Refining SQL Queries based on Why-Not Polynomials

Katerina Tzompanaki,
Nicole Bidoit, Melanie Herschel

Theory and Practice of Provenance

MacLean, 9/6/2016
Why-Not Questions / Provenance
Why-Not Questions / Provenance

- Trust assertion.
- Results verification.
- Data cleaning.
- Data integration.
- Data transformations debugging and repairing.
Why-Not Questions / Provenance

- Trust assertion.
- Results verification.
- Data cleaning.
- Data integration.
- Data transformations debugging and repairing.
Problem Description
Problem Setting

Input Query → Test → Missing results

- New Query
- Repair
- Debug

Final Query

no

yes
Problem Setting

Input Query → Test → Missing results

New Query → Repair

Debug

Final Query

no

yes
Problem Setting

Input Query → Test → Missing results → no → Final Query

Input Query → Test → New Query → Repair

Input Query → Test → Missing results → yes → Debug

Input Query → Test → New Query → Repair
Problem Setting

Input Query → Test → Missing results

New Query → Repair

Polynomial Explanations (TaPP14, CIKM15)

no → Final Query

yes
Problem Setting

Input Query → Test → Missing results

Test → Repair

New Query → Polynomial Explanations (TaPP14, CIKM15)

Final Query

no

yes
## Motivating Example

<table>
<thead>
<tr>
<th>Author</th>
<th>Birthday</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>1990</td>
<td>France</td>
</tr>
<tr>
<td>Mark</td>
<td>1985</td>
<td>UK</td>
</tr>
<tr>
<td>John</td>
<td>1980</td>
<td>USA</td>
</tr>
<tr>
<td>Susan</td>
<td>1975</td>
<td>USA</td>
</tr>
<tr>
<td>Frank</td>
<td>1980</td>
<td>Canada</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Olivier</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Frank</td>
</tr>
</tbody>
</table>

Countries of successful authors?
Motivating Example

<table>
<thead>
<tr>
<th>Author</th>
<th>Birthday</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>1990</td>
<td>France</td>
</tr>
<tr>
<td>Mark</td>
<td>1985</td>
<td>UK</td>
</tr>
<tr>
<td>John</td>
<td>1980</td>
<td>USA</td>
</tr>
<tr>
<td>Susan</td>
<td>1975</td>
<td>USA</td>
</tr>
<tr>
<td>Frank</td>
<td>1980</td>
<td>Canada</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>Olivier</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Frank</td>
</tr>
</tbody>
</table>

Countries of successful authors?
Motivating Example

SQL Query

```
SELECT A.Country
FROM Author A, Publication P
WHERE A.Birthday > 1978 AND
    P.Pages > 10 AND
    P.Citations ≥ 100 AND
    P.Year > 2006 AND
    A.Name = P.Author
```
Motivating Example

<table>
<thead>
<tr>
<th>Author</th>
<th>Birthday</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>1990</td>
<td>France</td>
</tr>
<tr>
<td>Mark</td>
<td>1985</td>
<td>UK</td>
</tr>
<tr>
<td>John</td>
<td>1980</td>
<td>USA</td>
</tr>
<tr>
<td>Susan</td>
<td>1975</td>
<td>USA</td>
</tr>
<tr>
<td>Frank</td>
<td>1980</td>
<td>Canada</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication</th>
<th>Author</th>
<th>Page</th>
<th>Citations</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>7</td>
<td>150</td>
<td>2007</td>
<td></td>
</tr>
<tr>
<td>John</td>
<td>6</td>
<td>100</td>
<td>2010</td>
<td></td>
</tr>
<tr>
<td>John</td>
<td>7</td>
<td>150</td>
<td>2008</td>
<td></td>
</tr>
<tr>
<td>John</td>
<td>5</td>
<td>60</td>
<td>2010</td>
<td></td>
</tr>
<tr>
<td>Susan</td>
<td>6</td>
<td>190</td>
<td>2011</td>
<td></td>
</tr>
<tr>
<td>Susan</td>
<td>8</td>
<td>200</td>
<td>2012</td>
<td></td>
</tr>
<tr>
<td>Susan</td>
<td>20</td>
<td>50</td>
<td>2013</td>
<td></td>
</tr>
<tr>
<td>Susan</td>
<td>12</td>
<td>90</td>
<td>2015</td>
<td></td>
</tr>
<tr>
<td>Frank</td>
<td>12</td>
<td>100</td>
<td>2016</td>
<td></td>
</tr>
</tbody>
</table>

Countries of successful authors?

