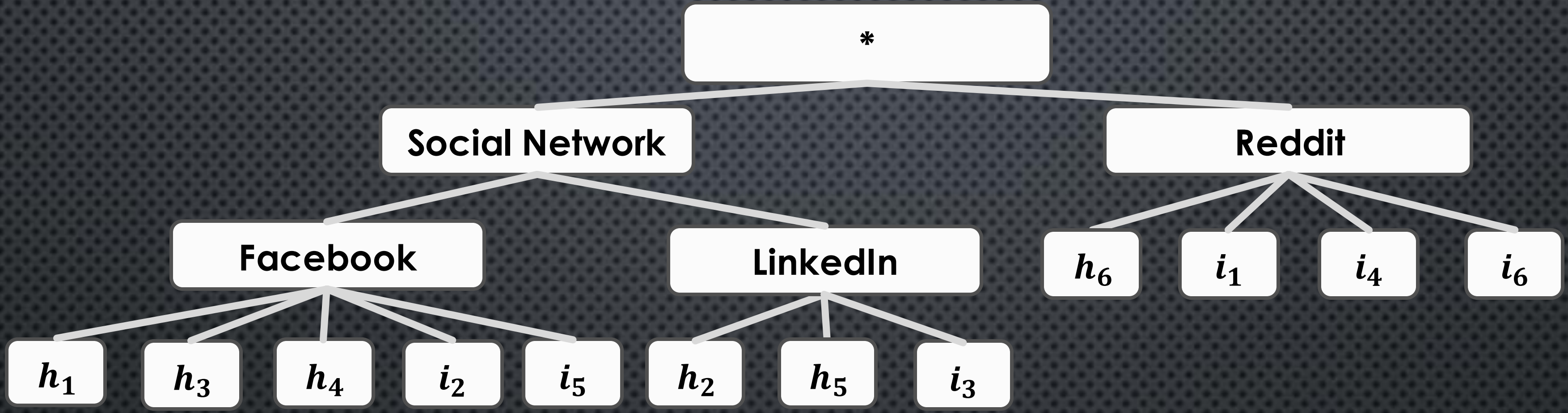
	PID	Hobby	Source
h_1	1	Dance	Facebook
h_2	2	Dance	LinkedIn
h_3	4	Dance	Facebook
h_4	1	Trips	Facebook
h_5	2	Trips	LinkedIn
h_6	3	Trips	Reddit

	PD	Name	Age
p_1	1	James T	27
p_2	2	Brenda P	31



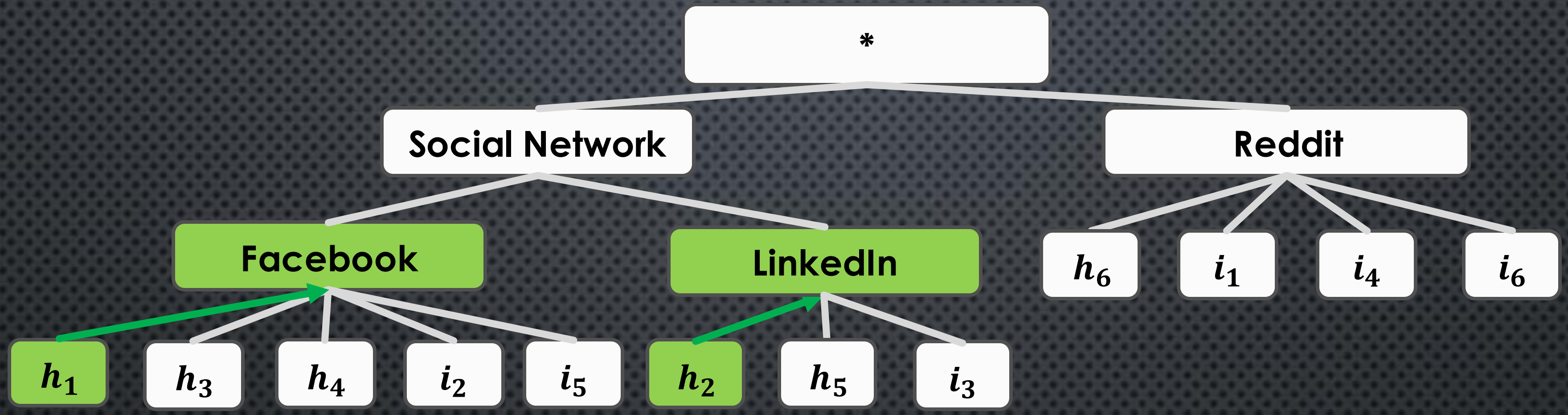
Deutch, Moskovitch, Rinetzky, "Hypothetical Reasoning via Provenance Abstraction". SIGMOD: pp. 537-554, 2019

PROVENANCE ABSTRACTION



Output	Provenance
1	$p_1 \cdot i_1 \cdot h_1$
2	$p_2 \cdot i_2 \cdot h_2$

PROVENANCE ABSTRACTION



Output	Provenance
1	$p_1 \cdot i_1 \cdot h_1$
2	$p_2 \cdot i_2 \cdot h_2$

→

Output	Provenance
1	$p_1 \cdot i_1 \cdot \boldsymbol{Facebook}$
2	$p_2 \cdot i_2 \cdot \boldsymbol{LinkedIn}$

THE PRIVACY OF AN ABSTRACT EXAMPLE USING K-ANONYMITY

Output	Provenance
1	$p_1 \cdot i_1 \cdot Facebook$
2	$p_2 \cdot i_2 \cdot LinkedIn$

ALL QUERIES WILL GENERATE THE PROVENANCE

Q(ID) :-PERSON(ID,NAME,AGE), HOBBIES(ID,'DANCE',SRC1), INTERESTS(ID,'MUSIC',SRC2)

$Q1$(ID) :- PERSON(ID,NAME,AGE), HOBBIES(ID,'TRIPS',SRC1), INTERESTS(ID,'MUSIC',SRC2)

$Q2$(ID) :- PERSON(ID,NAME,AGE), HOBBIES(ID,'DANCE',SRC1), INTERESTS(ID,'PARTIES',SRC2)

“GOOD” QUERIES FOR AN ABSTRACTED PROVENANCE EXAMPLE

Output	Provenance
1	$p_1 \cdot i_1 \cdot Facebook$
2	$p_2 \cdot i_2 \cdot LinkedIn$

Q(ID):-PERSON(ID,NAME,AGE), HOBBIES(ID,‘DANCE’,SRC1), INTERESTS(ID,‘MUSIC’,SRC2)

"GOOD" QUERIES FOR AN ABSTRACTED PROVENANCE EXAMPLE

Output	Provenance
1	$p_1 \cdot i_1 \cdot Facebook$
2	$p_2 \cdot i_2 \cdot LinkedIn$

Q(ID):-PERSON(**ID**,NAME,AGE), HOBBIES(**ID**,'DANCE',SRC1), INTERESTS(**ID**,'MUSIC',SRC2)

- CONNECTED

"GOOD" QUERIES FOR AN ABSTRACTED PROVENANCE EXAMPLE

Output	Provenance
1	$p_1 \cdot i_1 \cdot Facebook$
2	$p_2 \cdot i_2 \cdot LinkedIn$

Q(1):-PERSON(1,JAMES T,27), HOBBIES(1,'DANCE',FACEBOOK), INTERESTS(1,'MUSIC',REDDIT)

