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

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

Results

Why?

Duke University (Duke) ...

Duke is the organization of 63 authors who published 170 papers in 31 conferences in 2006 - 2015
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
**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:**
(pname, Duke) (aname, Jun Y.) (ptitle, iCheck...) (cname, SIGMOD) (pyear, 14')+ (pname, Duke) (aname, Jun Y.) (ptitle, Scalable...) (cname, VLDB) (pyear, 06')+ (pname, Duke) (aname, Jun Y.) (ptitle, Making...) (cname, VLDB) (pyear, 07')+ (pname, Duke) (aname, Brett W.) (ptitle, iCheck...) (cname, SIGMOD) (pyear, 14')+ (pname, Duke) (aname, Jun Y.) (ptitle, Cumulon...) (cname, SIGMOD) (pyear, 14')+ ...
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!
1. Li, F., Jagadish, H. V., “Constructing an Interactive Natural Language Interface for Relational Databases”. In: Proc. VLDB Endow. (2014), pp. 73–84
FROM MAPPINGS TO AN EXPLANATION

Return

organization
POS=NN, REL=dobj

of
POS=IN, REL=prep

the

authors
POS=NNS, REL=pobj

published
POS=VBD, REL=rcmod

who

papers

after
POS=IN, REL=prep

in

conferences
POS=NNS, REL=pobj

database
POS=NN, REL=nn

2005
POS=CD, REL=pobj

(ptitle, "iCheck...")

(pyyear, 2014)

(aname, Duke)

(aname, Jun Y.)

(cname, SIGMOD)
Return of organization POS=NN, REL=dobj
the of POS=IN, REL=prep
organization POS=NN, REL=dobj
authors POS=NNS, REL=pobj
published POS=VBD, REL=rcmod
who papers after POS=IN, REL=prep
2005 POS=CD, REL=pobj
conferences POS=NNS, REL=pobj
database POS=NN, REL=nn
(Return of organization (oname, Duke) of the of organization (oname, Jun Y.) published in 2005 conferences at SIGMOD)
Return

organization
POS=NN, REL=dobj

of
POS=IN, REL=prep

authors
POS=NNS, REL=pobj

published
POS=VBD, REL=rcmod

who

papers

after
POS=IN, REL=prep

2005
POS=CD, REL=pobj

conferences
POS=NNS, REL=pobj

database
POS=NN, REL=nn

Return

organization

Duke (is the)

of

Jun Y.

published

who

“iCheck”

in

2014

SIGMOD

county

SIGMOD

(cntry, SIGMOD)
Return of organization POS=NN, REL=dobj

of POS=IN, REL=prep

organization POS=NN, REL=nsubj

the POS=DT, REL=det

of POS=IN, REL=prep

authors POS=NNS, REL=pobj

published POS=VBD, REL=rcmod

papers POS=NNS, REL=pobj

who POS=PRON, REL=nsubj

after POS=IN, REL=prep

2005 POS=CD, REL=dt

conferences POS=NNS, REL=pobj

database POS=NN, REL=nn

Return

() organization

Duke (is the)

of

Jun Y.

published

who

“iCheck”

in

2014 SIGMOD

(database, SIGMOD)

(ptitle, "iCheck...")

(pyyear, 2014)

(name, Jun Y.)

(organization, Duke)
Duke is the organization of Jun Y. who published ‘iCheck...’ in SIGMOD in 2014.
**IDEA:** Use algebraic factorization to take out common values that appear in multiple assignments

\[
\text{Duke}\cdot\text{Jun Y.}\cdot\text{iCheck...}\cdot\text{SIGMOD}\cdot[2014]+ \\
\text{Duke}\cdot\text{Jun Y.}\cdot\text{Scalable...}\cdot\text{VLDB}\cdot[2006]+ \\
\text{Duke}\cdot\text{Jun Y.}\cdot\text{Making...}\cdot\text{VLDB}\cdot[2007]+ \\
\text{Duke}\cdot\text{Brett W.}\cdot\text{iCheck...}\cdot\text{SIGMOD}\cdot[2014]+ \\
\text{Duke}\cdot\text{Jun Y.}\cdot\text{Cumulon...}\cdot\text{SIGMOD}\cdot[2014]
\]

