EURASIP Journal on Information Security
a SpringerOpen Journal

RESEARCH Open Access

Taxonomy of social network data types

Christian Richthammer*†, Michael Netter†, Moritz Riesner†, Johannes Sänger† and Günther Pernul†

Abstract

Online social networks (OSNs) have become an integral part of social interaction and communication between people. Reasons include the ubiquity of OSNs that is offered through mobile devices and the possibility to bridge spatial and temporal communication boundaries. However, several researchers have raised privacy concerns due to the large amount of user data shared on OSNs. Yet, despite the large body of research addressing OSN privacy issues, little differentiation of data types on social network sites is made and a generally accepted classification and terminology for such data is missing. The lack of a terminology impedes comparability of related work and discussions among researchers, especially in the case of privacy implications of different data types. To overcome these shortcomings, this paper develops a well-founded terminology based on a thorough literature analysis and a conceptualization of typical OSN user activities. The terminology is organized hierarchically resulting in a taxonomy of data types. The paper furthermore discusses and develops a metric to assess the privacy relevance of different data types. Finally, the taxonomy is applied to the five major OSNs to evaluate its generalizability.

Keywords: Taxonomy; Privacy; Data types; Online social networks; Social identity management; Classification; Privacy relevance metric

1 Introduction

Online social networks (OSNs) have reached major importance due to their increased usage and ubiquity, rising membership, and presence in the media. Allowing their users to create custom profile sites, express relationships with other users, and explore the resulting social graph [1], they combine previously available communication and self-representation functions, such as personal blogs, forums, and instant messaging with novel social functions. Also, they allow reaching new contacts. The user base of OSNs is no longer restricted to private end users and college communities [2] but extends to professionals while serving as collaboration tools [3].

1.1 Privacy threats and the need for different data types

With the increased usage frequency and ubiquitous usage of OSNs, the quantity and sensitivity of user data that is stored on OSNs has grown tremendously as well. This is fostered by the availability of social networking services on mobile devices that provide location-based features and camera functions, for instance, allowing users to publish their current activity and location. For service providers, it is possible to derive rich profiles of their users [4], leading to *social footprints* [5].

Further privacy issues occur not only due to the service provider's data usage but also because other OSN users have access to user data. Similar to the physical world, OSNs need to offer means to create different images of the self, such as facets to cover the professional aspects of OSN-facilitated communication and further facets for family and friend-related representations of one's personality. Historically, creating and managing multiple facets of one's identity is not a OSN-specific phenomenon but part of everyday life. In historical records of ancient Greece, for instance, Plato refers to social interaction as the 'great stage of human life' (in Burns [6]). Sociologist Erving Goffman labels people's desire to control their appearances for different audiences as impression management [7]. Privacy is violated if information intended for a particular audience (such as one's family) unintentionally becomes available to another audience (such as one's employer). This understanding of privacy as respecting social norms of intended contexts is also referred to as contextual integrity [8].

On OSNs, on the one hand, the disembodied environment of these sites emphasizes the communication part

*Correspondence: christian.richthammer@wiwi.uni-regensburg.de
†Equal contributors
Department of Information Systems, University of Regensburg, Regensburg 93053, Germany

of social interaction. On the other hand, most communication is conducted in an asynchronous manner, i.e., OSNs *need to provide a variety of different data types* to adequately map social interaction onto the World Wide Web and to cover all aspects of social communication. This need for targeted and selective disclosure of personal information to create several facets of the self - representing different areas of the physical world - and keep them separated is also referred to as *Social Identity Management* (SIdM) [9].

1.2 Existing privacy-related research

Prompted by these developments, privacy concerns have been voiced by researchers. Numerous studies have been conducted on privacy issues on OSNs in general [10,11] as well as on people's awareness in this context [12,13] and on potential hazards [14,15]. Proposals for improving the user's understanding of disclosed information [16], enhanced access control models [17], and improved privacy protection on OSNs [18] have been made.

Observing the literature on privacy and user control in OSNs shows that there is little work describing data elements that are associated with the users, albeit surveying the application programming interface (API) of the popular site Facebook that reveals a large number of distinct data elements that can be associated with a user [19]. Related work in assessing the access control models in OSNs (e.g., [20,21]) does not differentiate between different attributes of the user identity, while others only focus on singular aspects such as the owner and creator of items [22]. Still, it is seldom or only briefly [23] considered that attributes on OSNs vary widely in implementation, semantics, applicable policies [24], and privacy controls [25] and thus carry far-reaching implications for the user.

This paper aims at tackling the lack of a generally accepted terminology for describing and differentiating data types on OSNs by developing and proposing a detailed taxonomy.[a] It is intended to benefit discussions among researchers, alleviate difficulties when comparing data elements within and across OSNs, and provide guidance for end users when assessing the privacy implications in dealing with particular OSN data types. To further ease the assessment of privacy threats, the paper introduces steps toward a metric to quantitatively assess the privacy relevance of particular data types.

The remainder of the paper is organized as follows. Work related to classifying OSN data types and privacy metrics is discussed and compared to our contribution in Section 2. The scope and methodology of the research are defined in Section 3. The proposed taxonomy is introduced in Section 4 accompanied by an analysis of related literature and a conceptualization of fundamental OSN user activities involving user data. Based on the data types identified in the taxonomy, a privacy relevance metric is developed in Section 5. The taxonomy is evaluated in Section 6 by applying it to five major OSNs before concluding the paper in Section 7.

2 Related work

This section provides an overview of important related work with respect to our work. We outline research regarding the study of user activities on OSNs, the conceptualization of data types, and the development of privacy metrics for OSNs. Furthermore, we point out how our paper distinguishes from related publications.

2.1 User activities on OSNs

Surma and Furmanek [26] and Zhang et al. [27] describe fundamental user activities on OSNs. The former work [26] focuses on user activities accustomed to a small community of OSNs, which is why it does not allow to draw generic conclusions. The latter work [27] is conducted on a high level of abstraction containing only three different entities and is used as a basis for the explanation of the variables of a heterogeneous network. Despite the unsuitable degree of abstraction and the completely different purposes, the study of user activities conducted in this paper in order to derive originating data types has been inspired by the user-centric approaches introduced above.