Result

<table>
<thead>
<tr>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
</tr>
</tbody>
</table>
### Why-Not question

#### Author

<table>
<thead>
<tr>
<th>Name</th>
<th>Birthday</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>1990</td>
<td>France</td>
</tr>
<tr>
<td>Mark</td>
<td>1985</td>
<td>UK</td>
</tr>
<tr>
<td>John</td>
<td>1980</td>
<td>USA</td>
</tr>
<tr>
<td>Susan</td>
<td>1975</td>
<td>USA</td>
</tr>
<tr>
<td>Frank</td>
<td>1980</td>
<td>Canada</td>
</tr>
</tbody>
</table>

#### Publication

<table>
<thead>
<tr>
<th>Author</th>
<th>Page</th>
<th>Citations</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>7</td>
<td>150</td>
<td>2007</td>
</tr>
<tr>
<td>John</td>
<td>6</td>
<td>100</td>
<td>2010</td>
</tr>
<tr>
<td>John</td>
<td>7</td>
<td>150</td>
<td>2008</td>
</tr>
<tr>
<td>John</td>
<td>5</td>
<td>60</td>
<td>2010</td>
</tr>
<tr>
<td>Susan</td>
<td>6</td>
<td>190</td>
<td>2011</td>
</tr>
<tr>
<td>Susan</td>
<td>8</td>
<td>200</td>
<td>2012</td>
</tr>
<tr>
<td>Susan</td>
<td>20</td>
<td>50</td>
<td>2013</td>
</tr>
<tr>
<td>Susan</td>
<td>12</td>
<td>90</td>
<td>2015</td>
</tr>
<tr>
<td>Frank</td>
<td>12</td>
<td>100</td>
<td>2016</td>
</tr>
</tbody>
</table>

**Result**

**Country**

Canada

---

*Why not USA?*
Why-Not question

<table>
<thead>
<tr>
<th>Author</th>
<th>Birthday</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>1990</td>
<td>France</td>
</tr>
<tr>
<td>Mark</td>
<td>1985</td>
<td>UK</td>
</tr>
<tr>
<td>John</td>
<td>1980</td>
<td>USA</td>
</tr>
<tr>
<td>Susan</td>
<td>1975</td>
<td>USA</td>
</tr>
<tr>
<td>Frank</td>
<td>1980</td>
<td>Canada</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Olivier</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Frank</td>
</tr>
</tbody>
</table>

Result

<table>
<thead>
<tr>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
</tr>
<tr>
<td>USA</td>
</tr>
</tbody>
</table>
Query-based explanation

SELECT A.Country
FROM Author A,
Publication P
WHERE A.Birthday > 1978 AND
   P.Pages > 10 AND
   P.Citations ≥ 100 AND
   P.Year > 2006 AND
   A.Name = P.Author

Result
Country
Canada
USA
Why-Not Polynomial [TaPP14, CIKM15]

\[ \bigtriangledown_{\text{Name=Author}} + 2\sigma_{\text{Pages}>10} + 2\sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10} + 2\sigma_{\text{Bday}>78} \sigma_{\text{Citations} \geq 100} + \ldots \\
+ \sigma_{\text{Pages}>10} \sigma_{\text{Bday}>78} \sigma_{\text{Citations} \geq 100} \bigtriangledown_{\text{Name=Author}} \]
Why-Not Polynomial [TaPP14, CIKM15]

\[ \bowtie_{\text{Name}=\text{Author}} + 2\sigma_{\text{Pages}>10} + 2\sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10} + 2\sigma_{\text{Bday}>78} \sigma_{\text{Citations}\geq100} + \ldots \]

\[ +\sigma_{\text{Pages}>10} \sigma_{\text{Bday}>78} \sigma_{\text{Citations}\geq100} \bowtie_{\text{Name}=\text{Author}} \]