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

**Shortest Factorization:**

\[
\text{Duke}. \\
\text{[SIGMOD]\cdot[2014].} \\
\text{[iCheck...].} \\
\text{([Jun Y.] + [Brett W.])} \\
\text{+ [Jun Y.]\cdot[Cumulon...]} \\
\text{+ [VLDB]\cdot[Jun Y.]} \\
\text{([2006]\cdot[Scalable...]} \\
\text{+ [2007]\cdot[Making...]} \\
\text{+ [SIGMOD]\cdot[2014].} \\
\text{([iCheck... + [Cumulon...])} \\
\text{+ [Brett W.]\cdot[iCheck...]\cdot[SIGMOD]\cdot[2014]}
\]
T-COMPATIBILITY

**NL Query:**
Return the organization of authors who published papers in database conferences after 2005

**Shortest Factorization:**

**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.
[Duke].
([SIGMOD].[2014].
([iCheck...].
([Jun Y.] + [Brett W.]))
+ [Jun Y.].[Cumulon...])
+ [VLDB].[Jun Y.]·
([2006].[Scalable...])
+ [2007].[Making...])
FINDING T-COMPATIBLE FACTORIZATION:

Algorithm:

- 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

Guarantee (Informal): The algorithm generates a T-compatible factorization, ensuring that the factorization can be used to generate an NL explanation.
SUMMARIZATION

TWO LEVELS OF SUMMARIZATION:

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
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.
**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**

---

**The general proprietary criterion for showing the ad**

<table>
<thead>
<tr>
<th>PID</th>
<th>Interest</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i_1$</td>
<td>1</td>
<td>Reddit</td>
</tr>
<tr>
<td>$i_2$</td>
<td>2</td>
<td>Facebook</td>
</tr>
<tr>
<td>$i_3$</td>
<td>3</td>
<td>LinkedIn</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PID</th>
<th>Hobby</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>Dance</td>
<td>Facebook</td>
</tr>
<tr>
<td>$h_2$</td>
<td>Dance</td>
<td>LinkedIn</td>
</tr>
<tr>
<td>$h_3$</td>
<td>Dance</td>
<td>Facebook</td>
</tr>
<tr>
<td>$h_4$</td>
<td>Trips</td>
<td>Facebook</td>
</tr>
<tr>
<td>$h_5$</td>
<td>Trips</td>
<td>LinkedIn</td>
</tr>
<tr>
<td>$h_6$</td>
<td>Trips</td>
<td>Reddit</td>
</tr>
</tbody>
</table>

---

Deutch, G., “Reverse-Engineering Conjunctive Queries from Provenance Examples”. In EDBT 2019, pp. 277-288
### PRIVACY-PRESERVING EXPLANATIONS

<table>
<thead>
<tr>
<th>PID</th>
<th>Interest</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i_1$</td>
<td>Music</td>
<td>Reddit</td>
</tr>
<tr>
<td>$i_2$</td>
<td>Music</td>
<td>Facebook</td>
</tr>
<tr>
<td>$i_3$</td>
<td>Music</td>
<td>LinkedIn</td>
</tr>
<tr>
<td>$i_4$</td>
<td>Parties</td>
<td>Reddit</td>
</tr>
<tr>
<td>$i_5$</td>
<td>Parties</td>
<td>Facebook</td>
</tr>
<tr>
<td>$i_6$</td>
<td>Movies</td>
<td>Reddit</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PID</th>
<th>Hobby</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>Dance</td>
<td>Facebook</td>
</tr>
<tr>
<td>$h_2$</td>
<td>Dance</td>
<td>LinkedIn</td>
</tr>
<tr>
<td>$h_3$</td>
<td>Dance</td>
<td>Facebook</td>
</tr>
<tr>
<td>$h_4$</td>
<td>Trips</td>
<td>Facebook</td>
</tr>
<tr>
<td>$h_5$</td>
<td>Trips</td>
<td>LinkedIn</td>
</tr>
<tr>
<td>$h_6$</td>
<td>Trips</td>
<td>Reddit</td>
</tr>
</tbody>
</table>