2.2 Data types on OSNs

Ho et al. [10] mention two different approaches for categorizing data. The first one is based on a survey in which users of OSNs were asked which data they would place on their profile. Consequently, the resulting classification only considers the items that were mentioned by the participants of the survey. In the second approach presented by Ho et al. [10], user data is divided by focusing on the data's impact on privacy. While reasonable for categorizing privacy settings, it is unsuitable for developing a general-purpose taxonomy as many other dimensions would be omitted. Similarly, Park et al. [4] also focus on certain aspects of data on OSNs. Data is categorized on the basis of its visibility (i.e., private or public) and its creator (i.e., the user himself or others). As a consequence, unlike this work, the categorization proposed by Park et al. [4] lacks a discussion of activity-related data types and solely focuses on the two dimensions mentioned before.

Beye et al. [28] follow a different approach that builds upon the definition of OSN by Boyd [sic] and Ellison [1] from which three data types are deduced. Additional six data types are derived by focusing on the goals of different OSNs. Compared to the approaches discussed so far,

Beye et al.'s work [28] contains a well-founded explanation on the origin of the data types. However, their definition (e.g., the definition of the data type *Messages*) can be considered too coarse-grained. No distinction is made concerning the item's visibility, its creator, and the domain in which it is created. These aspects are of major importance when analyzing the user's capabilities on modern OSNs.

Unlike other approaches which developed their classifications only as a basis for further examinations, Schneier [23] focuses solely on the task of establishing a taxonomy. However, his brief discussion lacks a structured methodology and does not mention any explanation on how he deduced the data types. Moreover, his taxonomy does not cover all important aspects of OSNs. For example, there are no data types in which the user's relationships or his connection-related attributes (e.g., Internet Protocol (IP) address) can be arranged. Årnes et al. [29] pick up the ideas proposed by Schneier [23]. Although being a meaningful extension to the work presented by Schneier [23], this approach also does not go into detail about the particular data types and is limited to a short definition and a list of examples for each category.

A major distinction between all previously discussed approaches and this paper is the level of granularity. Rather than aiming at a high-level classification, this work proposes a fine-grained taxonomy. In addition, the individual data types are arranged hierarchically, which is a common feature of taxonomies.

2.3 Privacy scoring for OSNs

Besides data types in OSNs and their classification, current research lacks an analysis of the privacy risks which single data types of the proposed taxonomies are accompanied with. A first step toward this direction has been made by Liu and Terzi [30] who proposed a framework for computing the privacy scores of users on OSNs. They developed mathematical models to assess the sensitivity and visibility of disclosed information. Their privacy score as an aggregation of combined sensitivity and visibility values provides a measure for the privacy risk a user faces. As the data basis for this score only involves the user's current privacy settings and the user's position in the social network, the authors assume that the user alone is responsible for the dissemination of his privacy-relevant data. However, as we will point out during the development of our taxonomy, privacy is significantly impacted by the creator and the publisher of data as well as the domain information is published in. The fact that privacy relevant data can be disclosed outside the user's domain bears additional potential risks. Further approaches to measure privacy risks include the work of Cutillo et al. [31] and Becker et al. [32]. The former assesses the achievable privacy degree in an OSN based on graph topology measures, such as degree distribution, clustering coefficient, and mixing time. Becker et al. [32], in contrast, measure the privacy risks attributed to direct social contacts. Similar to the work published by Liu and Terzi [30], these approaches do not make a differentiation between different data types but provide an aggregated value to measure the comprehensive privacy risks a user faces.

3 Research scope and approach

3.1 Problem scope

This work aims at developing a taxonomy for describing and classifying data types on OSNs, thereby benefiting three areas. The first goal is to improve comparability of user data within and across OSNs. Further, it intends to provide a clear terminology for discussions among researchers. Lastly, it aims at improving the understanding of attribute characteristics on OSNs and their implications by end users.

The goal of this paper is not to provide an exhaustive list of all attributes and data elements that are available or disclosed on current OSNs. Rather, it intends to develop a taxonomy to describe important characteristics of data types on OSNs and understand their differences, especially with regards to characteristics specific to OSNs and SIdM.

Note that this work focuses on centralized OSNs and only covers data types related to user actions that occur directly on them. External aspects like social plugins (e.g., Facebook's Like button) create extensive privacy issues. However, they have to be discussed separately and are out of the scope of this paper. Also, note that the subsequent discussions are solely based on facts and that no assumptions regarding the actions of OSN service providers are made.

3.2 Research approach

Aiming at delivering a taxonomy consisting of *constructs* for describing data types on OSNs that are used for abstracting from particular data types, a design-oriented research approach [33,34] is applicable to the problem scope. For conducting design-oriented research, a process model consisting of six steps - problem identification, elicitation of solution objectives, solution design, demonstration, evaluation, and communication - has been proposed [34].

The research approach employed in this work adapts the process proposed by Peffers et al. [34]. The first step has been performed in the previous two sections by identifying the problem and motivating the need for a taxonomy for data types on OSNs. Also, the corresponding research gap has been identified. Subsequently, the objectives of developing the taxonomy have been identified previously

in this section, thus constituting the second step of the design research process.

The research model depicted in Figure 1 shows the core steps performed in this paper. As a preparation for developing the taxonomy, the body of the related literature is analyzed with regards to possible elements of an OSN attribute taxonomy (Section 4). A conceptualization of fundamental user activities between the user, the OSN, and possibly the user's contacts that affect user data complements this analysis (step 1 in Figure 1). Based on these foundations, the proposed taxonomy is discussed thoroughly, which corresponds to the third step of the design research process model [34].

Evaluation is deemed as a *central and essential activity* [35] and a *key element* [36] in design-oriented research. Correspondingly, the design research process [34] contains both a demonstration and a dedicated evaluation step. The taxonomy is demonstrated (step 2 in Figure 1) by applying it to five major OSNs and identifying actually implemented data types for each element of the taxonomy (Section 6). Besides demonstration, the taxonomy is evaluated by identifying all data types of the OSN Facebook and iterating over all these data types and mapping them into the taxonomy (step 3 in Figure 1).

The presentation of results in this paper concludes one iteration of the design science process and corresponds to the communication step.

4 Proposed taxonomy

To arrive at a taxonomy for OSN data types, this section follows the previously outlined research model. In an initial step and based on Section 2, a thorough literature analysis reveals in essence the following three related approaches: Schneier [23], its refinement by Årnes et al. [29], and the classification by Beye et al. [28]. Figure 2 correlates the data elements of these approaches and the taxonomy proposed in this work, while the subsequent discussion of this section highlights conceptual similarities and deviations.

