From Data Flows to Privacy Issues: A User-Centric Semantic Model for Representing and Discovering Privacy Issues*

Yang Lu
School of Computing, University of Kent
Y.Lu@kent.ac.uk

Shujun Li
School of Computing, University of Kent
S.J.Li@kent.ac.uk

Abstract

In today's highly connected cyber-physical world, people are constantly disclosing personal and sensitive data to different organizations and other people through the use of online and physical services. Such data disclosure activities can lead to unexpected privacy issues. However, there is a general lack of tools that help to improve users' awareness of such privacy issues and to make more informed decisions on their data disclosure activities in wider contexts. To fill this gap, this paper presents a novel user-centric, data-flow graph based semantic model, which can show how a given user's personal and sensitive data are disclosed to different entities and how different types of privacy issues can emerge from such data disclosure activities. The model enables both manual and automatic analysis of privacy issues, therefore laying the theoretical foundation of building data-driven and user-centric software tools for people to better manage their data disclosure activities in the cyber-physical world.

1. Introduction

Living in a highly digitized and networked world and the wider cyber-physical space, people are interacting with organizations and other people more and more frequently via different kinds of online and offline (physical) services and products. For instance, through using travel agencies (e.g., Agoda and Booking.com) online or via physical means, people can arrange flight tickets, hotel rooms, transportation choices and tourist activities.In addition to providing basic services, it is a common practice for service providers to share customers' personal data with other third-party organizations, such as advertisers, insurers and relevant governmental bodies, due to legal requirements or some business reasons (e.g., to offer more personalized services). Furthermore, many people actively share information about their lives online with other people, e.g., on online social networks (OSNs) and web forums, which further extends the scale of data sharing. All such data sharing activities can lead to different kinds of privacy issues, caused by personal data flowing from the user (i.e., the data owner) to different entities in the cyber-physical world, directly or indirectly.

Certain privacy issues are actually caused by self-disclosures by the users themselves [1]. Past work was mostly designed to address "known events" such as decisions on data collection, access and processing, however insufficient work has been done towards privacy issues related to data flows unknown to users. To help reduce self-disclosures and associated privacy issues [2], it is necessary to keep users aware of data flows that can lead to possible privacy issues. In this context, many researchers have proposed to use a privacy related ontology or other conceptual models to systematically formalize knowledge about privacy by "explicit concepts and relations", in order to discover "implicit facts" (i.e., privacy issues or risks) [3]. With enhanced awareness, further privacy enhancement mechanisms can be adopted to help managing such privacy risks, e.g., adjusting access control or privacy policies, removing unused data, switching to more privacy-friendly services, and using privacy software tools to automatically block unwanted data disclosure. Specially, privacy nudging has also been proposed as a mechanism for a privacy-aware computing system to nudge users towards data disclosure decisions that protect their privacy better [4].

Most past theoretical work on privacy ontologies and concept modeling focuses either on high-level concepts or a narrow aspect or application domain (e.g., privacy policies, OSNs). So far, we have not seen any work focusing on user-centric data flows across different types of data consumers (services, organizations, other people, etc.). This paper fills this

*This is an extended version of the following paper: Yang Lu and Shujun Li, "From Data Flows to Privacy Issues: A User-Centric Semantic Model for Representing and Discovering Privacy Issues," accepted to HICSS 2020 (53rd Hawaii International Conference on System Sciences), to be held from January 7-10, 2020 in Hawaii, USA. Please cite the paper using the above citation information.

gap by proposing a novel user-centric and graph-based model for formalizing personal data flows that may lead to privacy issues. The model is generic enough to cover a wide range of data disclosure activities of people in the cyber-physical world. The model can be seen as an privacy-oriented data disclosure ontology, allowing manual and automatic analysis of known and unknown privacy issues represented as special topological patterns on a directed graph. The model lays the theoretical foundation of software tools that can be used by individual users (i.e., data owners rather than organizations and researchers) themselves to monitor their data disclosure activities and help provide opportunities to adapt their behaviors towards a better trade-off between privacy protection and values gained through data disclosures.

The rest of the paper is organized as follows. Section 2 defines the proposed model in details. A number of case studies in two application categories are discussed in Section 3, in order to demonstrate how the proposed model can be used to identify different types of privacy issues. In Section 4, we discuss how automated semantic reasoning can be done based on the proposed model, which can be implemented with existing web ontology tools. Other related works and possible future directions are discussed in Sections 6 and 7.

2. The proposed model

In this section, we first give two example scenarios about privacy issues related to data disclosures, to illustrate what real-world problems the proposed model aims at solving. Then, we formally explain basic concepts behind the proposed graph model. Finally, we show how privacy issues can be studied by analyzing different types of edges in the proposed graph model.

2.1. Example scenarios

As stated, due to the increased connectivity and digitization of the modern society, users are facing the unprecedented challenge on data privacy. While using online and physical services, users are disclosing a large amount of personal data to different external entities, which include service providers (organizations) and other people. the proposed model aims at empowering users with more knowledge (i.e., awareness) on their data disclosure activities and automated tools to detect potential privacy issues that will be neglected otherwise. Thus, it is expected that the model can be used to help users make more informed data disclosure decisions in different scenarios such as the following ones.

Scenario 1: Data released to service providers. Alice uses different travel services to arrange her trip to China. She has to share certain personal information with almost all such services without a clear understanding of what organizations behind those services actually see the data. Due to propagation among service providers, she worries her data containing sensitive attributes may end up with some organizations she distrusts. What's more, particular combinations of attributes may cause identification. She would like to prevent that from happening.

Scenario 2: Data released to other people. Alice uses online social media nearly every day to record her life. She interacts with her family members, colleagues, friends and other people on Facebook, Twitter, and Instagram by sharing various contents. Now she is traveling in China and is eager to share the experience but her accurate positions (She is privacy cautious.). She worries the propagation of posts will make the landmark photos (shared on Instagram) and her real-time locations at the city or country level (shared on Facebook) viewed by the same people connected on different platforms. Besides, she wants to post travel-related contents with a group of people who are not on the working contact list. It will be helpful to have a tool monitoring data flows so that she can decide what to do in future.

2.2. The model: basic concepts

At a higher level of conceptualization, our proposed model can be formalized as a directed graph describing how personal data of people can *possibly* flow through (i.e., may be disclosed to) different types of entities in a cyber-physical world, as shown in Fig. 1.[1] Mathematically, such a graph can be denoted by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{\mathcal{V}_i\}_{i=1}^{M}$ is a set of M nodes and each node $\mathcal{V}_i$ represents a specific type of entities with the same semantic meaning in our model (depicted by ellipses), and $\mathcal{E} = \{\mathcal{E}_j\}_{j=1}^{N}$ is a set of N edges and each edge $\mathcal{E}_j$ represents a specific type of relations[2] between two entity types. Edges in $\mathcal{G}$ can be categorized into two different groups: edges representing *semantic relations* and edges representing *data flows* (depicted by solid and dashed arrows, respectively, in Fig. 1). Note that in Fig. 1, when there is "..." included in the textual label of an edge there should actually be multiple edges

[1] Names of edges in Fig. 1 are not actually part of the conceptual model. They are used for enhancing readability and for informing naming of predicates in Table 1. The dashed edges are numbered to help discuss data flows in the rest of the paper.

[2] Terminology wise, both "relation" and "relationship" are used in the research literature. We chose to use the word "relation" because it is the one used in Web Ontology Language (OWL), which we used to implement the automatic reasoning part of the model in Section 4.