- CONNECTED
- CONSISTENT - GENERATES THE DESIRED PROVENANCE FOR EACH OF THE RESULTS IN ONE OF THE CONCRETE OPTIONS

"GOOD" QUERIES FOR AN ABSTRACTED PROVENANCE EXAMPLE

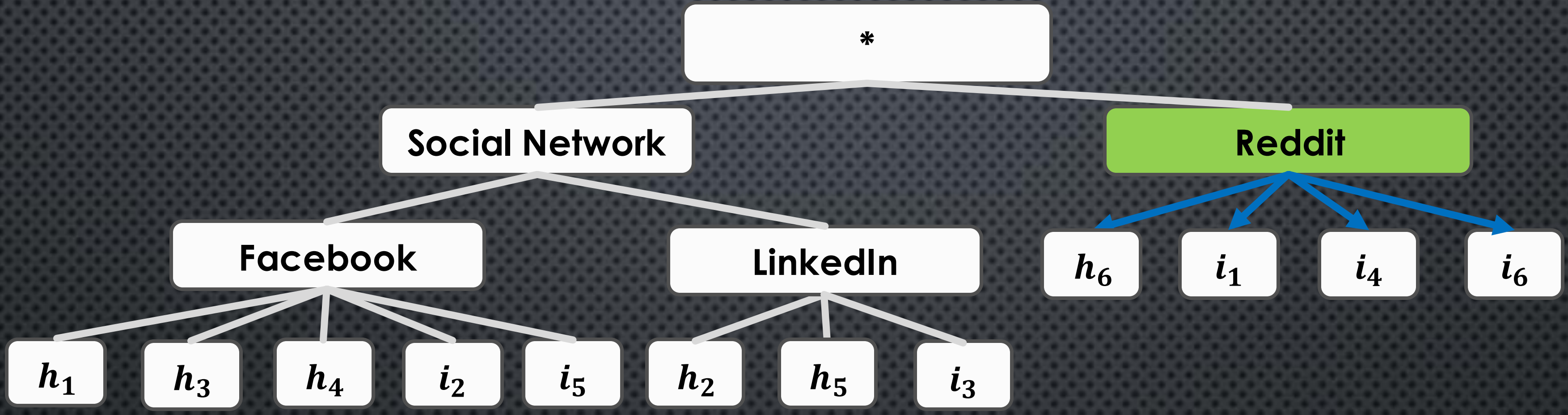
Output	Provenance
1	$p_1 \cdot i_1 \cdot Facebook$
2	$p_2 \cdot i_2 \cdot LinkedIn$

Q(ID):-PERSON(ID,NAME,AGE), HOBBIES(ID,'DANCE',SRC1), INTERESTS(ID,'MUSIC',SRC2)

IF WE HAVE K SUCH CANDIDATE QUERIES, WE SAY THAT THE ABSTRACTION HAS PRIVACY K

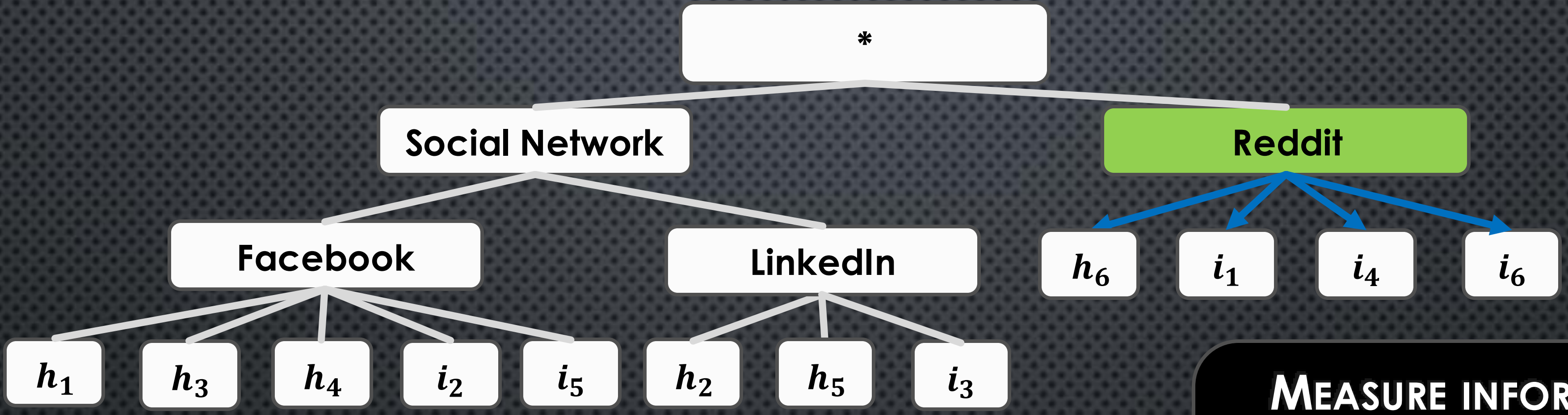
- CONNECTED
- CONSISTENT - GENERATES THE DESIRED PROVENANCE FOR EACH OF THE RESULTS IN ONE OF THE CONCRETE OPTIONS
- INCLUSION MINIMAL – NO OTHER CONSISTENT QUERY IS CONTAINED IN IT
  - Deutch, G., "Reverse-Engineering Conjunctive Queries from Provenance Examples". In EDBT 2019, pp. 277-288