The analysis of Figure 2 leads to several observations: Firstly, it reveals that to some extent terminology is not consistently used, such as the different understanding of behavioral data [23,29] and behavioral information [28]. Secondly, a general lack of granularity can be attributed to some existing data type definitions, as observable in Beye et al.'s generic conceptualization of profiles [28]. Consequently, it is difficult to precisely specify data elements as needed in scientific discussions. Lastly, some related work either does not cover all available data types, such as the missing specification of data related to the connection with other users in Schneier's proposal [23], or focuses on data elements whose existence is difficult to verify (e.g., the probability-based derived data in Schneier's [23] and in Årnes et al.'s [29] work that stems from the combination of several other data types).

Based on the analysis of existing literature, this work follows a user-centric approach by studying data that is created during possible user activities on OSNs. Figure 3 illustrates OSN entities and possible activities. As can be seen, most activities are either initiated by the user or one of his contacts. The subsequent elicitation of data types will refer to the numbered steps in Figure 3 to clarify the origin of a particular data element.

As a taxonomy is commonly regarded a hierarchical classification, this paper takes a top-down approach stepwise subdividing the set of data types into non-redundant partitions. The process is repeated until all data types are classified. At the first level, a distinction is made based on the stakeholder for whom a particular data type is of use. From a privacy point of view, two stakeholders are distinguishable [37]: *service providers* and *OSN users*. The former group offers OSN platforms and related services whereas personal data commonly provides the basis of their business model. For OSN users as the second

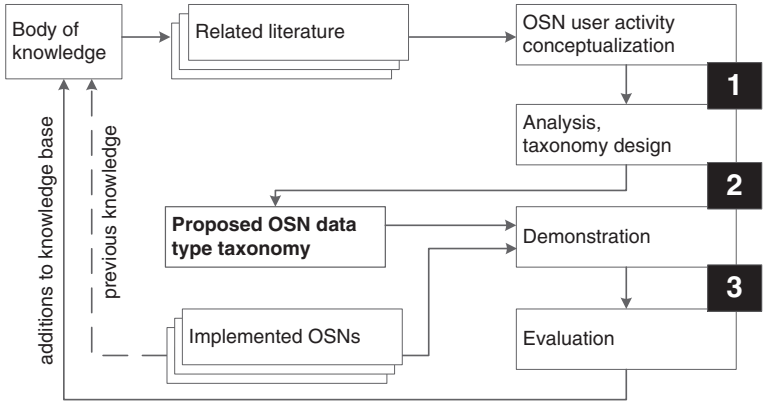

Figure 1 Research model.

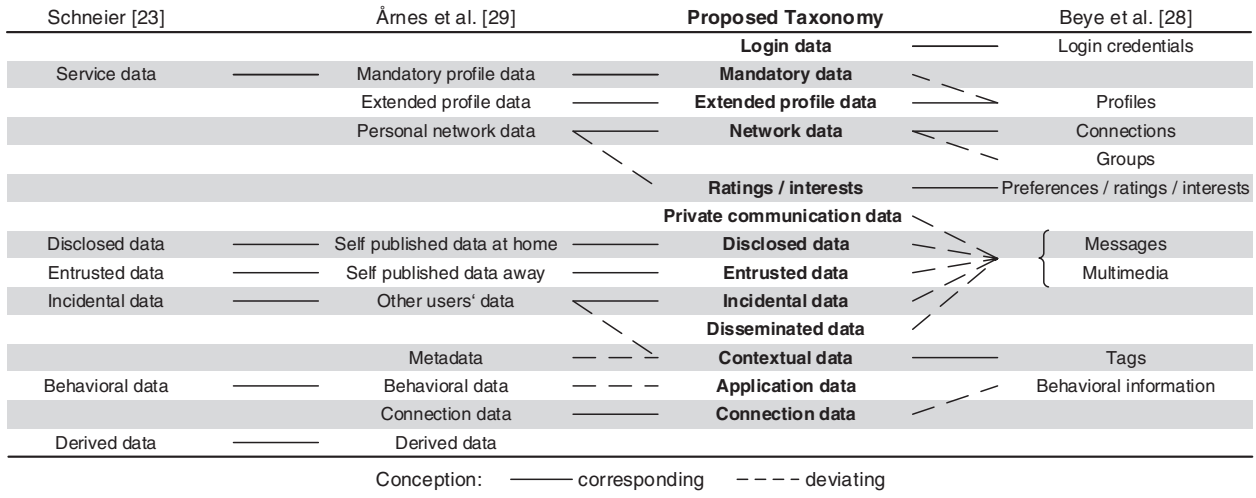

Figure 2 Mapping between data types of related classifications and this work.

stakeholder, personal data is used for the purpose of SIdM. In the following, accruing data types for each stakeholder are discussed in detail.

4.1 Service provider-related data types

Note that while service providers of centralized OSNs typically have access to personal data that is generated in user-related activities, this section discusses only data that originates from the service usage. Drawing on user activities identified in Figure 3, several service provider-related activities can be identified. In the following, data emerging from these activities is classified into three separate data types. To assess the privacy impact of service provider-related data types in general, it is valuable to recall that privacy is a social concept. In more detail, privacy is breached if other people that have a relationship with the user gain access to this data. For service providers, in contrast, such a relationship does commonly not exist. Moreover, data is typically processed by machines and algorithms without human interaction. Nevertheless, in the longer run, the data types which are subsequently discussed may lead to social privacy implications. For

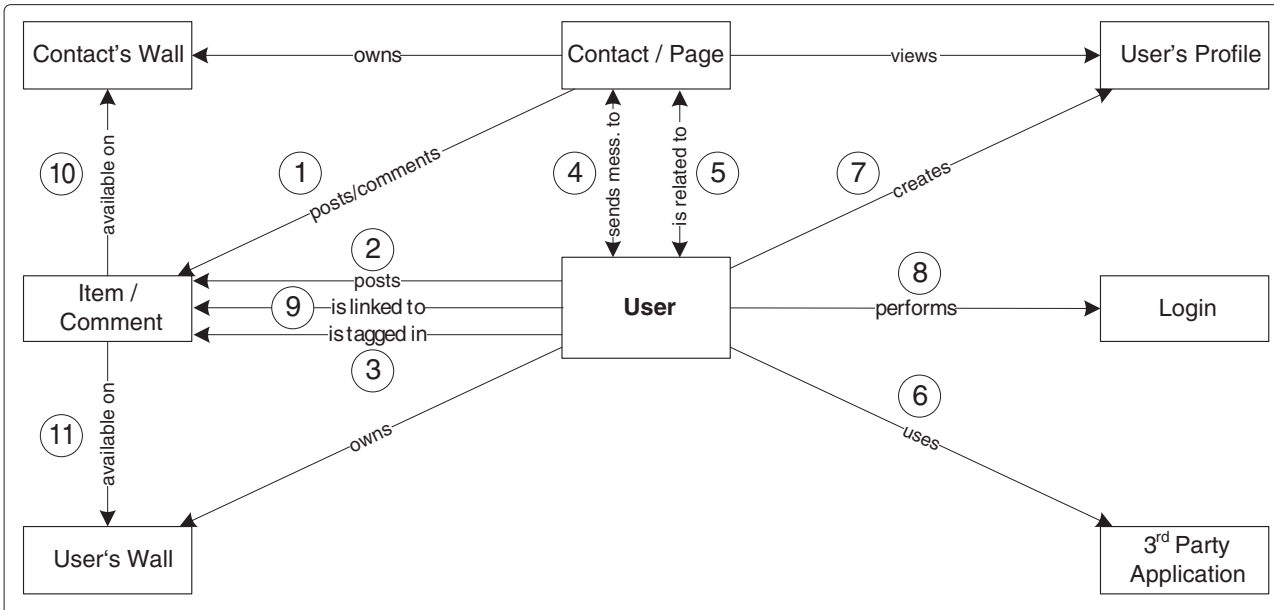

Figure 3 Fundamental user activities on OSNs.

instance, if such data becomes publicly available (e.g., through data breaches), people that do have a relation with the user can use this data to gain more knowledge and thereby invade his privacy. Consequently, protection is necessary and therefore discussed for each of the service provider-related data types.

4.1.1 Login data

OSN service usage requires prior user authentication to prevent identity theft, which is represented by activity 8 in Figure 3 and is consistent with Beye et al.'s respective data type [28] (*cf.* Figure 2). Consequently, login data is considered a data type that is required by the OSN service provider to provide evidence of a claimed identity. Common instances of this data type are identifiers such as username and email address as well as passwords used to verify an identity. From a privacy perspective, identifiers such as the user's email address may facilitate the linkability of different partial identities and may lead to the compilation of a more comprehensive profile. In addition, inadequate protection of login data may allow other users to access a user's profile and gain access to personal information that was not intended for them. This may cause a violation of privacy in the sense of contextual integrity (*cf.* Section 1).

4.1.2 Connection data

While not OSN-specific, requesting - i. e., connecting to and using - Internet-based services (activity 8 in Figure 3) leads to a variety of digital traces created by protocols on several layers of the OSI model. Figure 2 shows that the definition is consistent with the work presented by Årnes et al. [29], while a broader conceptualization is used by Beye et al. [28]. Instances include the user's IP address, the type of communication unit (such as mobile devices), information related to the browser and the operating system, and location (derived from the IP address or using GPS). Especially browser-related information and location are deemed sensitive and entail privacy implications when being available to OSN service providers, such as for acquiring detailed user information through cookies and browsing history or for creating a movement profile based on location data. Attacks, such as browser fingerprinting, have been successfully demonstrated, allowing users to be identified solely based on their HTTP headers even if they update their browser version [38]. In this case, it is not the actual data that is contained in this data type which is of privacy importance but rather the potential to use this data to de-anonymize users, link previously unrelated personal information, and thereby invade privacy.

4.1.3 Application data

Besides OSN platform usage, data originating from the use of third party services (activity 6 in Figure 3) running within the boundaries of the OSN platform or having API access can be differentiated. None of the related work explicitly focuses on this type of data. Common examples are player statistics of OSN games, application usage statistics, or in-app purchase data such as credit card information. This data type may entail both privacy and security risks. On the one hand, data security largely depends on the trustworthiness and protection mechanisms of the service provider and third party services. Breaches may lead to serious consequences such as credit card fraud. On the other hand, privacy may be threatened if third party usage statistics become available. For instance, an employer may notice that one of his employees is playing OSN games during working hours.

4.2 User-related data types

To model the diversity of a user's personality and his ways of social interaction, an OSN account offers a variety of means to express oneself and communicate with other users. Fundamentally, two classes of data can be distinguished: *semantically specified* and *semantically unspecified* data. The first category refers to data instances that have a clearly defined meaning and whose content is clearly understood. Examples include predefined attribute types of an OSN profile such as name, birthdate, and hometown. Yet, OSN service providers have acknowledged that it is difficult to force all aspects of a user's personality into well-specified conceptual boxes. Hence, semantically unspecified data types are provided to freely express some facets of one's personality, such as status posts whose content is not semantically predefined. Note that for some data types, a selective classification is difficult as parts of the data type are semantically specified while others remain unspecified. For this taxonomy, we focus on the data type's value for classification. For example, while the concept of a shared photograph is semantically defined, its value is difficult to interpret and consequently the data type is semantically unspecified. As another example, for friend lists, not only the concept itself but also its value (a set of contacts (which is also a clearly specified concept)) is semantically defined.

4.2.1 Semantically specified

Data elements available for self-description and expression of one's personality can be further subdivided into mandatory and optional data types.

4.2.1.1 Mandatory data Similar to the physical world, a minimal set of data is required to initiate social interaction. Consequently, this class covers data that is needed for an OSN service to be useful and to enable basic functionalities such as user discovery and verification purposes. Mandatory data refers to personal information that needs to be provided by the user during the registration or profile creation process (activity 7 in Figure 3), which -

except from the terminology used - corresponds with the works presented by Schneier [23] and Årnes et al. [29] (*cf.* Figure 2). A common example is the user's name serving as an identifier for other users to create a social graph. Due to age verification processes because of possibly inappropriate content and in order to preclude immature users, the user's birthday is also a frequently required attribute. Privacy implications for mandatory data depend on the concrete implementation by a OSN service provider. It needs to be examined whether mandatory data becomes part of the OSN user's profile and if privacy settings are available to restrict its visibility.

4.2.1.2 Optionally provided data Besides the mandatory data, several data types with clearly specified semantics exist on OSNs that are subsequently discussed.

Extended profile data: OSNs offer a variety of predefined attribute types that may be used to further describe particular aspects of one's personality. Note that extended profile data solely refers to the user's profile while other parts of an OSN account are covered by further data types. Consequently, properties of extended profile data are the following: profile-centricity, optionality, predefined attribute types with clear semantics, and in some cases predefined attribute values. Typically, the process of providing extended profile data (activity 7 in Figure 3) is guided by a form that contains input fields for attribute types, like address, education, favorite music, favorite films, hobbies, and interests. The profile picture, which is a common feature of OSNs, is also arranged in this category. According to Figure 2, this is in line with the conceptualizations presented by Schneier [23] and Årnes et al. [29], while the profiles' category presented by Beye et al. [28] is considered too coarse-grained. From the optionality of this data type, it follows that privacy risks are manageable as it is down to the user to decide whether to disclose a particular personal attribute. On closer examination, available privacy settings are to be considered as these define the granularity of the potential audience that may access an attribute.

Ratings/Interests: Besides the extended profile data that allows for a rich description, the study of user activities (activities 5 and 9 in Figure 3) reveals that binary or predefined multi-value attributes related to existing entities such as pages and shared items which are used to refine how one is seen by others (e.g., by liking favorite bands). Corresponding with the data type proposed by Beye et al. [28] in Figure 2, this class of data covers expressed interests such as Facebook's Like and Google's +1 and the rating of photos shared by other users. With regards to privacy, two aspects need to be discussed. On the one hand, privacy implications and privacy control depend on default or available visibility settings. On the other hand and in contrast to the structured listing of mandatory and extended profile data, ratings and interests are typically bound to items shared by others and hence are widely distributed across the OSN. Thus, awareness of previously expressed ratings and interests becomes increasingly important as it may allow others to draw inferences from all instances of this data type about the user's personality.

Network data: As social interaction is an inherent property of OSNs, users are encouraged to express their relationship with other users (activity 5 in Figure 3). The collection of all connections of a particular user is often referred to as his social graph [28] and describes data concerning the network the user has built around himself on the OSN, which conforms to the definition of Årnes et al. [29] as presented in Figure 2. From the viewpoint of a particular user, a single instance of network data has a binary value, i.e., a connection either exists or not. Network data may be uni- or bidirectional and differ in the strength of a connection. Common examples include the notions of friend, friend-of-friend, follower, and someone you are following. Depending on its concrete implementation, network data may be visible by default or access to it can be controlled by the user. Access to network data can be considered to have a significant impact on privacy. On the one hand, knowledge of a user's social graph allows to draw inferences about his identity and enables sybil attacks on other OSNs by forging the user's identity and connected identities [39]. On the other hand, it has been demonstrated that (partial) knowledge of a user's social graph (such as knowledge of groups he is member of) is sufficient to reveal his identity [40].

Contextual data: While some data shared on OSNs contains an atomic piece of information (such as the user's birthdate), other items such as pictures enclose a multitude of information. This class of data refers to a property of an existing item that is made explicit and provided with semantics, hence forming a new data type. Common examples include the tagging feature, allowing to make peoples' names (and eventually their identity) in an existing picture explicitly available to other OSN users (activity 3 in Figure 3). Further instances are the location of a picture and the relation of a shared item to an activity or an event. The comparison of existing taxonomies in Figure 2 shows that while corresponding with Beye et al.'s taxonomy [28], this type of data is only partly covered by Schneier [23] and Årnes et al. [29]. Two aspects of contextual data are particularly of importance when assessing the privacy risk. Firstly, contextual data is often a byproduct of its primary data type and as such is sometimes not explicitly visible to the user. As an example, a shared picture may contain a variety of information including the camera model, camera owner and location. Secondly, contextual data is often machine-processable (while its host data type can be semantically unspecified). Consequently,

it allows service providers to extrapolite from contextual data to the content of the host data type.

4.2.2 Semantically unspecified

Semantically unspecified data refers to data elements provided by the OSN where the data format is predefined but whose content is left to the user and cannot trivially be interpreted by machines. For instance, a photo album feature predefines the format (digital photos) but leaves the picture's content to the user (yet it is notable that OSN service providers are increasingly making progress on face recognition technologies). As a consequence, on the one hand, it is difficult to make generalizations on privacy risks associated with semantically unspecified data types where risks largely depend on the content. On the other hand, the lack of semantic specification impedes OSN service providers from automatically processing this data. To further refine the classification, a distinction can be made between data used in 1:1 and 1:*n* communication.

4.2.2.1 Private communication data This class covers data elements that originate from private communication (i.e., 1:1 communication) between two OSN users (activity 4 in Figure 3), which is only partly covered in Beye et al. [28] as illustrated in Figure 2. While private communication may comprise text messages as well as other media formats, their content is not semantically specified. Examples include private messages with or without attachments, private video chats as well as smaller interactions such as poking other users. Private communication data is not accompanied with privacy risks as long as the communication partner can be trusted, the OSN security mechanisms prevent third parties from gaining access, and the OSN service provider does not inspect the messages to an extent greater than roughly scanning them for illegal content.

4.2.2.2 1 : *n* communication Besides private communication between two users, data with semantically unspecified content can be shared with an audience of *n* other users where *n* defines the degree of publicness. Each of the subsequently discussed data types is concerned with semantically unstructured data such as photos, status messages, and comments, yet a differentiation is made between creator, publisher, and the domain in which the element is published (see Table 1). As can be seen from Figure 2, the first three data types subsequently discussed are based on the work presented by Schneier [23] and Årnes et al. [29].

Disclosed data: A frequent user activity on OSNs is to post information on one's wall (activities 2 and 11 in Figure 3). In conceptual terms, the data is generated and published by a user in his own domain. Privacy of disclosed data largely depends on the availability of visibility settings and the concrete implementation of the OSN service provider. Apart from this, privacy is not affected as the user is both the creator and publisher of the item (i.e., he intentionally wants the data instance to be visible to others), and it is shared within the domain of the user (no other users are affected with the shared item).

Table 1 Differences between disclosed data, entrusted data, incidental data, and shared data

	Creator	Publisher	Domain
Disclosed data	User	User	User
Entrusted data	User	User	Contact
Incidental data	Contact	Contact	User
Disseminated data	User	Contact	Contact

Entrusted data: In contrast, entrusted data refers to information that is both user-generated and user-published but in the domain of a contact (activities 2 and 10 in Figure 3), i.e., the former is able to shape the latter's representation on the OSN. Consequently, once the data is shared, control passes over to the domain owner that is from then on able to define its visibility. Whether this ability is only extended or shifts completely depends on the concrete OSN. Examples include posts and comments made on another user's wall or a similar space. Privacy implications mainly arise from the loss of control once the data element is published.

Incidental data: Incidental data originates from a contact sharing a data element on the user's wall (activities 1 and 11 in Figure 3), i.e., the contact is both the creator and publisher, however, the information is shared in the user's domain. In this scenario, a contact is able to shape the presentation of the user on the OSN. As a consequence, the user gains control over the item, whereas the extent depends on the concrete implementation.

