Compact Skeletons

- Assume tuples components are scattered over website
- We have a tagger that can tag all tuple components on website
  - Assume no noise for now
- Reconstruct relation

Website

Data Graph

Skeleton

Relation

Website

Dept \((D)\)

Title \((T)\)

Salary \((S)\)

Address \((A)\)

Welcome to Big Corp!

Join our team.

Jobs are available in these departments:

- R&D
- Corporate

The following jobs are open:

- Job #12345
- Job #12346

Send resumes to:

1200 Jose Blvd, CA

Job Title: Programmer
Salary: 100K
Must know Java....

R & D

Corporate

1200 Jose Blvd

Programmer

100K

Must know Java...
Compact Skeletons

- A skeleton is compact if all overlays are consistent
- Perfect if each node and edge of data graph is covered by at least one overlay
- Given a data graph G, does G have a Perfect Compact Skeleton (PCS)?
  - Not always
  - But if it exists it is unique

PCS Algorithm

- Work bottom-up:
  - Compute node signatures
  - Place nodes in equivalence classes based on signature
  - Construct skeleton from equivalence classes

Incomplete information

- Corporate
  - Admin Asst 60K CEO 400 7th Ave
- D
  - 60K

Incomplete information

- Corporate
  - Admin Asst 60K CEO 400 7th Ave
- D
  - 60K

Incomplete information

Partial Compact Skeletons

- For data graphs with incomplete information, we allow partial overlays
  - Results in nulls in relation
- If we can use consistent partial overlays to cover every node and edge of the graph, we have a partially perfect compact skeleton (PPCS)

Tuple subsumption

- **Tuple** \( t \) subsumes tuple \( u \) if \( t \) and \( u \) agree on every component of \( u \) that is not null

\[
\begin{array}{cccc}
T & S & D & A \\
\rightarrow & t_i & s_i & \perp & a_i \\
\leftarrow & u_i & \perp & \perp & a_i \\
\end{array}
\]

Noisy Data Graphs

- Real-life websites are **noisy**
  - False positives e.g., MS = degree, state or Microsoft?
  - Non-skeleton links e.g., featured products

Data graph for a retail website

Coverage of a skeleton

For simplicity: assume all nodes have a label
Skeletons for Noisy Data Graphs

- **Problem:**
  - Find skeleton K with optimal coverage, called the best-fit skeleton (BFS)
- **NP-complete**

**Greedy Heuristic for BFS**

<table>
<thead>
<tr>
<th>Label</th>
<th>Parent</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>I</td>
<td>3</td>
</tr>
<tr>
<td>A</td>
<td>I</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>I</td>
<td>C</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>R</td>
<td>1</td>
</tr>
</tbody>
</table>

Greedy skeleton
**Weighted Greedy Heuristic**

- **Simple Greedy** heuristic uses parent counts
  - "Memory-less"
- **Weighted Greedy** heuristic takes into account past selections to improve simple greedy selection
  - Computes "benefit" of each decision at every stage
Summary

Relation

Compact Skeleton

Data Graph

Website

Greedy skeleton
Coverage = 9

Weighted greedy skeleton
Coverage = 15