Ontology-based Negotiation and Enforcement of Privacy Constraints in Collaborative Knowledge Discovery

Lauri Tuovinen and Alan F. Smeaton
Insight Centre for Data Analytics, Dublin City University
Presented at PAP 2018 (ECML-PKDD), September 10, Dublin, Ireland

Background

- KDD-CHASER: a Marie Curie Individual Fellowship project dealing with knowledge discovery in data as a collaborative process
- February 2018 – January 2020 @ Insight Centre for Data Analytics, Dublin City University
- Aim: to develop a new software platform for data owners and data analysts to collaborate on knowledge discovery projects
- Focus: analysis of personal (lifelogging) data and enabling participation of data owners with no knowledge discovery expertise
- This presentation discusses a work-in-progress concept, only rudimentary implementation and validation done so far

Collaborative knowledge discovery

- Knowledge discovery has always been a collaborative process; what we're interested in is specifically *remote* or *online* collaboration
- The participants of a given collaboration may be located anywhere in the world, so an online platform is required to mediate the collaboration
- Initially, a user of the platform may not even know anyone they could collaborate with, so a key feature of the platform is the ability to search for and negotiate with potential partners
- Accommodating non-expert users involves special requirements concerning e.g. the usability of the platform and facilitation of human-human interactions

Collaboration for personal analytics

- Collecting data about your own life is fairly popular nowadays – think fitness and sleep trackers – but not many people have the skills and knowledge necessary to refine the data any further
- These products typically implement some KD algorithms, e.g. to convert accelerometer readings into more meaningful physical activity scores
- However, if you want to extract additional knowledge, e.g. by combining data from multiple sources, you have to do the implementation yourself
- Collaboration with knowledge discovery experts would enable non-expert data owners to achieve this, but existing collaborative KD systems do not address the particular challenges of personal analytics

The privacy trade-off

- One of the challenges in collaborative personal analytics is finding the perfect balance between privacy and utility
- The data owner is likely to want to keep some of their data private, especially when collaborating with partners they've only recently connected with through the collaboration platform
- However, withholding data reduces the utility of the dataset, which may prevent effective collaboration if the expert partners are not permitted to access enough data to do meaningful analysis
- To avoid losing either too much utility or too much privacy, a negotiation mechanism is needed to detect and resolve conflicts between data owners' privacy constraints and KD experts' data requirements

Proposed solution

- We propose to use an ontology to represent knowledge about data analysis tasks and privacy constraints in collaborative knowledge discovery
- The software platform provides a UI for data owners and data analysts to specify their requirements in terms of the classes and properties of the ontology
- By invoking a reasoner, the platform can detect conflicts where a data analyst would like to access some data items that the data owner would prefer to keep private
- The collaborators can then attempt to resolve the conflict, e.g. by applying a privacy-preserving transformation that retains some of the utility of the data

The ontology – core concepts

- **Collaboration**: a collaborative knowledge discovery project
- **Data Owner**, **Data Analyst**: the participants of a Collaboration
- **Dataset**, **Data Item**: the data to be analysed
- **Analysis Task**: a knowledge discovery operation (or set of operations) to be performed on a group of Data Items, specified by a Data Analyst
- **Privacy Constraint**, **Access Restriction**: a rule to restrict the level of access of a Data Analyst to a group of Data Items, specified by a Data Owner
- **Analysis Method**, **Protection Method**: computational methods that implement Analysis Tasks and Access Restrictions, respectively
- **Utility Reduction**: the effect of a Protection Method on the utility of an Analysis Method