Disseminated data: In the last case of Table 1, user-generated data elements that are considered are further disseminated by a contact within his own domain (activities 1 and 10 in Figure 3). This may include data elements that the user has initially shared with the contact or provided to him using other communication channels. In the first case, which is also discussed by Hu et al. [22], the OSN may prevent the contact from publishing the item with a larger than the user's intended audience and grant additional permissions to the user. However, in the second case, the contact remains the only person to control the visibility of the data element, raising serious privacy implications.

4.3 Summary

Figure 4 provides an overview of the proposed taxonomy based on the previous discussion. It comprises 13 data types that are integrated in a hierarchical structure. The

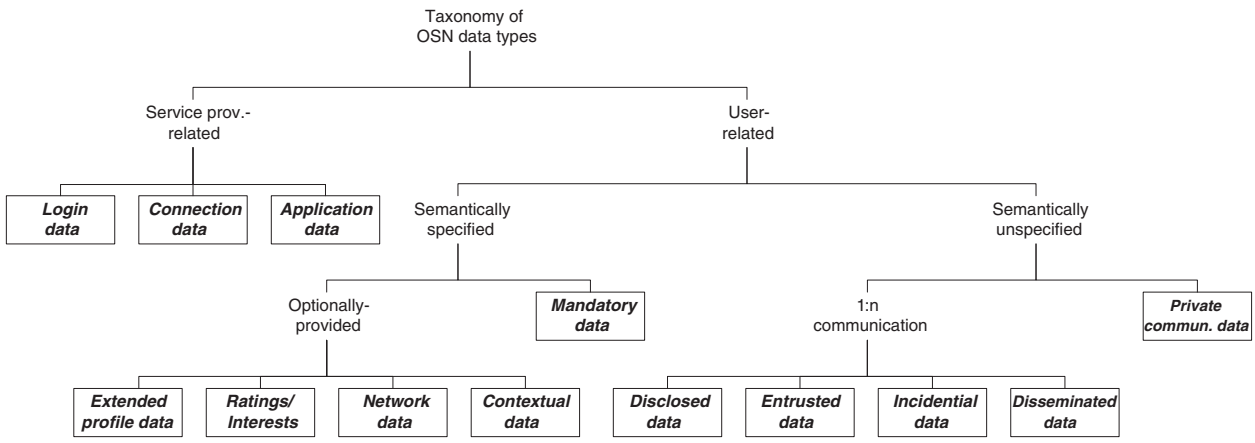

Figure 4 Proposed taxonomy of OSN data types.

analysis of privacy implications of data types revealed that privacy mainly depends on the interplay of a data element's content, the extent and granularity of user control, and its concrete implementation. The content may be easily accessible to service providers for data types with clear semantics, while semantically unspecified data requires human cognition for interpretation. Besides, each service provider decides whether the collection and visibility of a particular data type is user-controllable. If user control exists, its granularity largely depends on the concrete OSN implementation.

5 Toward a metric for assessing privacy relevance

Taking a bird's-eye view on our taxonomy, it becomes clear that the data types classified differ in the privacy risks they are accompanied with. Private communication data, for instance, is obviously not privacy critical as long as the communication partner is known and trusted. Disclosed data, in contrast, can be critical since one may easily lose track of who is able to read the content. Furthermore, permissions or social graph structures could change over time. As a result, a user could suddenly see messages that were originally not visible to him. In order to make the privacy relevance quantitatively measurable, a privacy relevance metric (PRM) is needed.

OSN data is always created within a specific context, which can be described through various attributes. The range comprises general factors such as audience size or audience composition as well as content-specific attributes like the topic of a message. These attributes form a finite context set $C = \{c_1, c_2, \ldots, c_n\}$. Based on this context set which serves as an input, our metric should calculate the privacy relevance of a data type. For this purpose, we compare the input context set (subsequently denoted as b) to a reference context set (subsequently denoted as a) which describes a scenario having no privacy issues. The outcome of the comparison should be a value within the range [0,1] with 0 denoting no privacy relevance (reference value) and 1 denoting maximum privacy relevance.

To this end, we first provide a generic PRM that can easily be adapted to a set of preselected context attributes. Secondly, we show how this generic PRM can be implemented for the three context attributes *audience size*, *domain*, and *creator and publisher*. In the evaluation section, we finally demonstrate our PRM by measuring the privacy relevance of the data types classified in our taxonomy for a fictive profile on Facebook.

To further ease the understanding, we will use the following worked example during the development of our metric: OSN user Jane posts a photograph within her domain (e.g., on her wall) and sets the visibility to all of her 150 friends (i.e., audience size $= 150$). Using the taxonomy, this is classified as *disclosed data*. For Jane, the size of the audience and who is creator and publisher is twice as important as the domain in which the photograph is posted.

5.1 Metric space and distance function

Context attributes can be modeled as a dimension in a vector space. Therefore, we define an n-dimensional metric space M, where each dimension is derived from one context attribute of the context set $C(|C| = n)$. We furthermore introduce n mapping functions $f_i(c)$ to range the context attributes to a predefined interval (either continuous or discrete). To measure the distance of points in the metric space M, various distance functions can be used. A function

$$d : M \times M \rightarrow \mathbb{R}$$

is called distance function if it satisfies the following conditions for all $a,b \in M$:

1. $d(a,b) \geq 0$
2. $d(a,b) = 0$ if and only $a = b$
3. $d(a,b) = d(b,a)$
4. $d(a,c) \leq d(a,b) + d(b,c)$

Exemplary distance measures are the Manhattan distance or the Euclidean distance which are two types of the generic Minowski distance. The decision which distance measure to use depends on the application area. In this work, we chose the Manhattan distance which is more robust to outliers.

To implement an exemplary PRM, we selected the three context attributes *audience size*, *domain*, and *creator and publisher*. These attributes form our context set $C = \{c_1, c_2, c_3\}$, where c_1 takes values of the discrete interval $C_1 = \{0, 1, 2, \ldots\}$, while c_2 and c_3 take values of the nominal sets $C_2 = \{\text{user, contact}\}$ and $C_3 = \{$ (user, user), (user, contact), (contact, contact)$\}$.