LOSS OF INFORMATION INCURRED BY PROVENANCE ABSTRACTION



Output	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{Reddit}$
2	$p_2 \cdot i_2 \cdot h_2$

LOSS OF INFORMATION INCURRED BY PROVENANCE ABSTRACTION



Measure information loss with Entropy =

$$-\sum_i P_X(x_i) ln(P_X(x_i)).$$

Output	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{Reddit}$
2	$p_2 \cdot i_2 \cdot h_2$

Outpu t	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{h_6}$
2	$p_2 \cdot i_2 \cdot h_2$

Outpu t	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{i_1}$
2	$p_2 \cdot i_2 \cdot h_2$

Outpu t	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{i_4}$
2	$p_2 \cdot i_2 \cdot h_2$

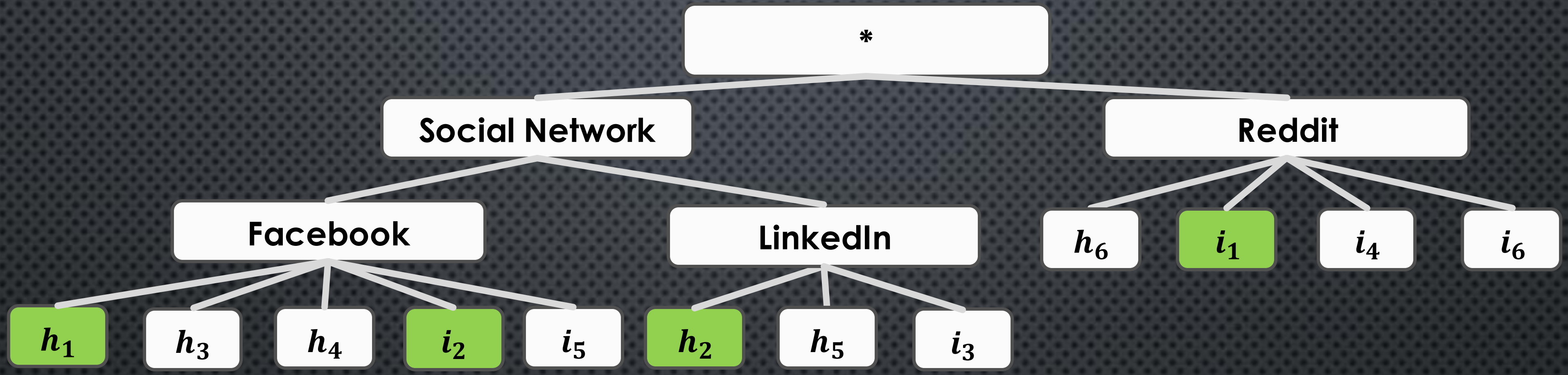
Outpu t	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{i_6}$
2	$p_2 \cdot i_2 \cdot h_2$

THE OPTIMAL PROVENANCE ABSTRACTION PROBLEM

Problem definition: Given an abstraction tree, a provenance example, and a privacy threshold k, find an abstraction for the example that achieves privacy ≥ k and incurs the minimum loss of information over all abstractions that achieve the privacy threshold k.