The ontology – class diagram
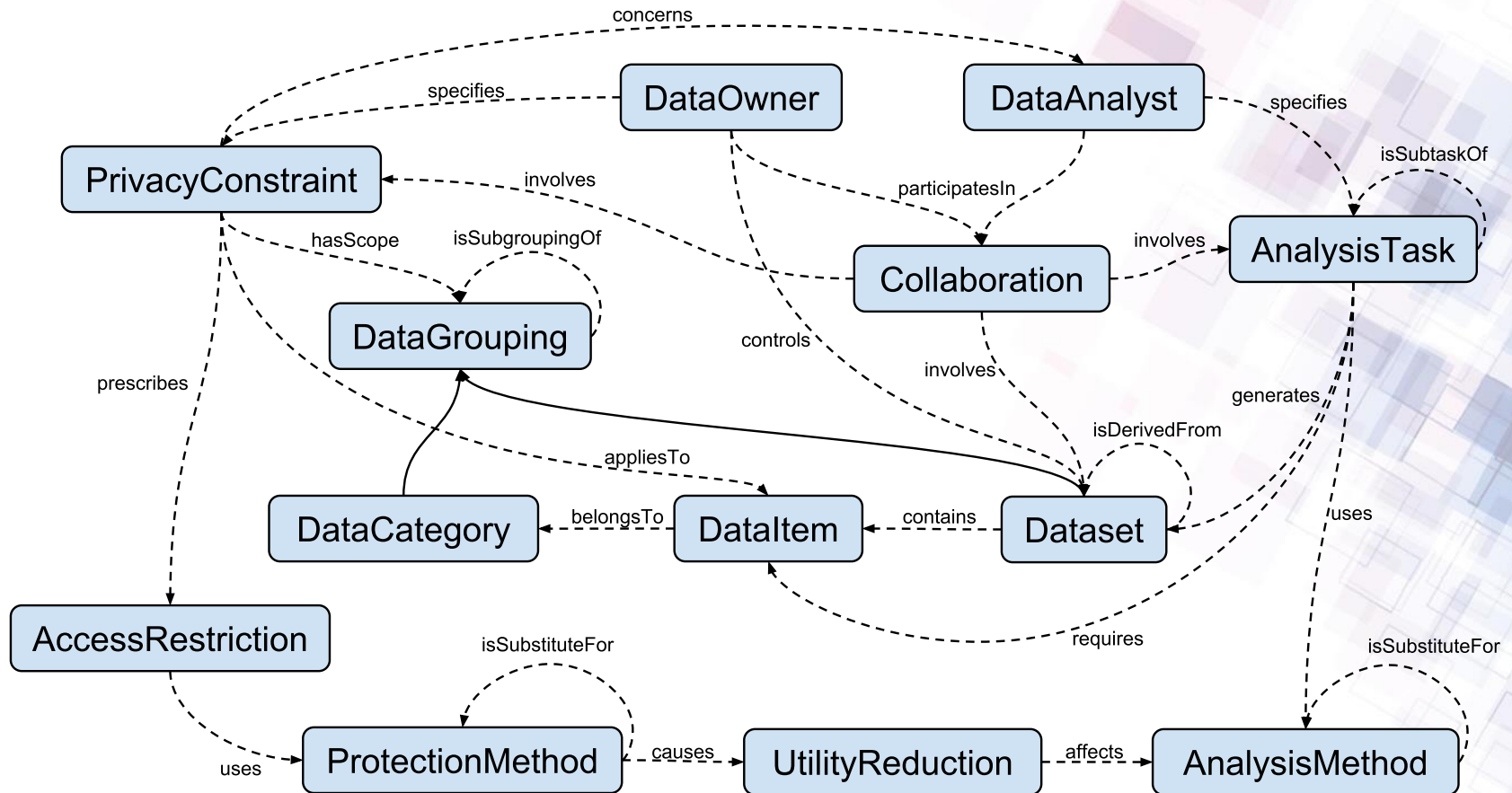
The ontology – reasoning

- Each Analysis Task has properties connecting it to the Data Analyst who specified it and the Data Items it requires
- Each Privacy Constraint has properties connecting it to the Data Analyst it concerns and the Data Items it applies to
- These properties enable the reasoner to identify Data Items that are both required by and denied to a given Data Analyst, resulting in a conflict
- Additionally, the outputs of Analysis Tasks are registered as Datasets and connected to the input Datasets as derivatives
- The original Data Owner becomes the owner of all derivative Datasets as well, enabling them to exert their control throughout the collaborative knowledge discovery process

Current status & future work

- The core concepts of the ontology have been implemented and tested using Protégé and FaCT++
  - Conflict detection works as intended
  - Data Owners are correctly identified as controllers of derivative Datasets
- Analysis Methods, Access Restrictions and Utility Reductions have yet to be modelled in any detail
  - As a result, the ontology is not (yet) well suited to representing cases where the restricted data is to be e.g. generalised rather than simply blocked
- The collaborative knowledge discovery software platform does not exist, but development of a proof-of-concept implementation is underway

Thanks!

- Get in touch:
  - Email: **lauri.tuovinen@insight-centre.org**
  - Twitter: **@LauriTuovinen**
  - Blog: **http://kddchaser.home.blog**
- Acknowledgments: