Raquel Hill

Semantic Privacy of Medical Survey Data

COLLABORATORS
Daryl Lopes, Jayalakshmi Sureshkumar
Joseph Cottam

INDIANA UNIVERSITY
SCHOOL OF INFORMATICS, COMPUTING, AND ENGINEERING
Uniqueness and Re-identification

Quasi-Identifier

Re-identification of William Weld

- Name
- Address
- Date registered
- Party affiliation
- Date last voted
- Voter List

Hospital Discharge Data
- Ethnicity
- Visit date
- Diagnosis
- Procedure
- Medication
- Total charge

Uniqueness in Medical Survey Responses

• Respondents of surveys can potentially have their identity and privacy compromised.
• unrestricted text responses pose a significant risk to user privacy,
• applying traditional privacy models that aggregate and suppress data may reduce data utility by a substantial amount
Semantic Aggregation

Discussion

UMLS specifies three different semantic relationship types:

- RN: Indicates a “narrower” relationship between disorders (e.g. child or descendant)
- RO: Indicates an “other” relationship between disorders, the disorders are not necessarily synonymous (e.g. sibling, cousin, etc)
- RB: Indicates a “broader” relationship between concepts (e.g. parent or ancestor)
Relating Unique Responses
Relating Unique Responses: Outliers and Aggregation
THANK YOU

ralhill@indiana.edu

www.cs.indiana.edu/~ralhill

Funded in part by NSF: TC1012081 “Anonymizing Textual Data and Its Impact on Utility”