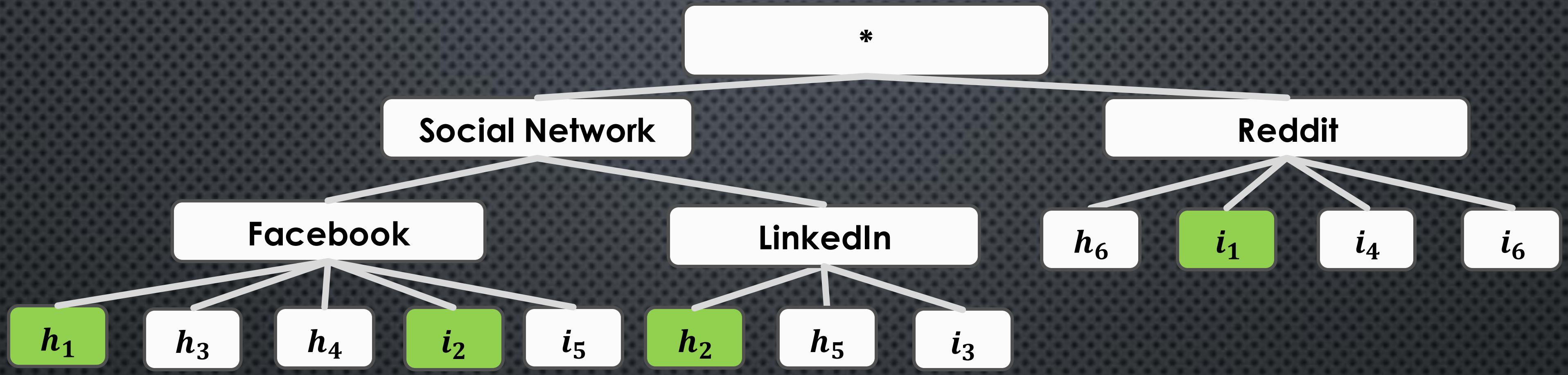
Proposition: The decision version of the optimal abstraction problem is **NP-hard**.

ABSTRACTION COMPUTATION



Output	Provenance
1	$p_1 \cdot i_1 \cdot h_1$
2	$p_2 \cdot i_2 \cdot h_2$

ABSTRACTION COMPUTATION



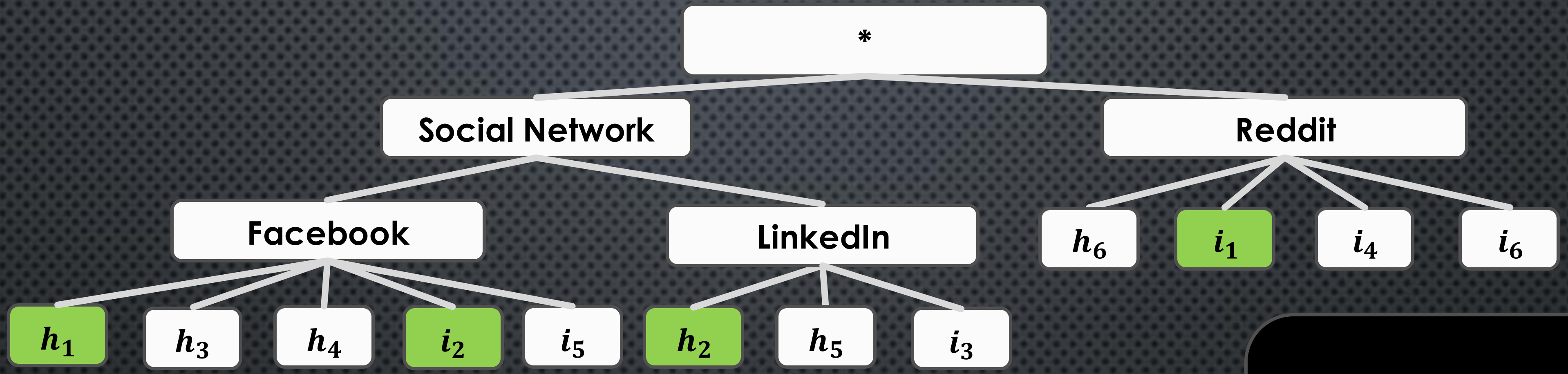
Output	Provenance
1	$p_1 \cdot i_1 \cdot h_1$
2	$p_2 \cdot i_2 \cdot h_2$