All correct query-based explanations

All erroneous query condition combinations
Why-Not Polynomial [TaPP14, CIKM15]

\[
\bigotimes_{\text{Name}=\text{Author}} + 2\sigma_{\text{Pages}>10} + 2\sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10} + 2\sigma_{\text{Bday}>78} \sigma_{\text{Citations}\geq 100} + \ldots \\
+ \sigma_{\text{Pages}>10} \sigma_{\text{Bday}>78} \sigma_{\text{Citations}\geq 100} \bigotimes_{\text{Name}=\text{Author}}
\]

Selections

- \(\sigma_{\text{Pages}>10}\)
- \(\sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10}\)
- \(\sigma_{\text{Bday}>78} \sigma_{\text{Citations}\geq 100}\)
- \(\sigma_{\text{Pages}>10} \sigma_{\text{Citations}\geq 100}\)

Joins

- \(\bigotimes_{\text{Name}=\text{Author}}\)

Selections & Joins

- \(\sigma_{\text{Pages}>10} \bigotimes_{\text{Name}=\text{Author}}\)
- \(\sigma_{\text{Bday}>78} \bigotimes_{\text{Name}=\text{Author}}\)
- \(\sigma_{\text{Citations}\geq 100} \bigotimes_{\text{Name}=\text{Author}}\)
- \(\sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10} \bigotimes_{\text{Name}=\text{Author}}\)
- \(\sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10} \sigma_{\text{Citations}\geq 100} \bigotimes_{\text{Name}=\text{Author}}\)
Query Refinement Problem

SELECT A.Country
FROM Author A, Publication P
WHERE ?

Result
Country
Canada
USA
Query Refinement Problem

- **Vast** number of possible query refinements!
- How do I (efficiently) **prune** this space and still find the **good** ones?
Query Refinement Problem

- **Vast** number of possible query refinements!
- How do I (efficiently) **prune** this space and still find the **good** ones?

✓ Why-Not polynomial tells us **exactly** which condition combinations to **change**.
✓ Selections: Use values from **chosen** tuples eliminated by the combination. **Efficient** tuple selection based on the explanation schema.
✓ Joins: Use **outer joins**.
Refining queries with **FixTed**
FixTed

Input
- Query
- Why-Not
- Polynomial
- Database

Minimum Distance

Preciser

Output
List of ordered refined queries.
**Input**
- Query
- Why-Not
- Polynomial
- Database

**Output**
List of ordered refined queries.

---

**Minimum Distance**
Find skyline tuples. Create one new (relaxed) query for each skyline tuple and each explanation.

**Preciser**

---

The diagram illustrates the process of finding minimum distance with explanations, involving selection, join, and explanation steps.
FixTed

Input
- Query
- Why-Not Polynomial
- Database

Output
List of ordered refined queries.

Minimum Distance
Find skyline tuples. Create one new (relaxed) query for each skyline tuple and each explanation.

Preciser
Restrict the refined queries. Add them to $L^s_{Q'}$.

$L^s_{Q'}$
Minimum Distance

Find skyline tuples. Create one new (relaxed) query for each skyline tuple and each explanation.

Preciser

Restrict the refined queries. Add them to $L_s^{Q'}$.

Create the query join graph. If possible, change the joins from the explanation to (left or right) outer-join.

Input
- Query
- Why-Not Polynomial
- Database

Output
List of ordered refined queries.
FixTed

**Input**
- Query
- Why-Not Polynomial
- Database

**Output**
List of ordered refined queries.

---

**Minimum Distance**
Find **skyline** tuples.
Create one new (**relaxed**) query for each skyline tuple and each explanation.

**Preciser**
*Restrict* the refined queries.
Add them to \( L^s_{Q'} \).

Create the query **join graph**.
If possible, change the joins from the explanation to (left or right) **outer-join**.