The space M^{prae} is defined as follows:

$$M^{\text{prae}} : = C_1 \times C_2 \times C_3$$

In M^{prae}, the single attributes are not comparable in their values. This might lead to a bias in the final context value. The *audience size*, for instance, is not necessarily bounded from above allowing very high distances between its members. The distances between members of the nominal sets *domain* and *creator and publisher* must be transformed before they can be measured. Hence, we define three mapping functions $f_i(c)$ to range the values within an interval of $[0,\gamma_i]$. γ_i is the weight factor that allows to weight single context attributes based on their importance. For $\gamma_1 = \gamma_2 = \gamma_3$, each dimension has the same impact on the global PRM. Thus, we define the final metric space M as

$$M : = [0, \gamma_1] \times [0, \gamma_2] \times [0, \gamma_3]$$

To make the privacy relevance measurable, we define the distance function $\tilde{d}$ in our metric space M as follows:

$$\tilde{d} : M \times M \rightarrow \left[0, \sum_{i=1}^{3} \gamma_i\right], \quad (a,b) \mapsto \sum_{i=1}^{3} (b_i - a_i)$$

A reasonable and intuitively understandable value range for the distance as a relevance measure is [0,1] with 1 denoting a maximum privacy relevance and 0 denoting no privacy relevance (absolute congruence to reference set). Therefore, $\tilde{d}$ is scaled by the inverse of $\tilde{d}_{\max} = \sum_{i=1}^{3} \gamma_i$ which quantifies the length of the diagonal in the metric space. That leads to the final distance function d:

$$d : M \times M \rightarrow [0, 1]$$

$$d(a,b) := \frac{\tilde{d}(a,b)}{\tilde{d}_{\max}} = \frac{\sum_{i=1}^{3} (b_i - a_i)}{\sum_{i=1}^{3} (\gamma_i)}$$

5.2 Context dimensions

The context dimensions depict the single attributes of the context set C. Since the context attributes consist of both structured and unstructured data of random values, modeling the mapping functions $f_i(c)$ can turn out to be difficult. However, this is an important step because parameters that vary greatly in size may have a much bigger impact on the distance than parameters that differ only slightly, even though slight distinctions could have an equal impact on the changes in context. In the following paragraphs, we define the mapping functions $f_i(c)$ for the three context attributes *audience size*, *domain*, and *creator and publisher*. In more detail, for each of the three attributes of input context set b, a mapping function $f_i(c)$ is defined. The same mapping functions are used for each of the three attributes of reference context set a.

5.2.1 Audience size

As stated above, the audience size is critical for privacy issues since an increased audience leads to a higher privacy risk. However, privacy relevance is not static and exactly the same for similar classes in different profiles. It depends on a user's number of friends. To capture this, we propose the following function:

$$f_1(c_1) = \gamma_1 * \left(1 - e^{\frac{c_1 * \ln(0.5)}{n_f}}\right)$$

with c_1 depicting the audience size and n_f quantifying a user's number of friends. The outcome of this function ranges in the interval of $[\,0, \gamma_1)$. Content that is only visible to the creator (audience size = 0) results in a value of 0, whereas content visible to the user's friends is reflected in a value of $0.5\gamma_1$. The value converges to γ_1 for content being visible to all friends-of-friends or all users. The logic modeled in this metric states that the growth of perceived privacy risk decreases with an increasing audience size. This coherence can be easily understood thinking of a user who shares a secret with one person. Having told two persons instead, the perceived privacy risk would have been notably higher. Now think of 1,000 persons. The difference of perceived privacy risk compared to sharing the private information with 1,001 persons instead would have been marginal.

Continuing the worked example introduced previously and assuming a weight factor of 2, the privacy relevance of Jane's audience size of 150 friends is calculated as follows:

$$f_1(c_1) = 2 * \left(1 - e^{\frac{150*\ln(0.5)}{150}}\right) = 1$$

5.2.2 Domain

Data can either be published in the user's or in a contact's domain. If shared in the own domain, the user is the data owner and can regulate the visibility and availability. Published in a contact's domain, in contrast, one passes these privileges. That may lead to increased privacy risk. We therefore model the domain function as follows:

$$f_2(c_2) = \begin{cases} 0, & \text{if } c_2 = \text{user} \\ \gamma_2, & \text{if } c_2 = \text{contact} \end{cases}$$

In the worked example, a weight factor for the domain of 1 was assumed. Hence, the privacy relevance of the domain attribute is calculated as follows:

$$f_2(c_2) = 1 * 0 = 0$$

5.2.3 Creator and publisher

Serious privacy concerns are raised if the user creates content that is further disseminated by a contact within his own domain. The user thereby loses control of his own content. The third function is designed to cover this case:

$$f_3(c_3) = \begin{cases} \gamma_3, & \text{if } c_3 = \{\text{user, contact}\} \\ 0, & \text{else} \end{cases}$$

A weight factor of 2 for the creator and publisher attribute in our worked example leads to the following calculation:

$$f_3(c_3) = 2 * 0 = 0$$

Based on these results, the overall privacy relevance for Jane in the worked example can be calculated. As previously defined, the privacy relevance for each attribute of the reference context set a is 0. Hence the overall privacy relevance is

$$d(a,b) = \frac{(1-0) + (0-0) + (0-0)}{2+1+2} = \frac{1}{5} = 0.2$$

6 Evaluation

Two approaches are used to evaluate the proposed taxonomy, namely, demonstration (which is described as a light-weight evaluation by Venable et al. [35]) of its efficiency to solve a certain problem [34] and evaluation of its suitability to map all data types of a given OSN. In the first part of the evaluation, five major OSNs - Facebook, Google+, Twitter, LinkedIn, and Instagram - are analyzed under the aspect of using the proposed taxonomy. Note that the intention of the analysis is to show the feasibility of the taxonomy in general and to present the most common and most important examples for each data type. With the help of these examples, the main differences between the inspected OSNs can be shown in a descriptive way that is comprehensible for casual OSN users as well. The differences are highlighted by referring to the availability and importance of the data types on the particular OSN but also by pointing out existing privacy implications and user control mechanisms. The second part of the evaluation takes the opposite direction, setting the starting point to the OSN (i.e., Facebook) with all its data types. These data types are mapped into the taxonomy in order to show that it is able to cover all of them. Finally, we demonstrate our privacy relevance metric and discuss its benefits and shortcomings.

6.1 Application of the taxonomy to OSNs

Table 2 gives an overview on the data types of the inspected OSNs as available on 4 March 2014.

6.1.1 Service provider-related data types

Login data can be found on all OSNs. Facebook, Google+, Twitter, and LinkedIn all provide a login via email and password. On Facebook, the phone number can replace the email. On Twitter, a login is alternatively possible via username and password. As opposed to the other four OSNs, Instagram does not provide a login via email but only via username.