Output	Provenance
1	$p_1 \cdot i_1 \cdot \boldsymbol{Facebook}$
2	$p_2 \cdot i_2 \cdot \boldsymbol{LinkedIn}$

Output	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{Reddit}$
2	$p_2 \cdot h_2 \cdot \boldsymbol{Facebook}$

Output	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{Reddit}$
2	$p_2 \cdot h_2 \cdot i_2$

ABSTRACTION COMPUTATION



Output	Provenance
1	$p_1 \cdot i_1 \cdot h_1$
2	$p_2 \cdot i_2 \cdot h_2$

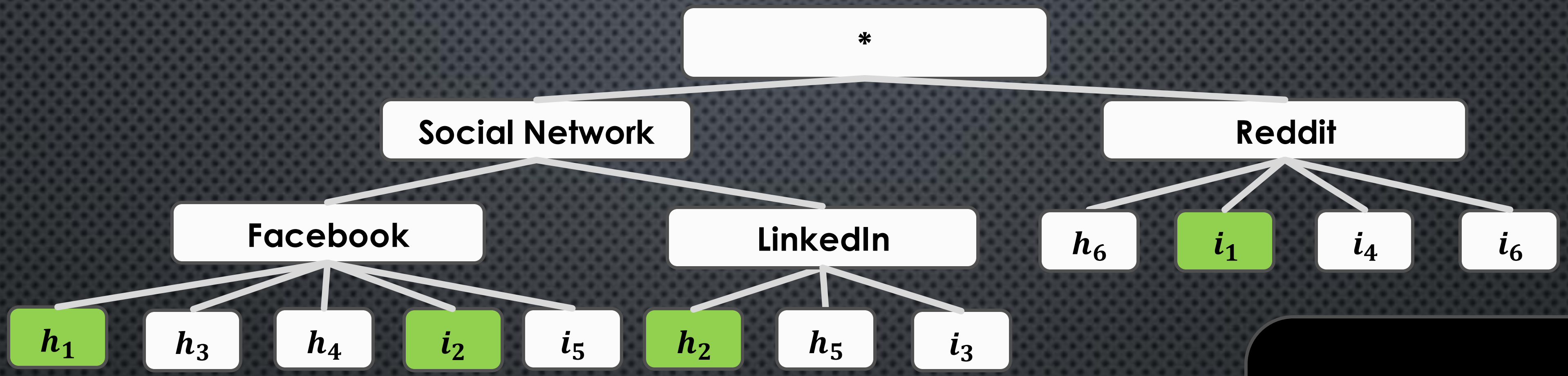
DOES NOT MEET THE PRIVACY THRESHOLD (K=2)

Output	Provenance
1	$p_1 \cdot i_1 \cdot \boldsymbol{Facebook}$
2	$p_2 \cdot i_2 \cdot \boldsymbol{LinkedIn}$

Output	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{Reddit}$
2	$p_2 \cdot h_2 \cdot \boldsymbol{Facebook}$

Output	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{Reddit}$
2	$p_2 \cdot h_2 \cdot i_2$