**Prune** & **order** by similarity & precision
Selection Repair

\[ \bowtie_{Name=Author} + 2\sigma_{Pages>10} + 2\sigma_{Bday>78} \sigma_{Pages>10} + 2\sigma_{Bday>78} \sigma_{Citations\geq 100} + \ldots \]

\[ + \sigma_{Pages>10} \sigma_{Bday>78} \sigma_{Citations\geq 100} \bowtie_{Name=Author} \]

selections

\[ \sigma_{Pages>10} \]
\[ \sigma_{Bday>78} \sigma_{Pages>10} \]
\[ \sigma_{Bday>78} \sigma_{Citations\geq 100} \]
\[ \sigma_{Pages>10} \sigma_{Citations\geq 100} \]

joins

\[ \bowtie_{Name=Author} \]

selections & joins

\[ \sigma_{Pages>10} \bowtie_{Name=Author} \]
\[ \sigma_{Bday>78} \bowtie_{Name=Author} \]
\[ \sigma_{Citations\geq 100} \bowtie_{Name=Author} \]
\[ \sigma_{Bday>78} \sigma_{Pages>10} \bowtie_{Name=Author} \]
\[ \sigma_{Bday>78} \sigma_{Pages>10} \sigma_{Citations\geq 100} \bowtie_{Name=Author} \]
## Selection Repair

<table>
<thead>
<tr>
<th>Author</th>
<th>Name</th>
<th>Birthday</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>1990</td>
<td>France</td>
<td></td>
</tr>
<tr>
<td>Mark</td>
<td>1985</td>
<td>UK</td>
<td></td>
</tr>
<tr>
<td>John</td>
<td>1980</td>
<td>USA</td>
<td></td>
</tr>
<tr>
<td>Susan</td>
<td>1975</td>
<td>USA</td>
<td></td>
</tr>
<tr>
<td>Frank</td>
<td>1980</td>
<td>Canada</td>
<td></td>
</tr>
</tbody>
</table>

## Publication

<table>
<thead>
<tr>
<th>Author</th>
<th>Page</th>
<th>Citations</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>7</td>
<td>150</td>
<td>2007</td>
</tr>
<tr>
<td>John</td>
<td>6</td>
<td>100</td>
<td>2010</td>
</tr>
<tr>
<td>John</td>
<td>7</td>
<td>150</td>
<td>2008</td>
</tr>
<tr>
<td>John</td>
<td>5</td>
<td>60</td>
<td>2010</td>
</tr>
<tr>
<td>Susan</td>
<td>6</td>
<td>190</td>
<td>2011</td>
</tr>
<tr>
<td>Susan</td>
<td>8</td>
<td>200</td>
<td>2012</td>
</tr>
<tr>
<td>Susan</td>
<td>20</td>
<td>50</td>
<td>2013</td>
</tr>
<tr>
<td>Susan</td>
<td>12</td>
<td>90</td>
<td>2015</td>
</tr>
<tr>
<td>Frank</td>
<td>12</td>
<td>100</td>
<td>2016</td>
</tr>
</tbody>
</table>

**Compatible tuples**
## Selection Repair

<table>
<thead>
<tr>
<th>Author</th>
<th>Name</th>
<th>Birthday</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>1990</td>
<td>France</td>
<td></td>
</tr>
<tr>
<td>Mark</td>
<td>1985</td>
<td>UK</td>
<td></td>
</tr>
<tr>
<td>John</td>
<td>1980</td>
<td>USA</td>
<td></td>
</tr>
<tr>
<td>Susan</td>
<td>1975</td>
<td>USA</td>
<td></td>
</tr>
<tr>
<td>Frank</td>
<td>1980</td>
<td>Canada</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>Olivier</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>John</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Susan</td>
</tr>
<tr>
<td>Frank</td>
</tr>
</tbody>
</table>