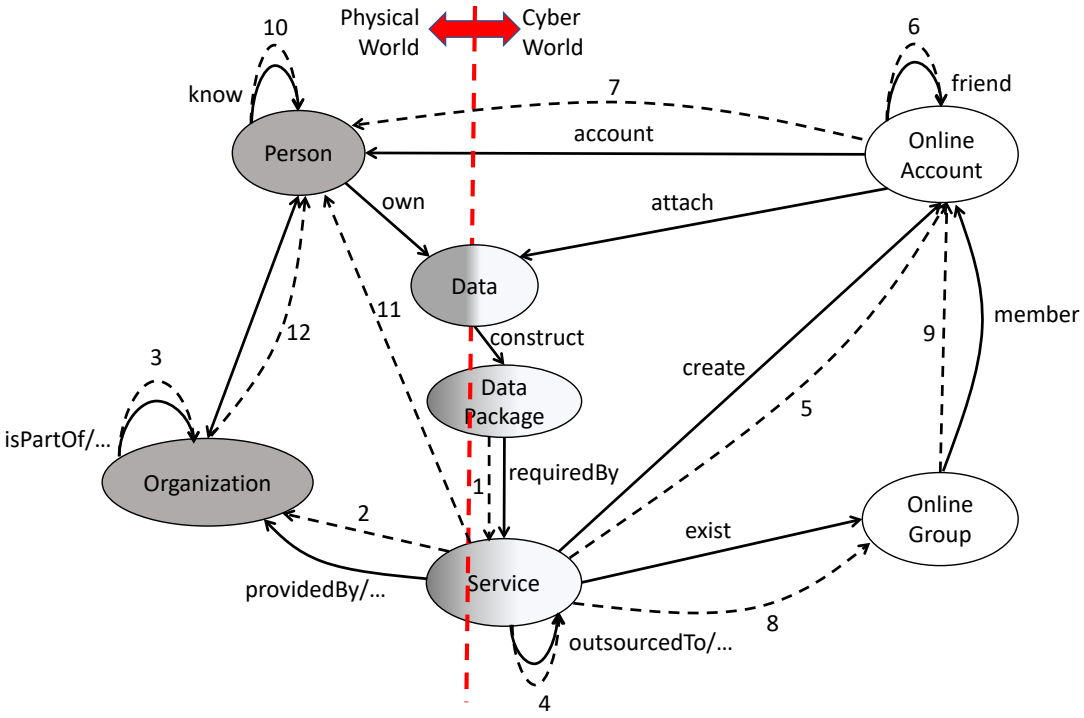

Figure 1: The entity-type graph of proposed model

(only one is shown for the sake of simplicity) due to the existence of multiple semantic relations between the two corresponding entity types (e.g., a service is provided by a company but owned by another, which have different implications on data flows). In our current model, we have $M = 7$ different entity types and a greater number of edge types between them[3].

The *entity type level* graph $\mathcal{G}$ can only show entity types and *possible* relations between different entities, but not the actual entities and relations (e.g., concrete data flows between two organizations/people) that are what we need to work with for detecting and analyzing privacy issues. To this end, we will need *entity level* graphs. Each of such graphs is a *different* directed graph $\boldsymbol{G} = (\mathbb{V}, \mathbb{E})$, where $\mathbb{V} = \{v | v \in \mathcal{V}_i, 1 \leq i \leq M\}$ is a set of nodes each representing an entity (i.e., an instance of a specific entity type / node in $\mathcal{G}$) and $\mathbb{E} = \{e | e \in \mathcal{E}_j, 1 \leq j \leq N\}$ is a set of edges each representing a relation (i.e., an instance of a specific relation type / edge in $\mathcal{G}$). Some concrete examples of such entity level models/graphs will be given in Section 3.

The entity types can be categorized into three groups: 1) physical entities that exist only in the physical world; 2) cyber entities that exist only in the cyber world (from user's perspective); 3) hybrid entities that may exist in both cyber and/or physical world. In Fig. 1, the 7 different entity types are colored differently to show which group(s) each entity type belongs to (gray: physical, white: cyber, gradient: hybrid). In the following we explain what these types represent.

Person (P) stands for natural people in the physical world. The model is *user-centric*, i.e., about a special P entity "me" – the user for whom the model is built. The model will include other people as well because privacy issues of "me" can occur due to data flows to other people who interact directly or indirectly with "me".

Data (D) refers to *atomic* data items about "me" (e.g., "my name"). Data entities may be by nature in the physical world, or in the cyber world, or in both worlds.

Service (S) refers to different physical and online services that serve people for a specific purpose (e.g., a travel agent helping people to book flights).

Data Package (DP) refers to specific combinations of data entities required by one or more services. In this model, DP entities can be seen as encapsulated data disclosed in a single transaction.

Organization (O) refers to organizations that relate to one or more services (e.g., service providers).

Online Account (OA) refers to "virtual identities" existing on online services. Note that even for physical services, there are often online accounts created automatically by the service providers to

[3]These numbers will change in enhanced versions of the model. See Section 5 for how the model can be possibly enhanced.

allow electronic processing and transmission of data, sometimes hidden from the users.

Online Group (OG) refers to "virtual groups" of online accounts that exist on a specific online service.

2.3. The model: edges

As stated before, each edge (i.e., relation type) in the entity level graph $\mathcal{G}$, and hence each edge (i.e., relation of a specific type) in an entity level graph $\boldsymbol{G}$, belongs to one of two groups of edges (relations). We explain these two edge groups in greater details below.

The first edge group is about **semantic relations** that may or may not relate directly to personal data flows. For instance, the edge connecting entity types P and D means that the special P entity "me" owns some personal data items. Unlike the second group of edges that can cause immediate privacy impacts, the first group of edges help modeling the "evidence" about how and why data may flow among these entities.

The second edge group is about **data flows** from a source entity to a destination entity. Most edges in this group are accompanied by semantic relation edges in the first group because the latter constructs the reason why a data flow can possibly occur.

To facilitate future discussions on data flows, we introduce a more loosely defined concept "data flow edge type" (and simply "edge type" when ambiguity or confusion will not arise) denoted by E_j, the set of *all* data flow edges between a specific pair of entity types labeled by the same number j in Fig. 1. Accordingly, we use $e_{j\text{-}k}$ to denote the k-th edge of the loose edge type E_j in an entity level graph $\boldsymbol{G}$, in order to give each individual edge in $\boldsymbol{G}$ a unique label. Note that E_j can cover multiple edges in $\mathcal{G}$ and $\boldsymbol{G}$ (e.g., data flows between S and O entities) and it conceptually differs from $\mathcal{E}_j$ as the latter refers to both Types 1 and 2 edges and also cover edges without a numeric edge label (e.g., edges between P and D entity types in Fig. 1).

The first data flow edge normally happens between DP and S entities, denoted by E_1. This is because before a data package is submitted to a service, no privacy issue can occur.The edge type E_{12} refers to potential *bidirectional* data flows between P and O entities, mapped to different types of semantic relations between P and O entities, e.g., a person owns a company. The edge types E_5 and E_8 refer to data flows from an S entity to an OA or an OG entity. The edge type E_7 refers to data flows from an OA to a P entity (i.e., a human user of an online account). The edge type E_{10} refers to data flows caused by social relationships among people (e.g., friendship and familial ties). The edge type E_{11} refers to data flows from an S entity directly to a person (i.e, not via an OA entity), e.g., a person can see public tweets on Twitter.

The relations and data flows represented by edges between people (P), services (S) and organizations (O) can be complicated in real world due to the complexity of how the business world works. Particularly, in Fig. 1 for each edge (between S and O, from S to S and from O to O) there can be multiple different semantic relations and data flows, e.g., a service is provided by an organization (i.e., a service provider), a service is *outsourced to*, *supplied by* or *powered by* another service, an organization is *part of*, *in partnership with* or *invested by* another organization. In this work we do not intend to cover a complete list of such complicated business relations, but focus on the conceptual abstraction needed to capture all such relations. Given semantic relation edges defined on OS and SP entities, identified data flows include:

- E2: (S, O) flows from S to O entities due to the existence of semantic relation edges *providedBy* in between.
- E3: (O, O) flows between O entities given the fact that one O entity has some relation with another, e.g., *isPartOf*, *invest* or *collabrateWith*.
- E4: (S, S) flows between S entities due to data sharing relations between them, e.g., *suppliedBy*, *poweredBy* or *outsourcedTo*.

Due to the data collection by service providers, it may be the case that data flows to the physical scope and lose effective control. This is undoubtedly a challenge to all of stakeholders on preserving user privacy in both physical and cyber spaces.

Unlike privacy issues caused by data collection activities of services, privacy issues of online communities (such as OSNs) are mostly related to how well users manage the visibility of personal data [5]. For instance, with "friends only" and "members only" as privacy settings, contents shared on private spaces can be viewed by friends and group members only. In our proposed model, the edges between OA, OG and P entities ($E_5, \ldots, E_{10}$) describe how personal data can possibly flow among such entities. Such data flow edges are cased by semantic relations, e.g., a person has access to an online account, an online account is befriended with another account, a person is a friend of another person, an online account is a member of an online group.

Given semantic relations defined for OA and OG entities, identified data flows include:

- E5: (S, OA) flows from S to OA entities due to the existence of semantic relation edges *create* in between.
- E6: (OA, OA) flows between OA entities given the fact that one online account is the *friend* of the other.
- E7: (OA, P) flows from OA to P entities due to the existence of semantic relation edges *account* in between.
- E8: (S, OG) flows from S to OG entities due to the semantic relation edges *exist* in between.
- E9: (S, P) flows from S to P entities is due to a service platform providing public data sets.
- E10: (P, P) flows between P entities due to the existence of semantic relation edges *know* in between.

2.4. "Topological" privacy issues

For a given user "me", if we can construct an *entity level* graph $\boldsymbol{G}$, which shows relevant entities, semantic relations and data flows, we will be able to study a number of different types of privacy issues concerning this given user, e.g., if the user is disclosing too much information to a single service or organization, if the user has disclosed too much personal information to other people or the general public. Even when the graph $\boldsymbol{G}$ is incomplete, which is likely the case for most scenarios due to the lack of complete details about the user, some privacy issues may still be identified.

Within the proposed model, we can define an important concept: a "data-flow path" is a sequence of consecutive data flows (edges in an entity level graph $\boldsymbol{G}$). This concept allows us to map different "privacy issues" to certain *topological* patterns that are formed by one or more data-flow paths. Different privacy issues may share the same topological pattern but follow different edges or different edge types, e.g., one privacy issue may be related to one organization while another to a different organization. Beyond using the model to detect privacy issues, we can also try to quantify the risk of a given privacy issue and provide possible solutions to the user. Some concrete examples about such privacy issues will be discussed in the next section with a number of imaginary but realistic case studies. In addition to investigating privacy issues, it deserves mentioning that the proposed model can also find applications in other contexts, e.g., studying how personal data are consumed by online services (even if there are no privacy issue for any particular user).

3. Case studies

In this section, we use realistic examples in two broad categories to illustrate how entity level graphs can be built based on our proposed model and how privacy issues can be possibly identified.

3.1. Privacy issues related to service providers

Figure 2 shows the simplest model involving S and O entities: an online service <service 1> connects to a service provider <provider 1> by semantic relation edge *providedBy*, denoted by *providedBy*(service 1, provider 1). For instance, an E_1 flow $e_{1\text{-}1}$ at the beginning could cause an E_2 flow $e_{2\text{-}1}$ from <service 1> to <provider 1>, denoted as $e_{1\text{-}1}$(item 1, service 1) and $e_{2\text{-}1}$(service 1, provider 1) respectively. As a result, there is only one path $p_1 = (e_{1\text{-}1}, e_{2\text{-}1})$ found from the source data <item 1> to the service provider <provider 1> in the physical world[4]. Such a simple path does not normally lead to any privacy issue since it merely describes what data items are needed for a service to happen. In the following examples, we will show how non-trivial real privacy issues can be identified on more complicated data flow graphs.

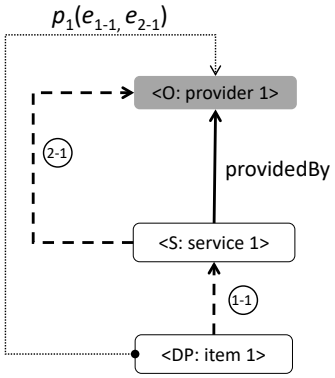

Figure 2: Example entity graph showing a data flow

In real world, data flows can take place within a corporate family (connected by the semantic relation *isPartOf*). Therefore, it may be the case that different data items flow among multiple service providers and aggregate at a single organization, which may be unknown to the user thus leading to a privacy issue. For instance, in Fig. 3, as <item 1> and <item 2> flow to <service 1> and <service 2> separately, E_2 flows $e_{2\text{-}1}$(service 1, provider 1) and $e_{2\text{-}2}$(service 2, provider 2) take place. Then, E_3 flows follow such as

[4]The path is shown as a dotted line in Fig. 2 from the source to the destination, ignoring the entities in the middle. The same hereinafter for other figures.

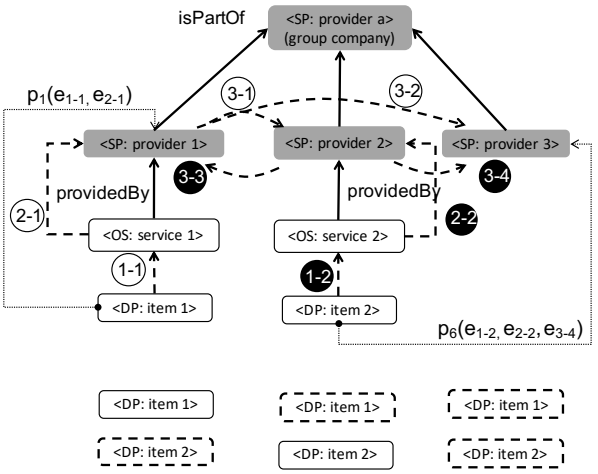

Figure 3: Entity graph in provider hierarchies

$e_{3\text{-}1}$(provider 1, provider 2), $e_{3\text{-}2}$(provider 1, provider 3), $e_{3\text{-}3}$(provider 2, provider 1) and $e_{3\text{-}4}$(provider 2, provider 3). Similarly, paths can be found from data packages <item 1> and <item 2> to service providers, <provider 1>, <provider 2> and <provider 3>, such as $p_1 = (e_{1\text{-}1}, e_{2\text{-}1})$ and $p_6 = (e_{1\text{-}2}, e_{2\text{-}2}, e_{3\text{-}4})$. Here we use black and white edge labels to distinguish flows about different data packages containing two different data items. Inspecting the data flow graph, we see both data packages flow to the organization <provider a>, which may cause unknown disclosure of personal data.

Complex business models exist in the real world. Figure 4 shows data flows among some business partners who jointly support online services. As shown in Fig. 4a), an E_4 data flow $e_{4\text{-}1}$(service a, service b) can be found among the business partners connected by an *outsourcedTo* semantic relation edge. Based on an E_2 flow $e_{2\text{-}1}$(service b, provider b) and the service ownership expressed with the semantic relation edge *belongTo*, an E_3 flow $e_{3\text{-}1}$(provider b, provider a) can be identified. Similarly, Figure 4b) shows E_4 flows that would incur due to the semantic relation edge *poweredBy* between online services, e.g., $e_{4\text{-}1}$(service a, service 1) and $e_{4\text{-}2}$(service a, service 2), while in Fig. 4c), the only E_4 flow $e_{4\text{-}1}$(service 1, service 2) is due to the semantic relation edge *suppliedBy* in between. If any of business relations between S and O entities are unknown, privacy concerns can arise.

To further illustrate how data flows in an entity level graph can be used to identify privacy issues, Figure 5 shows a scenario where a customer (a P entity) books flight tickets and hotels via online services provided by organizations Booking.com and Agoda. Privacy restrictions may be given to data items on pre-defined labels, such as *sensitive data items are not allowed to share with more than 5 organizations*. For this purpose, data entities are categorized in the following groups: **Profile** (Name, Age, Gender, and Email), **Event** (Itinerary, Companion, Dates, and Spending), **Location** (Destination, Landmark), **Sensitive** (Health), and **Entertainment** (Tour, Food). Sensitive data such as medical certificates may be required and shared with third-party suppliers, in case travelers need special medical assistance during travel. As a result, data package <item 1> will flow to eleven service providers along with paths p_1 to p_{11}. For instance, paths $p_1 = (e_{1\text{-}1}, e_{4\text{-}1}, e_{2\text{-}1})$, $p_2 = (e_{1\text{-}1}, e_{4\text{-}1}, e_{2\text{-}1}, e_{3\text{-}1})$ and $p_{10} = (e_{1\text{-}1}, e_{4\text{-}1}, e_{2\text{-}1}, e_{3\text{-}9})$ can respectively lead data package <item 1> to <GoToGate>, <Booking> and <SuperSaver>. Besides, the Agoda hotel booking service may incur data flows to seven service providers (led by paths p_{12} to p_{18}), such as $p_{12} = (e_{1\text{-}2}, e_{2\text{-}2})$ and $p_{13} = (e_{1\text{-}2}, e_{2\text{-}2}, e_{3\text{-}11})$ running to <Agoda> and <Kayak>. This may cause location privacy leakage if an O entity has the access to the user's <name> and <destination> simultaneously.

3.2. Unwanted disclosures to other people

In addition to privacy issues raised from data collection by service providers and data shared among services and organizations, online privacy issues may also be caused by unwanted data disclosures to other people e.g. on OSNs. Figure 6 is an entity level graph showing how the P entity <me> connects with other people through online and offline relations. Based on the friend relations between <fb_abc> and <ig_abc>, E_6 data flows such as $e_{6\text{-}4}$(fb_abc, fb_edward), $e_{6\text{-}5}$(ig_abc, ig_ed1989) could take place in the cyber space when "I" use Facebook and Instagram services and generate data flows $e_{1\text{-}1}$, $e_{5\text{-}1}$, $e_{1\text{-}2}$ and $e_{5\text{-}2}$. Given the account ownership, E_7 flows such as $e_{7\text{-}4}$(fb_edward, edward) and $e_{7\text{-}5}$(ig_ed1989, edward) will follow. Along with paths $p_4 = (e_{1\text{-}1}, e_{5\text{-}1}, e_{6\text{-}4}, e_{7\text{-}4})$ and $p_5 = (e_{1\text{-}2}, e_{5\text{-}2}, e_{6\text{-}5}, e_{7\text{-}5})$, it shows that both data packages <item 1> and <item 2> will be disclosed to <edward>. Therefore, "my" current location may be inferred from the itinerary post on Facebook and landmark photos shared on Instagram during the trip.

Data visibility can be managed by privacy policies related to online friendships and memberships. As a result, privacy leakage could be caused when "I" permit unwanted access requests. Figure 7 shows a scenario where online data are propagated across groups that have members in common. Through E_9 flows $e_{9\text{-}1}$(fb_travel, fb_alice) and $e_{9\text{-}2}$(fb_travel, fb_bob), Alice and Bob can view <item 3> once "I" send it to the travel group. In some situations, <item

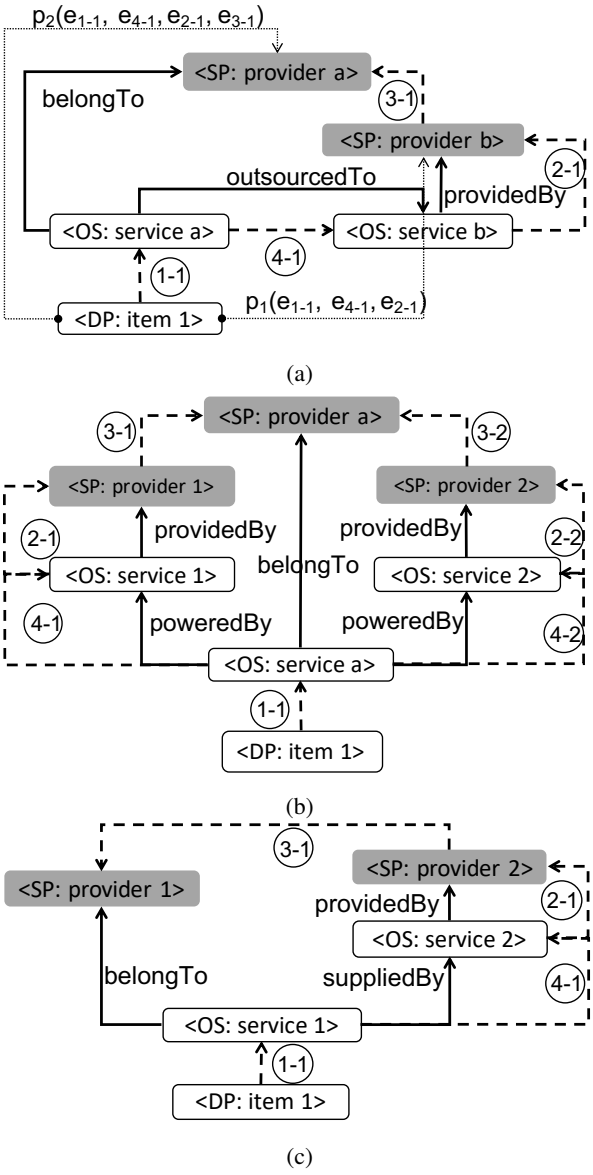

Figure 4: Example entity graphs of supply chains

3> can be resent to other groups and cause the E_8 flows, such a $e_{8\text{-}2}$(fb_bob, fb_writing) and $e_{8\text{-}3}$(fb_carol, fb_work). Through the following E_9 and E_7 flows, <item 3> may be disclosed wrong people through $p_4 = (e_{1\text{-}3}, e_{8\text{-}3}, e_{9\text{-}4}, e_{7\text{-}4})$.

4. Automated reasoning of privacy issues

Web ontology language (OWL) and semantic web rule language (SWRL) are widely utilized in specifying security and privacy policy constraints on data usage [6–9]. In this section, we use OWL and SWRL to formalize our model and show how reasoning can be done to detect privacy issues *automatically*. For the sake of simplicity, in this section we will focus on a subset of the entity types and relations. We will also focus on only online services (OS) and service providers (SP), so will use OS for services (S) and SP for organizations (O).

4.1. Semantic formalization

Following OWL and SWRL, different components in the proposed model can be defined as classes, predicates (with domains and values) and instances, as shown in Table 1. With the ontology and semantic rules (Rules 1-10) developed in Protégé 4.0 we can implement an automated semantic reasoning engine. Through running the reasoner Pellet [10] and description logic (DL) queries [11] on the knowledge base, implicit relations (i.e., data flows) could be identified for privacy assessment and decision making purposes. Assuming that data flows to physical entities are likely causing privacy issues, privacy questions can be made to look for *finalFlowTo* (or *access*) in the result sets.

In dealing with scenarios related to service providers, DL queries are utilized to answer the following questions: *"where the sensitive information flows to?"* and *"who can access the user profile and location at the same time?"* Through reasoning on the semantic graph of Fig. 5, the engine shows that the number of service providers can be reduced by changing <flight_booking> to <flight_agoda> as the sensitive item <item 1> will be shared with one single corporate group, as shown in Fig. 8. In a scenario about purchasing travel service packages, Figure 9 shows the result of comparing two service packages by running queries to answer *"who can access the user profile and location at the same time?"* Given the demand for booking "flights + hotels", the result sets show that adopting Package 2 can better control the privacy risks. In this case, query services can enhance user privacy by splitting personal details contained in data flows.

Towards the privacy requirements in the scenarios concerning unwanted data disclosures to other people, DL queries can be applied to check things such as if someone else can access certain data combinations or if entertainment-related messages are disclosed to colleagues. As illustrated in Fig. 10, through querying on recipients *who can access two data types during the same period*, the system is expected to provide privacy suggestions such as *blocking Facebook account fb_edward so as to stop such disclosure to Edward in the real world* (see Fig. 6). Similarly, a DL query can be made to check if certain data will flow to unwanted groups (recipients). As shown in Fig. 11, it shows <item 3> has breached personal privacy and thus demands

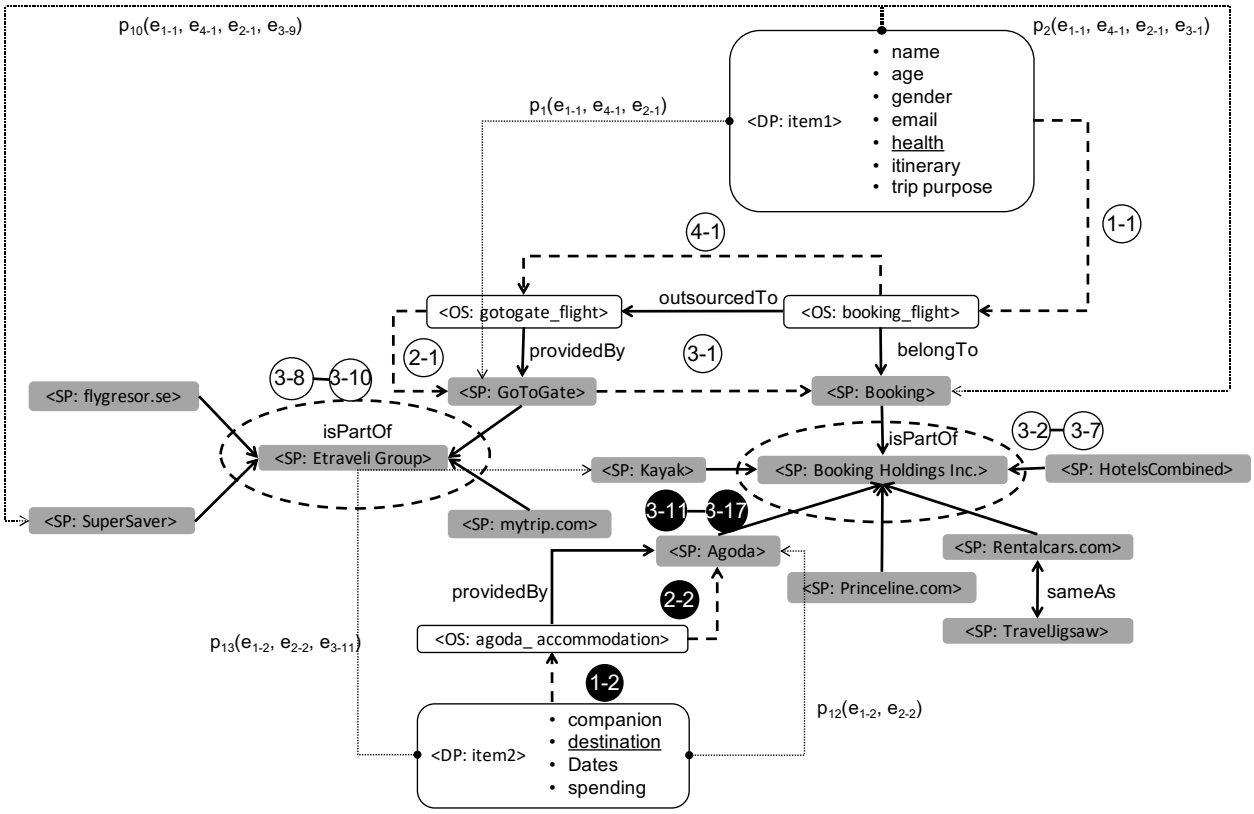

Figure 5: An example entity graph about data sharing in the travel context

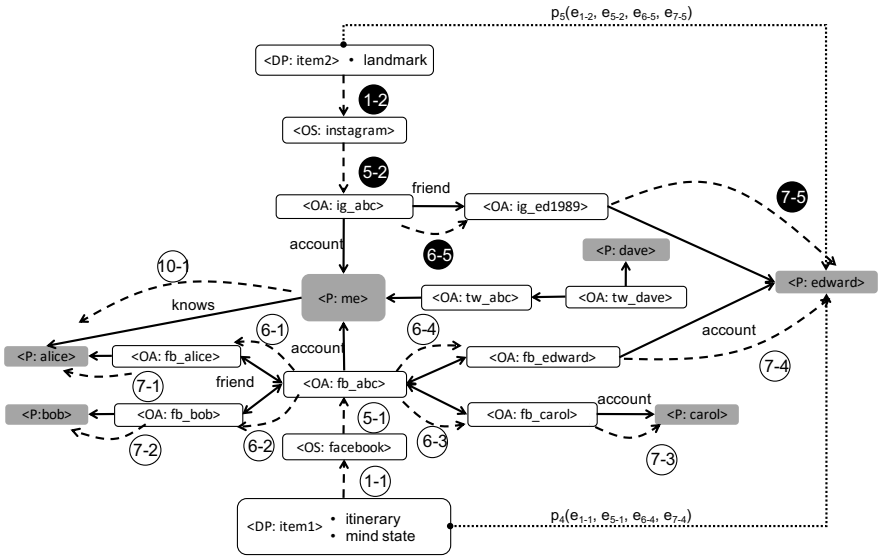

Figure 6: An example entity graph showing unwanted data disclosure on OSNs

for extra modification, like removing entertainment information from the Facebook post to <fb_travel>.

1. DP(?d), flowTo(?d, ?s), providedBy(?s, ?p) →finalFlowTo(?d, ?p)

2. DP(?d), flowTo(?d, ?s), outsourcedTo(?s, ?s1), providedBy(?s1, ?p) →finalFlowTo(?d, ?p)

Table 1: Definitions of classes, predicates and instances to represent different components of the proposed model

Class (Domain)	Predicate	Range	Instance
Data_Package(DP)	*flowTo*	OA, OG, OS	item1, item2, item3, ...
	finalFlowTo	P, SP	
	has	D	
Data(D)	*construct* ($\leftrightarrow$ *has*)	DP	itinerary, email, name, date_of_birth, ...
Online_Account(OA)	*account*	P	fb_alice, tw_dave, ig_ed1989, ...
	friend	OA	
Online_Group(OG)	*member*	OA	fb_travel, fb_writing, fb_work, ...
Online_Service(OS)	*belongTo*	SP	flight_booking, accommodation_agoda, facebook, twitter, instagram, ...
	providedBy	SP	
	outsourcedTo	OS	
	poweredBy	OS	
	suppliedBy	OS	
	create	OA	
	exist	OG	
Service_Provider(SP)	*isPartOf*	SP	Booking, Agoda, TripAdvisor, ...
	access ($\leftrightarrow$ *finalFlowTo*)	DP	
Person(P)	*know*	P	alice, bob, me, dave, edward, ...
	access ($\leftrightarrow$ *finalFlowTo*)	DP	

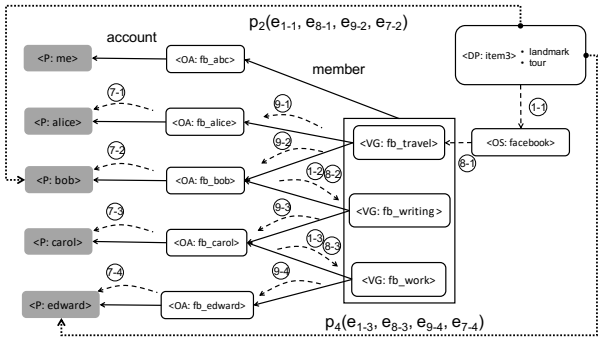

Figure 7: Entity graph of cross-group data disclosure

3. DP(?d), flowTo(?d, ?s), poweredBy(?s, ?s1), providedBy(?s1, ?p) $\rightarrow$finalFlowTo(?d, ?p)
4. DP(?d), flowTo(?d, ?s), suppliedBy(?s, ?s1), providedBy(?s1, ?p) $\rightarrow$ finalFlowTo(?d, ?p)
5. SP(?p), isPartOf(?p, ?q), isPartOf(?r, ?q), finalFlowTo(?d, ?p) $\rightarrow$ finalFlowTo(?d, ?r)
6. DP(?d), flowTo(?d, ?s), finalFlowTo(?d, ?p1), belongTo(?s, ?p) $\rightarrow$ finalFlowTo(?d, ?p)
7. DP(?d), flowTo(?d, ?a), account(?a, ?p) $\rightarrow$finalFlowTo(?d, ?p)
8. DP(?d), finalFlowTo(?d, ?p), know(?p, ?p1) $\rightarrow$finalFlowTo(?d, ?p1)
9. DP(?d), flowTo(?d, ?s), create(?s, ?a), friend(?a, ?a1) $\rightarrow$flowTo(?d, ?a1)
10. DP(?d), flowTo(?d, ?g), member(?g, ?a) $\rightarrow$flowTo(?d, ?a)

5. More discussion and future work

The proposed model is generic enough to cover a wide range of applications and privacy issues. There are a number of key areas for further development of the proposed model, which we leave as our future work.

More entity types and relations. Our proposed model currently covers 7 entity types and relations between them. There are other entity types we may add, e.g., physical groups of people and groups of organizations.

More complicated business models. As mentioned before, the business world is actually very complicated and we have considered only some simple business relations between services and organizations. Therefore, graphs should be built based on more complicated real-world business models and related data flows.

More complicated inter-personal relations. Similar to the above, there can be more complicated relationships among people as well. Therefore, current relations to person (P) entities need to be refined to capture more semantic information from real-world human relations, e.g., family, friends, colleagues, carers. According to the semantic relations between P entities, data flows can be differentiated by quantities and thus improve the accuracy of detection results. For instance, to avoid potential privacy issues caused by other people, the central user can exchange recorded data flows with "friends" to see if s/he has overly disclosed data to them.

More complicated data structures. Our current

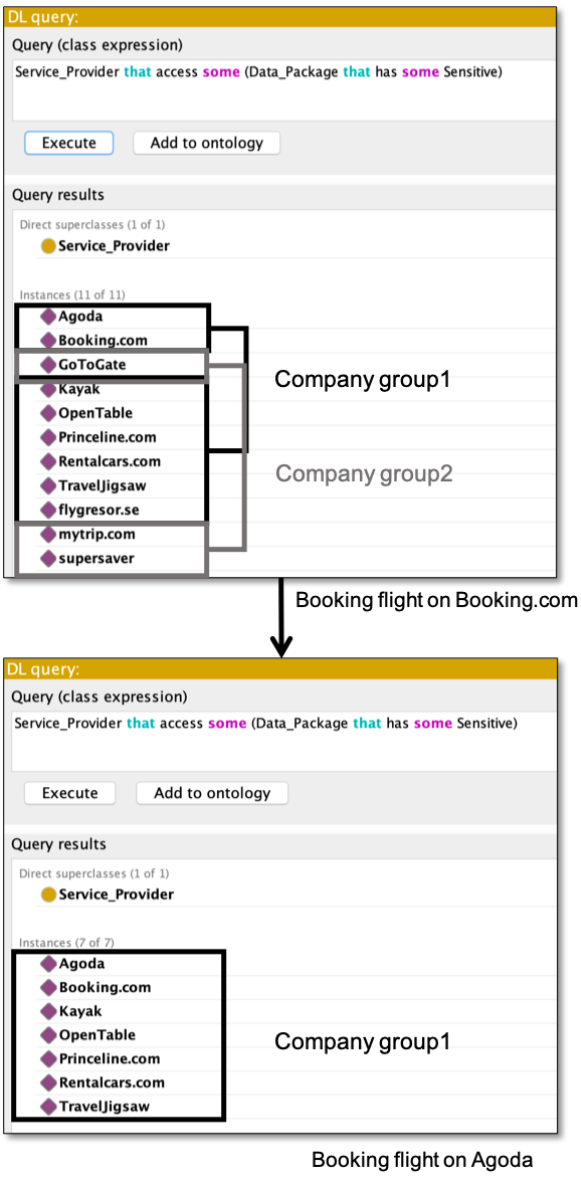

Figure 8: Example query on sensitive data disclosures

model abstracts data using Data (D) and Data Package (DP) entity types related with *construct*. In reality, many data entities often include complicated attributes, which may be important for analysing privacy issues as well. For instance, a travel itinerary contains multiple destinations visited at specific times, transportation types, points of interest, etc. Similar issues exist in email, date of birthday, etc.

Invisible or implicit data flows. This work mainly focuses on data flows caused by visible data sharing, i.e., all data flows are explicit and visible to the user concerned. However, it is necessary to monitor invisible or implicit data disclosures that can happen without

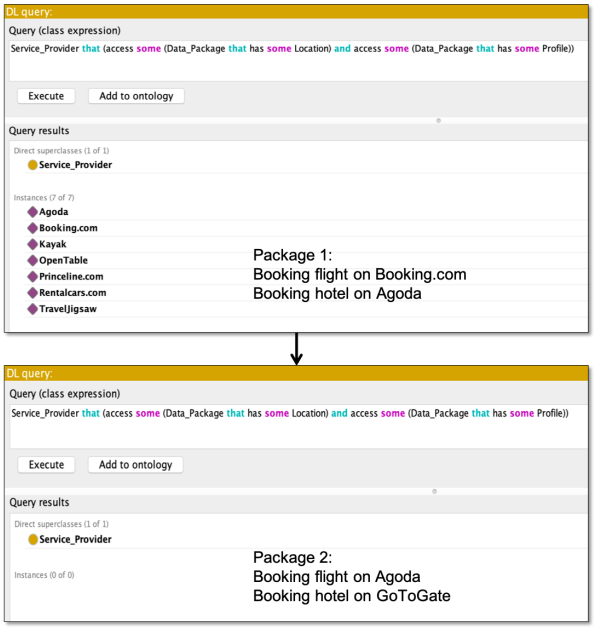

Figure 9: Example query on combined data disclosures

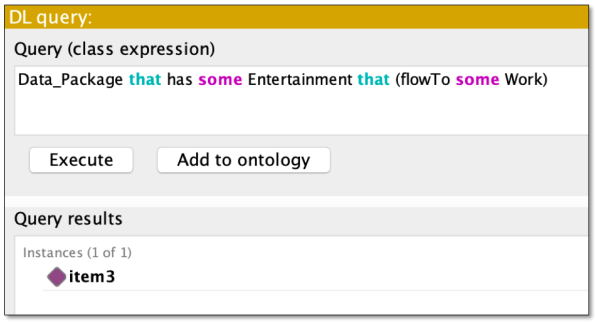

Figure 10: Example query on unintended disclosures

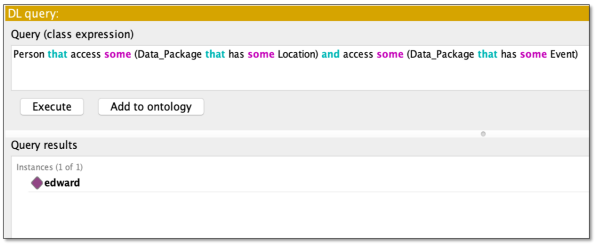

Figure 11: Example query regarding common recipients on OSNs

users' explicit knowledge. For instance, a user's IP address is often disclosed to service providers without a separate explicit notice, which however can be captured as invisible and implicit data flows by extending our model.

More explicit benefit returns. The proposed model

focuses on data flows and privacy issues only. However, it is well-known in the literature that a privacy paradox exists, i.e., a trade-off between privacy and utility to be considered in real world, if a privacy issue is considered in real-world cases. The proposed model implicitly covers some benefits, e.g., disclosing data to a service has a defined aim to get a desired service by return. However, more quantitative and explicit benefit/value returns can be added to allow consider privacy issues in a more contextualized manner and to do better reasoning.

Legal framework for data protection and privacy laws. The proposed model can be further enhanced by including a legal framework regarding legality, consequences and users' rights as data subjects. This can be added as attributes and constraints to data flows and relations. There has been some related work on formalizing such legal frameworks, e.g., on the new EU GDPR (General Data Protection Regulations) [12].

Connecting multiple models together. As a user-centric model, the entity level graph has a special entity "me" at the centre of everything. Given a number of users, it is possible to connect their user-centric graphs to form a larger graph showing how privacy issues change from person to person, which will help study larger-scale privacy issues, e.g., how privacy issues of one user propagate to his/her fiends on OSNs.

There are also some useful tools that will make it easier to use the proposed model. We give some examples below.

Automatic and dynamic building of the entity level graph. The cyber-physical system (CPS) is not static itself and by nature large scale. Therefore, it is unsuitable to build semantic-based graphs by manual generation. To better manage the ever-changing world and to cover as many relevant information as possible, an automation tool is necessary for building large-scale graphs where data flows are produced all the time.

Interactive visualization of data flow paths and privacy issues. As mentioned before, each privacy issue can be represented by a specific topological pattern involving one or more data flow paths. It will be helpful to develop some visualization tools to show such paths and topological patterns, possibly with animation.

Automatic comparison of data disclosure options. Given an data flow graph and a number of options for data disclosure, we can automatically compare all such options to compare them and determine which options provide better privacy protection. After benefit/value returns are added, such comparison can be done to balance two main objectives: privacy and utility.

Automatic discovery of OSN accounts that belong to the same organization or individual. More potential privacy issues can be detected if we have more information about the physical entities (organizations or people) behind OSN accounts. Some automatic tools can be developed to detect OSN accounts belonging to the same organization or person to allow a more complete data flow graph related to such accounts, therefore exposing more potential privacy issues.

As part of the research project PriVELT (`https://privelt.ac.uk/`) that made the reported work possible, we will also try to incorporate the proposed conceptual model into a user-centric framework for providing privacy protection and value enhancement for leisure travelers.

6. Related work

The most related area is privacy ontologies, which often involve a graph-based model. Most work on this topic mainly focuses on specifying conditions of data access by the controllers. For instance, ontological models can be built to incorporate privacy causes, impacts and contextual factors. Sacco and Passant (2011) proposed a privacy preference ontology (PPO) to allow users specify fine-grained conditions of using of their RDF data [13]. To effectively combine data (or knowledge) of different sources in the cyber security domain, a knowledge graph STUCCO was built up with data from 13 structured sources [14]. To ensure privacy criteria of different stakeholders are properly implemented, Kost et al. integrated an ontology into privacy policy specifications and the evaluation of privacy constraints [15]. Michael et al. proposed a privacy ontology to support the provision of privacy and derive the privacy levels associated with e-commerce transactions and applications [3]. To guarantee business processes are performed securely, Ioana et al. designed a semantic annotation tool to assist users in specifying security and privacy constraints onto different business process models [16]. As far as we know, no existing ontologies consider how likely privacy issues are caused from user-centric data flows like we report in this paper.

Reasoning from background knowledge on human relationships, content types and contextual factors can support decision making on authorization and privacy preservation. Passant et al. [17] utilized semantic vocabularies such as FOAF (friend of a friend) and SIOC (Semantically Interlinked Online Communities) to establish a trust and privacy layer to restrict publishing, sharing or browsing data by various social behaviors. By categorizing privacy violations of OSNs as endogenous and exogenous information disclosures in a direct or an indirect way, an agent-based representation was proposed based on users' privacy requirements on their generated contents

[18]. Considering that limited privacy requirements can be expressed through access control policies, semantic data models have been suggested to assist in authorization to reduce leakage risks [19]. To anonymize e-health records with statistical disclosure control (SDC) methods, the healthcare terminology SNOMED CT[5] was incorporated into a privacy ontology to mask categorical attributes and preserve information utility [20]. To help designers understand security mechanisms and how well they are aligned with corporate missions, the ontology is also modelled about information systems and settings on permission, delegation, and trust at the organizational level [21].

Another closely related research area is OSN (structural) anonymity. Focusing on OSN data protection, Qian et al. [22] proposed individual network snapshots. In case sensitive attributes are inferred by attackers, distance between published data and background knowledge needs to be controlled in a safe range. Noticing that anonymized graphs may incur identification attacks, Peng et al. [23] developed a two-staged algorithm: constructing a sub-graph of users (seed) and connecting to the rest (grow) to show the feasibility. User similarities are shared among "neighbors". As a result, knowing neighbor nodes and attached attributes can increase the probability of identification central users [24]. In addition to static relations, "contact graphs" are formalized with contextual factors in mobility [25]. Similarly, graph representations storing user interactions over OSNs should be protected against privacy attacks [26]. Singh and Zhan analyzed the vulnerability to identity attacks based on topological properties [27]. Instead of modeling network graphs, Li et al. converted tabular data in data graphs, including original datasets, anonymity datasets and background knowledge of attackers [28]. Instead of direct anonymity on graphs, our goal is to offer users a knowledge graph about data flows to reflect their activities in the wider business world (online and offline). Since our approach effectively combines the ontological formalization about data flows, graph-based structures of service providers and people as well as a knowledge base with semantic meanings to support automatic reasoning on potential issues individual users care about, we believe that this model can support further development of user-centric privacy-enhancement applications on personal devices, for the purposes such as monitoring data-related activities through different mobile apps.

[5] http://www.snomed.org/snomed-ct/five-step-briefing

7. Conclusions

In this paper, we propose a user-centric, graph-based semantic model to identify data flows produced from a given user's online and offline activities that can potentially lead to privacy issues. In the conceptual model, privacy issues concerning the given user can be represented as specific topological patterns involving one or more data-flow paths. The model is generic enough to be applied to a wider range of scenarios, some of which were given in this paper to illustrate how it can be used. We also demonstrate that the model can be easily implemented using OWL tools to enable automatic semantic reasoning of privacy issues.

Acknowledgements

The authors' work was supported by the research project, PRIvacy-aware personal data management and Value Enhancement for Leisure Travellers (PriVELT, https://privelt.ac.uk/), funded by the EPSRC (Engineering and Physical Sciences Research Council)in the UK, under grant number EP/R033749/1.

References

[1] J. Ge, J. Peng, and Z. Chen, "Your privacy information are leaking when you surfing on the social networks: A survey of the degree of online self-disclosure (DOSD)," in *Proceedings of 2014 IEEE 13th International Conference on Cognitive Informatics and Cognitive Computing*, pp. 329–336, IEEE, 2014.

[2] H. Krasnova, E. Kolesnikova, and O. Guenther, ""It won't happen to me!": Self-disclosure in online social networks," in *Proceedings of 15th Americas Conference on Information Systems*, pp. 343–354, AISeL, 2009.

[3] M. Hecker, T. S. Dillon, and E. Chang, "Privacy ontology support for e-commerce," *IEEE Internet Computing*, vol. 12, no. 2, pp. 54–61, 2008.

[4] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, and Y. Agarwal, "Your location has been shared 5,398 times! A field study on mobile app privacy nudging," in *Proceedings of 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 787–796, ACM, 2015.

[5] H. Hu, G.-J. Ahn, and J. Jorgensen, "Multiparty access control for online social networks: Model and mechanisms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1614–1627, 2013.

[6] T. Finin, A. Joshi, L. Kagal, J. Niu, R. Sandhu, W. Winsborough, and B. Thuraisingham, "R*OWL*BAC - representing role based access control in *OWL*," in *Proceedings of 13th ACM Symposium on Access Control Models and Technologies*, pp. 73–82, ACM, 2008.

[7] H. Muhleisen, M. Kost, and J.-C. Freytag, "SWRL-based access policies for linked data," in *Proceedings of 2nd Workshop on Trust and Privacy on the Social and Semantic Web*, 2010.

[8] Y. Lu and R. O. Sinnott, "Semantic security for e-health: A case study in enhanced access control," in *Proceedings of 2015 IEEE 12th International Conference on Autonomic and Trusted Computing*, pp. 407–414, IEEE, 2015.

[9] Y. Lu and R. O. Sinnott, "Semantic-based privacy protection of electronic health records for collaborative research," in

Proceedings of 15th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pp. 519–526, IEEE, 2016.

[10] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz, "Pellet: A practical OWL-DL reasoner," *Web Semantics: science, services and agents on the World Wide Web*, vol. 5, no. 2, pp. 51–53, 2007.

[11] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati, "DL-Lite: Tractable description logics for ontologies," in *Proceedings of 20th National Conference on Artificial Intelligence*, vol. 5, pp. 602–607, AAAI, 2005.

[12] C. Bartolini and L. Robaldo, "PrOnto: Privacy ontology for legal reasoning," in *Electronic Government and the Information Systems Perspective: 7th International Conference, EGOVIS 2018, Regensburg, Germany, September 3–5, 2018, Proceedings*, pp. 139–152, Springer, 2018.

[13] O. Sacco and A. Passant, "A privacy preference ontology (PPO) for linked data," in *Proceedings of WWW 2011 Workshop on Linked Data on the Web*, 2011.

[14] M. D. Iannacone, S. Bohn, G. Nakamura, J. Gerth, K. M. Huffer, R. A. Bridges, E. M. Ferragut, and J. R. Goodall, "Developing an ontology for cyber security knowledge graphs," in *Proceedings of 10th Annual Cyber and Information Security Research Conference*, 2015.

[15] M. Kost, J.-C. Freytag, F. Kargl, and A. Kung, "Privacy verification using ontologies," in *Proceedings of 2011 6th International Conference on Availability, Reliability and Security*, pp. 627–632, IEEE, 2011.

[16] I. Ciuciu, G. Zhao, J. Mülle, S. von Stackelberg, C. Vasquez, T. Haberecht, R. Meersman, and K. Böhm, "Semantic support for security-annotated business process models," in *Enterprise, Business-Process and Information Systems Modeling: 12th International Conference, BPMDS 2011*, pp. 284–298, Springer, 2011.

[17] A. Passant, P. Kärger, M. Hausenblas, D. Olmedilla, A. Polleres, and S. Decker, "Enabling trust and privacy on the social web," in *Proceedings of W3C Workshop on the Future of Social Networking*, pp. 15–16, W3C, 2009.

[18] N. Kökciyan and P. Yolum, "PriGuard: A semantic approach to detect privacy violations in online social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2724–2737, 2016.

[19] F. Paci and N. Zannone, "Preventing information inference in access control," in *Proceedings of 20th ACM Symposium on Access Control Models and Technologies*, pp. 87–97, ACM, 2015.

[20] S. Martínez, D. Sánchez, and A. Valls, "A semantic framework to protect the privacy of electronic health records with non-numerical attributes," *Journal of Biomedical Informatics*, vol. 46, no. 2, pp. 294–303, 2013.

[21] F. Massacci, J. Mylopoulos, and N. Zannone, "An ontology for secure socio-technical systems," in *Handbook of Ontologies for Business Interaction*, pp. 188–206, IGI Global, 2008.

[22] J. Qian, X.-Y. Li, C. Zhang, L. Chen, T. Jung, and J. Han, "Social network de-anonymization and privacy inference with knowledge graph model," *IEEE Transactions on Dependable and Secure Computing*, 2017.

[23] W. Peng, F. Li, X. Zou, and J. Wu, "A two-stage deanonymization attack against anonymized social networks," *IEEE Transactions on Computers*, vol. 63, no. 2, pp. 290–303, 2014.

[24] B. Zhou and J. Pei, "Preserving privacy in social networks against neighborhood attacks," in *Proceedings of 2008 IEEE 24th International Conference on Data Engineering*, pp. 506–515, IEEE, 2008.

[25] M. Srivatsa and M. Hicks, "Deanonymizing mobility traces: Using social network as a side-channel," in *Proceedings of 2012 ACM Conference on Computer and Communications Security*, pp. 628–637, ACM, 2012.

[26] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava, "Class-based graph anonymization for social network data," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 766–777, 2009.

[27] L. Singh and J. Zhan, "Measuring topological anonymity in social networks," in *Proceedings of 2007 IEEE International Conference on Granular Computing*, pp. 770–774, IEEE, 2007.

[28] X.-Y. Li, C. Zhang, T. Jung, J. Qian, and L. Chen, "Graph-based privacy-preserving data publication," in *Proceedings of 2016 35th Annual IEEE International Conference on Computer Communications*, IEEE, 2016.