Towards a **Scalable and Robust Entity Resolution**

- **BLOCKING UNDER RELATIONAL CONSTRAINTS**

MINGYUAN CUI  U5323288
SUPERVISED BY DR QING WANG & DR HUIZHI LIANG
Outline

- Background
- Overview
- Motivation & example
- Goal of the Project
- Methods/Techniques
- Evaluation
**Background**

**What is entity resolution and why do we need it?**

**Entity Resolution** is the process of matching records across data sets that refer to the same entity in the real world.

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>POST CODE</th>
<th>CITY OF RESIDENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>William James</td>
<td>2602</td>
<td>Canberra</td>
</tr>
<tr>
<td>2</td>
<td>Bill James</td>
<td>2602</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>William Pitt</td>
<td>2606</td>
<td>ACT</td>
</tr>
</tbody>
</table>
Overview

Database → Blocking/Indexing → Matching → Clustering → Database

Propagation
Motivation

What is the blocking process for?

Comparing between all of the records is impractical: a rule for computing pairwise similarities has a time complexity of $O(n^2)$.

Blocking breaks the dataset in blocks, which are possibly overlapping.

Record pairs in the same block are matched.
## Example

### What is the blocking process for?

<table>
<thead>
<tr>
<th>Last Name</th>
<th>First Name</th>
<th>Post Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roger</td>
<td>Jim</td>
<td>2602</td>
</tr>
<tr>
<td>Roger</td>
<td>James</td>
<td>2606</td>
</tr>
<tr>
<td>Tom</td>
<td>James</td>
<td>90024</td>
</tr>
<tr>
<td>Tom</td>
<td>Jimmy</td>
<td>90025</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Last Name</th>
<th>First Name</th>
<th>Post Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roger</td>
<td>Jim</td>
<td>2602</td>
</tr>
<tr>
<td>Roger</td>
<td>James</td>
<td>2606</td>
</tr>
<tr>
<td>Tom</td>
<td>James</td>
<td>90024</td>
</tr>
<tr>
<td>Tom</td>
<td>Jimmy</td>
<td>90025</td>
</tr>
</tbody>
</table>

- 10^5
- 120
- 230
Goal of the Project

Implementing a framework that incorporates relational constraints into the blocking process to achieve **scalable** and **robust** entity resolution.

To conduct a literature review on constraint-based entity resolution methods, and block techniques.

To incorporate relational **constraints** into the blocking process.

To analyze the efficiency and effectiveness of the developed approach, and evaluate the approach over real data sets.
Methods/Techniques

- Blocking
- Constraints
  - Hard constraints
  - Soft constraints
Evaluation

- Programming Language
  Java

- Test Dataset
  Cora dataset
  NC voter dataset

- Evaluation Metrics
  Time
  Accuracy
  Memory Usage
Any Questions?