### 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.

---

**WOOHOO!!!**
SPJU QUERIES

<table>
<thead>
<tr>
<th>PID</th>
<th>Interest</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>i₁</td>
<td>Music</td>
<td>Reddit</td>
</tr>
<tr>
<td>i₂</td>
<td>Music</td>
<td>Facebook</td>
</tr>
<tr>
<td>i₃</td>
<td>Music</td>
<td>LinkedIn</td>
</tr>
<tr>
<td>i₄</td>
<td>Parties</td>
<td>Reddit</td>
</tr>
<tr>
<td>i₅</td>
<td>Parties</td>
<td>Facebook</td>
</tr>
<tr>
<td>i₆</td>
<td>Movies</td>
<td>Reddit</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PID</th>
<th>Hobby</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>h₁</td>
<td>Dance</td>
<td>Facebook</td>
</tr>
<tr>
<td>h₂</td>
<td>Dance</td>
<td>LinkedIn</td>
</tr>
<tr>
<td>h₃</td>
<td>Dance</td>
<td>Facebook</td>
</tr>
<tr>
<td>h₄</td>
<td>Trips</td>
<td>Facebook</td>
</tr>
<tr>
<td>h₅</td>
<td>Trips</td>
<td>LinkedIn</td>
</tr>
<tr>
<td>h₆</td>
<td>Trips</td>
<td>Reddit</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PD</th>
<th>Name</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>p₁</td>
<td>James T</td>
<td>27</td>
</tr>
<tr>
<td>p₂</td>
<td>Brenda P</td>
<td>31</td>
</tr>
</tbody>
</table>

\( Q(\text{id}) : \neg \text{Person}(\text{id}, \text{name}, \text{age}), \text{Hobbies}(\text{id}, '\text{Dance}', \text{src1}), \text{Interests}(\text{id}, '\text{Music}', \text{src2}) \)

**Return the ID of a person whose hobby is `Dance' and whose interest is `Music'**
Q(1): $\text{Person}(1, \text{James T, 27}), \text{Hobbies}(1, \text{‘ Dance’, Facebook}), \text{Interests}(1, \text{‘ Music’, Reddit})$

Output: 1

Provenance: $p_1 \cdot i_1 \cdot h_1$

### PROVENANCE EXAMPLE FOR SPJU QUERY RESULTS

#### Provenance Example with Two Tuples

<table>
<thead>
<tr>
<th>PID</th>
<th>Interest</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i_1$</td>
<td>1</td>
<td>Music</td>
</tr>
<tr>
<td>$i_2$</td>
<td>2</td>
<td>Music</td>
</tr>
<tr>
<td>$i_3$</td>
<td>3</td>
<td>Music</td>
</tr>
<tr>
<td>$i_4$</td>
<td>1</td>
<td>Parties</td>
</tr>
<tr>
<td>$i_5$</td>
<td>2</td>
<td>Parties</td>
</tr>
<tr>
<td>$i_6$</td>
<td>4</td>
<td>Movies</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PID</th>
<th>Hobby</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>1</td>
<td>Dance</td>
</tr>
<tr>
<td>$h_2$</td>
<td>2</td>
<td>Dance</td>
</tr>
<tr>
<td>$h_3$</td>
<td>4</td>
<td>Dance</td>
</tr>
<tr>
<td>$h_4$</td>
<td>1</td>
<td>Trips</td>
</tr>
<tr>
<td>$h_5$</td>
<td>2</td>
<td>Trips</td>
</tr>
<tr>
<td>$h_6$</td>
<td>3</td>
<td>Trips</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PD</th>
<th>Name</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_1$</td>
<td>James T</td>
<td>27</td>
</tr>
<tr>
<td>$p_2$</td>
<td>Brenda P</td>
<td>31</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_1 \cdot i_1 \cdot h_1$</td>
</tr>
<tr>
<td>2</td>
<td>$p_2 \cdot i_2 \cdot h_2$</td>
</tr>
</tbody>
</table>
PROVENANCE ABSTRACTION

