Two applications

(this is NOT an intro to provenance)

Dan Suciu
University of Washington
Two Applications

1. Provenance summaries for query answering using probabilistic views
   • With Chris Re
   • Status: ongoing

2. Provenance for privacy in RFID applications
   • With Vibhor Rastogi
   • Status: preliminary
Application 1: Provenance summaries

Query Answering Using Views

\[ V(x) = R(x,y), S(x,y,z), T(x,z) \]

Materialize:

\[ V = \]

Query:

\[ q = R(x,y), S(x,y,z), T(x,z), U(x,v), K(v,w) \]

Rewrite to:

\[ q = V(x), U(x,v), K(v,w) \]

More efficient!
Using *Probabilistic Views*

\[ V(x) = R^p(x,y), S^p(x,y,z), T^p(x,z) \]

<table>
<thead>
<tr>
<th></th>
<th>R^p:</th>
<th></th>
<th>S^p:</th>
<th></th>
<th>T^p:</th>
<th></th>
<th>V^p:</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
<td>P</td>
<td>x</td>
<td>y</td>
<td>z</td>
<td>P</td>
<td>x</td>
</tr>
<tr>
<td>a</td>
<td>m</td>
<td>0.3</td>
<td>a</td>
<td>m</td>
<td>s</td>
<td>0.1</td>
<td>a</td>
</tr>
<tr>
<td>a</td>
<td>n</td>
<td>0.2</td>
<td>a</td>
<td>n</td>
<td>s</td>
<td>0.5</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>m</td>
<td>0.4</td>
<td>b</td>
<td>m</td>
<td>t</td>
<td>0.4</td>
<td>b</td>
</tr>
<tr>
<td>b</td>
<td>p</td>
<td>0.1</td>
<td>b</td>
<td>p</td>
<td>t</td>
<td>0.9</td>
<td></td>
</tr>
</tbody>
</table>

**Application 1: Provenance summaries**
Using *Probabilistic Views*

\[ V(x) = R^p(x,y), S^p(x,y,z), T^p(x,z) \]

### Marginal probabilities

<table>
<thead>
<tr>
<th>x</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.1</td>
</tr>
<tr>
<td>b</td>
<td>0.5</td>
</tr>
</tbody>
</table>

---

**Application 1: Provenance summaries**
Using Probabilistic Views

\[ V(x) = R^p(x,y), S^p(x,y,z), T^p(x,z) \]

\[ q = V(x), U(x,v), K(v,w) \]

Application 1: Provenance summaries

Marginal probabilities

Marginal Prob in \( V^p \) insufficient
Enter Provenance

Application 1: Provenance summaries

V(x) = \( R^p(x,y), S^p(x,y,z), T^p(x,z) \)
Enter Provenance

\[ V(x) = R^p(x,y), S^p(x,y,z), T^p(x,z) \]
Application 1: Provenance summaries

Enter Provenance

\[ V(x) = R^p(x,y), S^p(x,y,z), T^p(x,z) \]

\[ q = V(x), U(x,v), K(v,w) \]
### Application 1: Provenance summaries

#### “Provenance Summary”

$$V^p:$$

<table>
<thead>
<tr>
<th>x</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>E_1 &amp; F_1 &amp; G_1 \lor E_2 &amp; F_2 &amp; G_1</td>
</tr>
<tr>
<td>b</td>
<td>E_3 &amp; F_3 &amp; G_3 \lor E_4 &amp; F_4 &amp; G_3</td>
</tr>
</tbody>
</table>

\[\Rightarrow\]

<table>
<thead>
<tr>
<th>x</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>H_1</td>
</tr>
<tr>
<td>b</td>
<td>H_2</td>
</tr>
</tbody>
</table>

A very concise summary of the provenance
Application 1: Provenance summaries

“Provenance Summary”

$V^p$: $\begin{array}{c|c}
    x & E \\
    \hline
    a & E_1 \land F_1 \land G_1 \lor E_2 \land F_2 \land G_1 \\
    b & E_3 \land F_3 \land G_3 \lor E_4 \land F_4 \land G_3 \\
\end{array}$

Now we *know* we can use the marginals

A very concise summary of the provenance
“Provenance Summary”

\[ V^p: \]

\[
\begin{array}{c|c}
 x & E \\
\hline
 a & E_1 \land F_1 \land G_1 \lor E_2 \land F_2 \land G_1 \\
 b & E_3 \land F_3 \land G_3 \lor E_4 \land F_4 \land G_3 \\
\end{array}
\]

Now we *know* we can use the marginals

A very concise summary of the provenance

**Status**: deciding if a view V has independent tuples is \( \Pi^p_2 \) complete

**Open**: find a *minimal provenance summary*
Application 2: RFID Security

RFID Ecosystem at UW

[Welbourne'2007]
RFID Data

Base table

SIGHTINGS(tagID, antennaID, time)
EnteredRoom(personTagID, room, time)
CarriesObject(personTagID, objectTagID, time)
Meeting(personTagID1, personTagID2, time)
............

Derived tables (views)
Application 2: RFID Security

Privacy w. Authorization Views

Alice’s query

\[ q(x) = \text{EnteredRoom}(x, "Rm552", t), \text{Yesterday}(t) \]

Authorization view

System answers the query if it can be rewritten in terms of views; else deny

[Rizvi’2004]
Privacy and Provenance

• Issue 1: the data *itself* is a materialized view. How can we make access control decisions based on how the data was derived?

• Issue 2: the *authorization views* are probabilistic. How can we grant access with probability, say, 75%?
Questions ?