(join)
Selection Repair

Author

<table>
<thead>
<tr>
<th>Name</th>
<th>Birthday</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>1990</td>
<td>France</td>
</tr>
<tr>
<td>Mark</td>
<td>1985</td>
<td>UK</td>
</tr>
<tr>
<td>John</td>
<td>1980</td>
<td>USA</td>
</tr>
<tr>
<td>Susan</td>
<td>1975</td>
<td>USA</td>
</tr>
<tr>
<td>Frank</td>
<td>1980</td>
<td>Canada</td>
</tr>
</tbody>
</table>

Publication

<table>
<thead>
<tr>
<th>Author</th>
<th>Page</th>
<th>Citations</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olivier</td>
<td>7</td>
<td>150</td>
<td>2007</td>
</tr>
<tr>
<td>John</td>
<td>6</td>
<td>100</td>
<td>2010</td>
</tr>
<tr>
<td>John</td>
<td>7</td>
<td>150</td>
<td>2008</td>
</tr>
<tr>
<td>John</td>
<td>5</td>
<td>60</td>
<td>2010</td>
</tr>
<tr>
<td>Susan</td>
<td>6</td>
<td>190</td>
<td>2011</td>
</tr>
<tr>
<td>Susan</td>
<td>8</td>
<td>200</td>
<td>2012</td>
</tr>
<tr>
<td>Susan</td>
<td>20</td>
<td>50</td>
<td>2013</td>
</tr>
<tr>
<td>Susan</td>
<td>12</td>
<td>90</td>
<td>2015</td>
</tr>
<tr>
<td>Frank</td>
<td>12</td>
<td>100</td>
<td>2016</td>
</tr>
</tbody>
</table>

Explanations

- $\sigma_{\text{Pages}>10}$
- $\sigma_{\text{Citations}\geq100}$
- $\sigma_{\text{Bday}>78}$
- $\sigma_{\text{Pages}>10}\sigma_{\text{Citations}\geq100}$
- $\sigma_{\text{Bday}>78}\sigma_{\text{Pages}>10}$
- $\sigma_{\text{Bday}>78}\sigma_{\text{Citations}\geq100}$
FixTed

Input
- Query
- Why-Not Polynomial
- Database

Output
- List of ordered refined queries.

Minimum Distance

Find skyline tuples.
Create one new (relaxed) query for each skyline tuple and each explanation.

Preciser

Restrict the refined queries.
Add them to \( L^{s_{Q'}} \).

Create the query join graph.
If possible, change the joins from the explanation to (left or right) outer-join.

Prune & order by similarity & precision

\( L^{s_{Q'}} \)

\( L^{s_{Q'}} \)
One selection refinement

Pages $> 10$

- Compatible
- Correct
- Wrong

Pages

5 6 7 9 10 13 16 20

35
One selection refinement

$\sigma_{\text{Pages}>10}$

- Compatible
- Correct
- Wrong

Pages: 5, 6, 7, 9, 10, 13, 16, 20
One selection refinement

Refine with 7

\[ \sigma_{\text{Pages} > 10} \quad \rightarrow \quad \sigma_{\text{Pages} \geq 7} \]
Two selections refinement

\( \sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10} \)

- **Compatible**
- **Correct**
- **Wrong**

Birthday:
- 7, 1990
- 6, 1975
- 8, 1975
- 12, 1980

Pages:
- 5, 9, 10, 13, 16, 20

Dates and pages highlighted:
- 7, 1990 (compatible)
- 6, 1975 (correct)
- 8, 1975 (correct)
- 12, 1980 (correct)
Two selections refinement

\[ \sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10} \]
Two selections refinement

\[ \sigma_{Bday>78} \sigma_{Pages>10} \]
Two selections refinement

Refine with

\( \sigma_{\text{Bday} \geq 1975} \sigma_{\text{Pages} \geq 8} \)

(8, 1975)
Minimum Distance Refinements

<table>
<thead>
<tr>
<th>Refinement ( (L^s_Q) )</th>
<th>New conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Q_1 )</td>
<td>( \sigma_{\text{Pages}} \geq 7 )</td>
</tr>
<tr>
<td>( Q_2 )</td>
<td>( \sigma_{\text{Bday}} \geq 1975 \sigma_{\text{Pages}} \geq 8 )</td>
</tr>
<tr>
<td>( Q_3 )</td>
<td>( \sigma_{\text{Bday}} \geq 1975 \sigma_{\text{Citations}} \geq 90 )</td>
</tr>
</tbody>
</table>
Input
- Query
- Why-Not Polynomial
- Database

Output
List of ordered refined queries.