Output | Provenance
-------|------------
1       | $p_1 \cdot i_1 \cdot h_1$
2       | $p_2 \cdot i_2 \cdot h_2$
PROVENANCE ABSTRACTION

Social Network

Facebook

LinkedIn

Reddit

*h* · *i* · *h*

Output | Provenance
--- | ---
1 | \( p_1 \cdot i_1 \cdot h_1 \)
2 | \( p_2 \cdot i_2 \cdot h_2 \)

Output | Provenance
--- | ---
1 | \( p_1 \cdot i_1 \cdot \text{Facebook} \)
2 | \( p_2 \cdot i_2 \cdot \text{LinkedIn} \)
THE PRIVACY OF AN ABSTRACT EXAMPLE USING K-ANONYMITY

<table>
<thead>
<tr>
<th>Output</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_1 \cdot i_1 \cdot \text{Facebook}$</td>
</tr>
<tr>
<td>2</td>
<td>$p_2 \cdot i_2 \cdot \text{LinkedIn}$</td>
</tr>
</tbody>
</table>

$Q(id) :- \text{Person}(id, \text{name}, \text{age}), \text{Hobbies}(id, \text{Dance}', src1), \text{Interests}(id, \text{Music}', src2)$

$Q1(id) :- \text{Person}(id, \text{name}, \text{age}), \text{Hobbies}(id, \text{Trips}', src1), \text{Interests}(id, \text{Music}', src2)$

$Q2(id) :- \text{Person}(id, \text{name}, \text{age}), \text{Hobbies}(id, \text{Dance}', src1), \text{Interests}(id, \text{Parties}', src2)$
“GOOD” QUERIES FOR AN ABSTRACTED PROVENANCE EXAMPLE

<table>
<thead>
<tr>
<th>Output</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( p_1 \cdot i_1 \cdot Facebook )</td>
</tr>
<tr>
<td>2</td>
<td>( p_2 \cdot i_2 \cdot LinkedIn )</td>
</tr>
</tbody>
</table>

\[ Q(\text{id}) : \neg \text{PERSON(\text{id, name, age})}, \ \text{Hobbies(\text{id, ‘Dance’, src1})}, \ \text{Interests(\text{id, ‘Music’, src2})} \]
"GOOD" QUERIES FOR AN ABSTRACTED PROVENANCE EXAMPLE

<table>
<thead>
<tr>
<th>Output</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_1 \cdot i_1 \cdot Facebook$</td>
</tr>
<tr>
<td>2</td>
<td>$p_2 \cdot i_2 \cdot LinkedIn$</td>
</tr>
</tbody>
</table>

$Q(\text{ID}): \neg \text{PERSON(} \text{ID}, \text{NAME, AGE}\text{)}, \ \text{HOBBIES(} \text{ID}, \text{‘DANCE’, src1}\text{)}, \ \text{INTERESTS(} \text{ID}, \text{‘Music’, src2}\text{)}$

- Connected
“GOOD” QUERIES FOR AN ABSTRACTED PROVENANCE EXAMPLE

<table>
<thead>
<tr>
<th>Output</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_1 \cdot i_1 \cdot Facebook$</td>
</tr>
<tr>
<td>2</td>
<td>$p_2 \cdot i_2 \cdot LinkedIn$</td>
</tr>
</tbody>
</table>

$Q(1): \text{Person(1, James T, 27), Hobbies(1, ‘Dance’, Facebook), Interests(1, ‘Music’, Reddit)}$