ABSTRACTION COMPUTATION



Output	Provenance
1	$p_1 \cdot i_1 \cdot h_1$
2	$p_2 \cdot i_2 \cdot h_2$

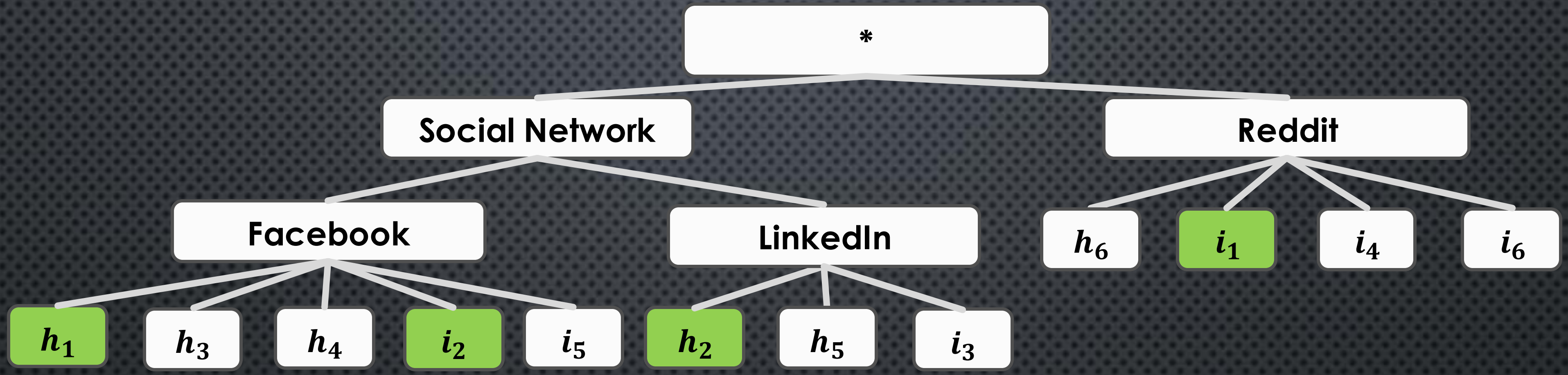
DOES NOT MINIMIZE THE LOSS OF INFORMATION

Output	Provenance
1	$p_1 \cdot i_1 \cdot \boldsymbol{Facebook}$
2	$p_2 \cdot i_2 \cdot \boldsymbol{LinkedIn}$

Output	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{Reddit}$
2	$p_2 \cdot h_2 \cdot \boldsymbol{Facebook}$

Output	Provenance
1	$p_1 \cdot h_1 \cdot \boldsymbol{Reddit}$
2	$p_2 \cdot h_2 \cdot i_2$