FixTed

Minimum Distance
Find skyline tuples. Create one new (relaxed) query for each skyline tuple and each explanation.

Preciser
Restrict the refined queries. Add them to $L_{Q'}^s$.

Create the query join graph. If possible, change the joins from the explanation to (left or right) outer-join.

Prune & order by similarity & precision

$L_{Q'}^s$
Preciser

<table>
<thead>
<tr>
<th>Refinement ((L^s_Q))</th>
<th>New conditions</th>
<th>False positive tuples</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Q_1)</td>
<td>(\sigma_{\text{Pages} \geq 7})</td>
<td>1</td>
</tr>
<tr>
<td>(Q_2)</td>
<td>(\sigma_{\text{Bday} \geq 1975 \land \text{Pages} \geq 8})</td>
<td>0</td>
</tr>
<tr>
<td>(Q_3)</td>
<td>(\sigma_{\text{Bday} \geq 1975 \land \text{Citations} \geq 90})</td>
<td>0</td>
</tr>
</tbody>
</table>
Preciser

Year

2006 2007 2008 2016

compatible  correct  wrong
Preciser

Year

2006 2007 2008 2016

compatible correct wrong
### Preciser

<table>
<thead>
<tr>
<th>Refinement ($L^s_Q$)</th>
<th>New conditions</th>
<th>False positive tuples</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_1$</td>
<td>$\sigma_{\text{Pages} \geq 7}$</td>
<td>1</td>
</tr>
<tr>
<td>$Q_2$</td>
<td>$\sigma_{\text{Bday} \geq 1975} \sigma_{\text{Pages} \geq 8}$</td>
<td>0</td>
</tr>
<tr>
<td>$Q_3$</td>
<td>$\sigma_{\text{Bday} \geq 1975} \sigma_{\text{Citations} \geq 90}$</td>
<td>0</td>
</tr>
<tr>
<td>$Q_4$</td>
<td>$\sigma_{\text{Pages} \geq 7} \sigma_{\text{Year} \geq 2008}$</td>
<td>0</td>
</tr>
<tr>
<td>$Q_5$</td>
<td>$\sigma_{\text{Pages} \geq 7} \sigma_{\text{Bday} \geq 1980}$</td>
<td>0</td>
</tr>
</tbody>
</table>
FixTed

Input
- Query
- Why-Not Polynomial
- Database

Output
List of ordered refined queries.

Minimum Distance
Find skyline tuples. Create one new (relaxed) query for each skyline tuple and each explanation.

Preciser
Restrict the refined queries. Add them to $L_s^{Q'}$.

Create the query join graph. If possible, change the joins from the explanation to (left or right) outer-join.
Join Repair

\[ \bowtie_{\text{Name}=\text{Author}} + 2\sigma_{\text{Pages}>10} + 2\sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10} + 2\sigma_{\text{Bday}>78} \sigma_{\text{Citations} \geq 100} + \ldots \]

\[ + \sigma_{\text{Pages}>10} \sigma_{\text{Bday}>78} \sigma_{\text{Citations} \geq 100} \bowtie_{\text{Name}=\text{Author}} \]

---

**selections**

\[ \sigma_{\text{Pages}>10} \]
\[ \sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10} \]
\[ \sigma_{\text{Bday}>78} \sigma_{\text{Citations} \geq 100} \]
\[ \sigma_{\text{Pages}>10} \sigma_{\text{Citations} \geq 100} \]

**joins**

\[ \bowtie_{\text{Name}=\text{Author}} \]