- **Connected**
- **Consistent -** generates the desired provenance for each of the results in one of the concrete options
“GOOD” QUERIES FOR AN ABSTRACTED PROVENANCE EXAMPLE

\[
Q(\text{id}) : -\text{PERSON}(\text{id}, \text{name}, \text{age}), \ \text{Hobbies}(\text{id}, \text{‘Dance’}, \text{src1}), \ \text{Interests}(\text{id}, \text{‘Music’}, \text{src2})
\]

<table>
<thead>
<tr>
<th>Output</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(p_1 \cdot i_1 \cdot \text{Facebook})</td>
</tr>
<tr>
<td>2</td>
<td>(p_2 \cdot i_2 \cdot \text{LinkedIn})</td>
</tr>
</tbody>
</table>

- **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      | \( p_1 \cdot h_1 \cdot \text{Reddit} \)
2      | \( p_2 \cdot i_2 \cdot h_2 \)
LOSS OF INFORMATION INCURRED BY PROVENANCE ABSTRACTION

Measure information loss with Entropy = 
\[- \sum_i P_X(x_i) \ln(P_X(x_i)) \].

<table>
<thead>
<tr>
<th>Output</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( p_1 \cdot h_1 \cdot Reddit )</td>
</tr>
<tr>
<td>2</td>
<td>( p_2 \cdot i_2 \cdot h_2 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( p_1 \cdot h_1 \cdot i_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( p_2 \cdot i_2 \cdot h_2 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( p_1 \cdot h_1 \cdot i_4 )</td>
</tr>
<tr>
<td>2</td>
<td>( p_2 \cdot i_2 \cdot h_2 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( p_1 \cdot h_1 \cdot i_6 )</td>
</tr>
<tr>
<td>2</td>
<td>( p_2 \cdot i_2 \cdot h_2 )</td>
</tr>
</tbody>
</table>
**Problem definition:** Given an abstraction tree, a provenance example, and a privacy threshold \( k \), find an abstraction for the example that achieves privacy \( \geq k \) 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 \textbf{NP-hard}. 
**ABSTRACTION COMPUTATION**

![Diagram of abstraction computation]

<table>
<thead>
<tr>
<th>Output</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$p_1 \cdot i_1 \cdot h_1$</td>
</tr>
<tr>
<td>2</td>
<td>$p_2 \cdot i_2 \cdot h_2$</td>
</tr>
</tbody>
</table>
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 | \( p_1 \cdot h_1 \cdot Reddit \)
2 | \( p_2 \cdot h_2 \cdot Facebook \)

Output | Provenance
---|---
1 | \( p_1 \cdot h_1 \cdot LinkedIn \)
2 | \( p_2 \cdot h_2 \cdot i_2 \)
ABSTRACTION COMPUTATION

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

Output | Provenance
---|---
1 | $p_1 \cdot h_1 \cdot Reddit$
2 | $p_2 \cdot h_2 \cdot Facebook$

Output | Provenance
---|---
1 | $p_1 \cdot h_1 \cdot Reddit$
2 | $p_2 \cdot h_2 \cdot i_2$

DOES NOT MEET THE PRIVACY THRESHOLD ($k=2$)
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       | $p_1 \cdot h_1 \cdot Reddit$
2       | $p_2 \cdot h_2 \cdot Facebook$

Output | Provenance
-------|------------
1       | $p_1 \cdot h_1 \cdot Reddit$
2       | $p_2 \cdot h_2 \cdot i_2$

DOES NOT MINIMIZE THE LOSS OF INFORMATION
**Guarantee (Informal):** The algorithm finds an optimal abstraction.
RUNTIME AS A FUNCTION OF THE PRIVACY THRESHOLD

SAMPLE EXPERIMENTAL RESULTS
TAKEAWAYS

1. **There are different ways to manipulate raw provenance, including:**
   
   I. **Factorization and summarization**
   
   II. **Abstraction**

2. **Factorization and summarization can help make provenance understandable and "easier to digest" for creating explanations**

3. **Abstraction can help preserve the privacy of the query while providing explanations**

4. **Tradeoff: smaller factorization/higher privacy threshold = less informative explanations**