ABSTRACTION COMPUTATION



Output	Provenance
1	$p_1 \cdot i_1 \cdot h_1$
2	$p_2 \cdot i_2 \cdot h_2$

2	$p_2 \cdot i_2 \cdot \textbf{LinkedIn}$

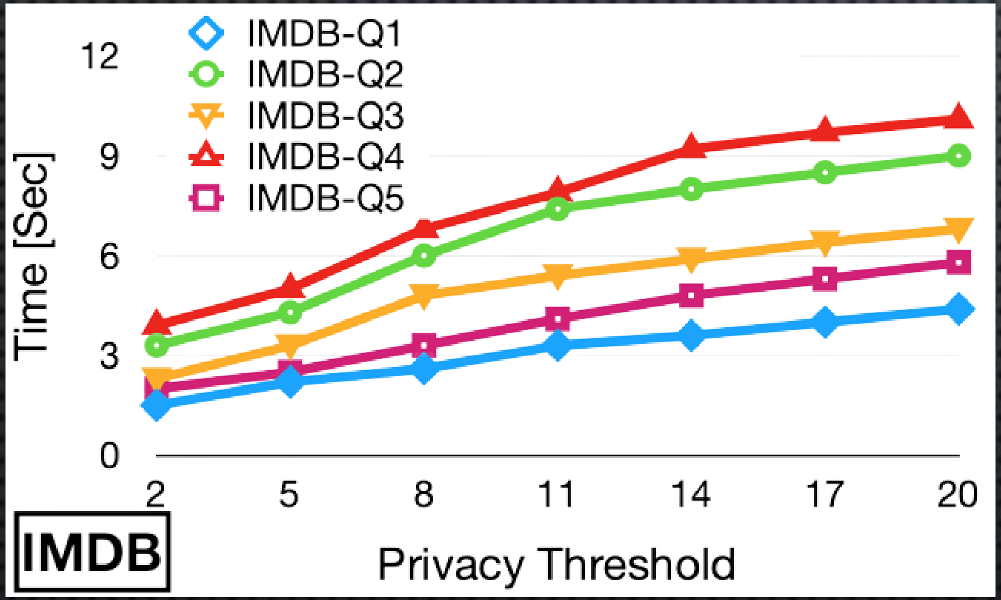
2	$p_2 \cdot h_2 \cdot \textbf{Facebook}$

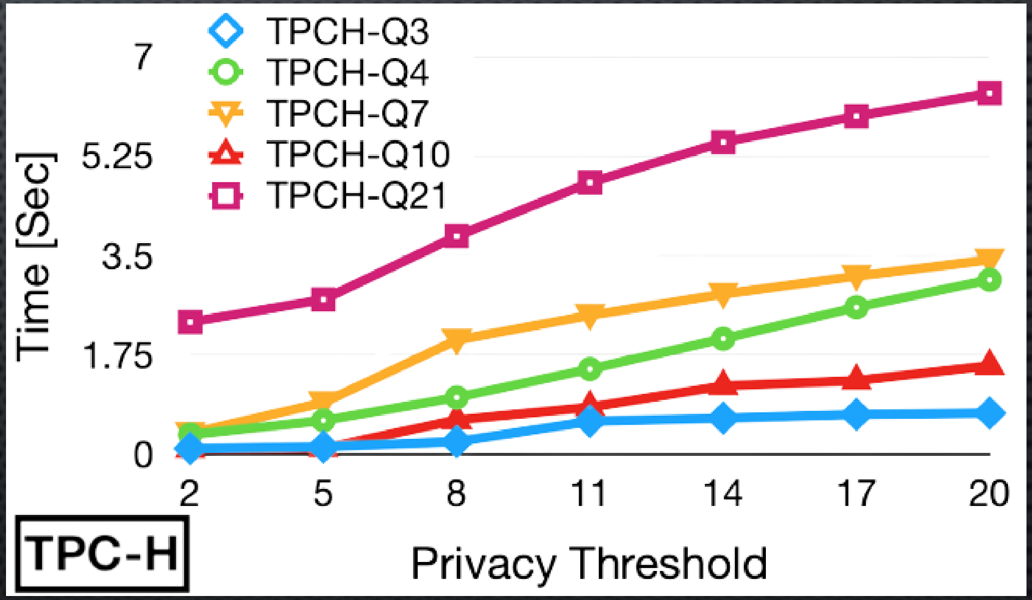
2	$p_2 \cdot h_2 \cdot i_2$

GUARANTEE (INFORMAL): THE ALGORITHM FIND AN OPTIMAL ABSTRACTION.

SAMPLE EXPERIMENTAL RESULTS

Runtime as a function of the privacy threshold





Takeaways

1. There are different ways to manipulate raw provenance, including:
   I. Factorization and summarization
   II. Abstraction
2. <u>Factorization and summarization</u> can help make provenance understandable and "easier to digest" for creating explanations
3. <u>Abstraction</u> can help preserve the privacy of the query while providing explanations
4. <u>Tradeoff:</u> smaller factorization/higher privacy threshold = less informative explanations