Provenance Management In Knowledge Graphs

Prof. Arnab Bhattacharya
Dr. Srikanta B Jagannath
Garima Gaur

April 6, 2017
Knowledge Graph

- Graphical way of representing knowledge
- Belongs to the category of semantic networks.
- Directed or undirected graph with concepts as vertices and relationships between concepts as edges.
Due to new facts coming in – resulting in deletion and insertion.

Semantics of affecting an answer set is based on query type and the nature of change in KG:

- Top-k queries: The value of parameter under consideration might have changed.
- Descriptive query: Particular item in answer set doesn’t satisfy query condition.
- Shortest path query: Answer is no more correct.
- Analytical Query: Evaluated value is no more valid.
WHY DOES IT MATTER?

- Critical decision making based on query result.
**Why does it matter?**

- Critical decision making based on query result.

- Ever growing Knowledge Graphs.

Recompute Query!!
Handling Descriptive Query

- Trying to answer a simple question
  Does the deletion/insertion of an edge $e$ affects the query result $R$?
Handling Descriptive Query

- Trying to answer a simple question
  Does the deletion/insertion of an edge $e$ affects the query result $R$?
- Metadata provides better insight.
- Provenance — origin of something.
- Various perspectives under one umbrella —
  - *Why*-provenance: Comprises of the data involved.
  - *How*-provenance: Concerns with the derivation process.
In need of Provenance Model

Query:
Select ?actor
where{
  ?actor ActedIn ?movie.
  ?movie Genre "Com".
}

"Com"

"Drama"

A1
ActedIn

A3
ActedIn

A2
ActedIn

A4
ActedIn

M1
Genre

M2
ActedIn

M3
Genre

M4
Genre

A1
ActedIn

M1
Genre

M2
ActedIn

M3
Genre

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre

M3
ActedIn

M4
Genre
IN NEED OF PROVENANCE MODEL

Query:
Select ?actor
where{
  ?actor ActedIn ?movie.
  ?movie Genre "Com".
}

Result R
A1
A2
**Obvious Choice – Lineage**

Query:
Select `?actor`
where{
  `?actor ActedIn ?movie`.
  `?movie Genre "Com"`.
}

![Graph diagram](image-url)
Query:
Select ?actor
where{
?actor ActedIn ?movie.
?movie Genre "Com".
}

Result R

<table>
<thead>
<tr>
<th></th>
<th>Lineage</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>{ e1, e4 }</td>
</tr>
<tr>
<td>A2</td>
<td>{ e2, e3, e4, e5 }</td>
</tr>
</tbody>
</table>
Let’s delete an edge

Assume that edge $e_1$ gets deleted.

**Result R**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>${ e_1, e_4 }$</td>
</tr>
<tr>
<td>A2</td>
<td>${ e_2, e_3, e_4, e_5 }$</td>
</tr>
</tbody>
</table>

**Updated Result R’**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td>${ e_2, e_3, e_4, e_5 }$</td>
</tr>
</tbody>
</table>
IS LINEAGE SUFFICIENT?

What if edge e3 gets deleted?

Result $R'$

A2 $\{e2, e3, e4, e5\}$
**IS LINEAGE SUFFICIENT?**

What if edge $e_3$ gets deleted?

**Result R’**

$\text{A2} \{ e_2, e_3, e_4, e_5 \}$

- A2 still an answer!!
- Need a provenance model which can capture derivation process.
Provenance Polynomial encodes the interaction of involved edges.

- Each term of polynomial is self-sufficient.
HOW DELETION WORKS

- Membership-function $M : E \rightarrow \{0, 1\}$, where $E$ is the set of edge variables.

$$M(e) = \begin{cases} 
1 & \text{if } e \text{ is part of KG} \\
0 & \text{if } e \text{ is deleted}
\end{cases}$$
How deletion works

- Membership-function $M : E \rightarrow \{0, 1\}$, where $E$ the is set of edge variables.

$$M(e) = \begin{cases} 
1 & \text{if } e \text{ is part of } \text{KG} \\
0 & \text{if } e \text{ is deleted}
\end{cases}$$

- Substitute the edge variables by $M(e)$ and evaluate the polynomial.

- Result set persists all the answers whose corresponding provenance polynomial evaluates to non-zero value.
**How Deletion Works**

- Membership-function $M : E \rightarrow \{0, 1\}$, where $E$ the is set of edge variables.

$$M(e) = \begin{cases} 
1 & \text{if } e \text{ is part of KG} \\
0 & \text{if } e \text{ is deleted} 
\end{cases}$$

- Substitute the edge variables by $M(e)$ and evaluate the polynomial.

- Result set persists all the answers whose corresponding provenance polynomial evaluates to non-zero value.

- On deletion of edge $e_3$,

| A2 | $e_3.e_5 + e_2.e_4 = 0.1 + 1.1 = 1$ |
Our System

- 2-step process
  - Search-step: Find candidate queries.
  - Confirmation-step: Evaluate polynomials to confirm the changes.

YAGO Dataset: 5.8M nodes, 22.5M edges and 39 relations.

Achieved an update time of 2.6% of RDF query execution time.
Our System

- 2-step process
  - Search-step: Find candidate queries.
  - Confirmation-step: Evaluate polynomials to confirm the changes.

- YAGO Dataset: 5.8M nodes, 22.5M edges and 39 relations.
- Achieved an update time of 2.6% of RDF query execution time.
Questions?
Thanks!