---

**selections & joins**

\[ \sigma_{\text{Pages}>10} \bowtie_{\text{Name}=\text{Author}} \]
\[ \sigma_{\text{Bday}>78} \bowtie_{\text{Name}=\text{Author}} \]
\[ \sigma_{\text{Citations} \geq 100} \bowtie_{\text{Name}=\text{Author}} \]
\[ \sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10} \bowtie_{\text{Name}=\text{Author}} \]
\[ \sigma_{\text{Bday}>78} \sigma_{\text{Pages}>10} \sigma_{\text{Citations} \geq 100} \bowtie_{\text{Name}=\text{Author}} \]
Join Repair

Query join graph

Author

Name=Author

Publication
Join Repair

Why not Author.Country=USA?
Join Repair

*Direct relation*

Why not \textit{Author.Country=} \textit{USA}?
Join Repair

Author <-> Publication

Refine with

Author <-> Publication

(USA)
Join Repair

Original query

SELECT   A.Country
FROM      Author A, Publication P
WHERE   A.Birthday > 1978  AND
        P.Pages > 10           AND
        P.Citations ≥ 100  AND
        P.Year > 2006         AND
        A.Name = P.Author

Refined query

Q6

SELECT   A2.Country
FROM      ( SELECT A.Name, A.Country
            FROM Author A
            WHERE A.Birthday > 1978 ) as A2
LEFT OUTER JOIN ( SELECT P.Author
                   FROM Publication P
                   WHERE P.Pages > 10 AND P.Citations ≥ 100)
                   as P2
ON   A2.Name = P2.Author
Input
- Query
- Why-Not Polynomial
- Database

Output
List of ordered refined queries.

Minimum Distance
Find skyline tuples. Create one new (relaxed) query for each skyline tuple and each explanation.

Preciser
Restrict the refined queries. Add them to $L_s^{Q'}$.

Create the query join graph. If possible, change the joins from the explanation to (left or right) outer-join.

Prune & order by similarity & precision

$L_s^{Q'}$
Prune and Order

Metrics

Similarity:

Number of changed conditions

Distance between the original and changed condition values

Number of added conditions

\[
\text{Precision: } \frac{|Q'[D]| - FP_{Q'}}{|Q'[D]|}
\]

Skyline of refined queries based on the metrics.
### Prune and Order

#### All Refinements

<table>
<thead>
<tr>
<th>Q1</th>
<th>σPages≥7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2</td>
<td>σBday≥1975σPages≥8</td>
</tr>
<tr>
<td>Q3</td>
<td>σBday≥1975σCitations≥90</td>
</tr>
<tr>
<td>Q4</td>
<td>σPages≥7σYear≥2008</td>
</tr>
<tr>
<td>Q5</td>
<td>σPages≥7σBday≥1980</td>
</tr>
<tr>
<td>Q6</td>
<td>Name=Author</td>
</tr>
</tbody>
</table>

#### Pruned & Ordered Refinements

<table>
<thead>
<tr>
<th>Q2</th>
<th>σBday≥1975σCitations≥90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>σPages≥7</td>
</tr>
<tr>
<td>Q6</td>
<td>Name=Author</td>
</tr>
</tbody>
</table>

*Most similar*
Contributions and Future Work

Contributions of *FixTed*:

- Query refinement involving *join* and *selection* reparation.
- Leverage of *explanations* given by the *Why-Not polynomial*, to *optimize* the computation.
- Computation of the *most similar* refinement.
- Efficiency *improvement* of the *skyline* tuples computation by:
  - Reducing the number of considered attributes.
  - Reducing the number of considered input tuples.
- Join reparation using *outer joins*. 
Contributions and Future Work

Future Work:

• Thorough experimental evaluation.

• Further investigation for join reparation
  (e.g., by using integrity constraints).
Contributions of *FixTed*:

Query refinement involving **join** and **selection** reparation.

Leverage of **explanations** given by the **Why-Not polynomial**, to **optimize** the computation.

Computation of the **most similar** refinement.

Efficiency **improvement** of the **skyline** tuples computation by:

- Reducing the number of considered attributes.
- Reducing the number of considered input tuples.

Join reparation using **outer joins**.

**Future Work:**

- Thorough experimental evaluation.
- Further investigation for join reparation.