Connection data is collected by all OSNs. In order to inspect the items arranged in this category, the privacy policies of the five OSNs have been analyzed. It is important to state that these policies do not list every single data item collected through the use of the platform. For example, Google tries to arrange the collected data into categories (e.g., device information, log information, location information) and then mentions the most important examples with the help of expressions like 'such as' and 'may include.' However, splitting up connection data in the three data types mentioned above does not lead to better results regarding the taxonomy because the analyzed providers do not define them in the same way. Moreover, the five inspected OSNs differ in the examples they list and their level of detail. Nevertheless, all OSNs collect similar data items that can be arranged in the categories used by Google, which is why we also employ them in Table 2.

Application data is available on all five inspected OSNs because for all of them, there are connectors for external websites or unofficial smartphone apps. On Facebook and Google+, the number of third party applications

Table 2 Demonstration of the taxonomy on Facebook, Google+, Twitter, LinkedIn, and Instagram

Data types	Facebook	Google+	Twitter	LinkedIn	Instagram
Login data	Email, phone, password	Email, password	Email, username, password	Email, password	Username, password
Connection data	Device, log, and location information, cookies	Device, log, and location information, cookies	Device, log, and location information, cookies	Device, log, and location information, cookies	Device, log, and location information, cookies
Application data	Usage statistics, credit card data	Usage statistics, credit card data	Usage statistics	Usage statistics	Usage statistics
Mandatory data	Name, email, birthday, gender	Name, email, birthday, gender	Name, email	Name, email, job status, country, postal code	Name, email
Extended profile data	General-purpose input fields	General-purpose input fields	Bio, location, website	Professionally related input fields	Phone, gender, bio, website
Ratings/interests	Page, status/photo/video	Page, status/photo/video	Verified account, Tweet	Organization, status	Photo/video
Network data	Unidirectional, bidirectional	Unidirectional	Unidirectional	Bidirectional	Unidirectional
Contextual data	Tag in status/comment, on photo, at location	Tag in status/comment, on photo, at location	Mention in Tweet	Tag in status/comment	Tag on photo
Private commun. data	Private message, video chat, poke	Private message, video chat	Private message	Private message, InMail	N/A
Disclosed data	Text post, photo (album), video, check-in	Text post, photo (album), video, check-in	Text post, single photo	Text post, single photo, file attachments	Single photo, video
Entrusted data	*See disclosed data*	*Restricted to comments*	N/A	*Restricted to comments*	*Restricted to comments*
Incidental data	*See disclosed data*	*Restricted to comments*	N/A	*Restricted to comments*	*Restricted to comments*
Disseminated data	*See disclosed data*	*See disclosed data*	*See disclosed data*	*See disclosed data*	*See disclosed data*

is bigger by far as there are a lot of providers for games. As mentioned in Section 4, games may process credit card information because of in-app purchases, whereas website connectors and smartphone apps do not collect additional data (i.e., in addition to the data already available without using the application) except for the usage statistics. An important characteristic of application data is its optionality, i.e., the user decides about the use of third party applications. In the majority of cases, confirmation for requested permissions is required before being able to use an application. Consequently, user control is implemented on a binary decision basis.

6.1.2 User-related data types

As all OSNs include profiles, mandatory data and extended profile data always exist. Basic items of *mandatory data* are name, email, birthday, and gender. The first two items are mandatory on all inspected OSNs; the latter two items are only required on Facebook and Google+. There is a peculiarity regarding email being mandatory on Google+. Users do not have to provide an email account in the first place but Google will automatically create one for them, which is why we treat email also as mandatory here. LinkedIn forces a new user to indicate his country and postal code for networking purposes. Moreover, his job status is mandatory as well, which is motivated by the way LinkedIn describes itself - as a network for professionals. Note that email, birthday, and gender can usually be hidden from other users, giving the user the ability to alleviate certain threats (e.g., social engineering attacks with the help of personalized emails). If mandatory attributes are hidden, they are only used for internal purposes, such as using the user's gender in order to address him with the correct pronouns.

Which *extended profile data* is ultimately present in addition to the profile photo and the cover photo - each inspected OSN uses at least the concept of the profile photo - depends on whether the OSN is a platform for general purposes (e.g., Facebook, Google+) or for rather specialized ones (e.g., LinkedIn). Facebook and Google+ offer the user the ability to provide a variety of attributes such as basic info, contact info, work, education, and living. Similarly, LinkedIn offers additional data elements to refine one's profile but with a professional focus (such as experience, skills, publications, and awards). In contrast to the three OSNs mentioned before, Twitter does not focus on this detailed self-presentation in one's profile and only offers three single input fields for extended profile data. Instagram also only offers the input fields of Twitter and adds the phone number as a fourth data item. Although the provision of extended profile data is optional on all OSNs, only Facebook and Google+ offer a selective disclosure of attribute values. On Twitter, LinkedIn, and Instagram, they are either publicly visible or only available to oneself.

Ratings/Interests is a category that possesses differing importance on OSNs but can be observed on all of them. On Facebook and Google+, it is possible to express one's preference for all kinds of pages (e.g., persons, products, sports). On Twitter and LinkedIn, the pages mainly resemble verified accounts of well-known persons and organizations, respectively. Moreover, the focus lies more on staying informed about these pages rather than publicly demonstrating certain interests. Instagram is quite similar to Twitter with regards to following pages of well-known persons in order to stay up-to-date on them. However, there are currently no indicators that officially verify a celebrity page. Besides the pages mentioned above, the inspected OSNs all provide mechanisms to express one's favor for the items that are available as disclosed data. When focusing on the user's control over the visibility of his preferences, pages and disclosed data have to be discussed separately. For disclosed data, the visibility of one's favor always depends on the visibility of the corresponding item, whereas for pages, the visibility options are different on the inspected platforms. As following pages on Twitter and Instagram is done by establishing a unidirectional connection to them, the visibility properties are them same as for network data (see below), which means that the pages users follow are publicly visible. Users of Facebook and Google+ have the option to hide their preferences.

As the term *Online Social Network* already indicates, OSNs always include *network data*. The main difference between OSNs is whether the connections are bidirectional (e.g., Facebook, LinkedIn) or unidirectional (e.g., Google+, Twitter, Instagram). As shown in Table 2, Facebook is the only OSN that supports both types of connections at the same time. However, the unidirectional connections have to be enabled by the user before others are able to follow him without befriending him. Another important difference can be observed when analyzing the user's ability to hide his social graph from other users. Facebook, Google+, and LinkedIn implement this feature, whereas Twitter and Instagram always reveal your followers and the users you are following.

Further differences between the inspected OSNs can be observed when analyzing the presence of *contextual data*. On Facebook and Google+, the user has the ability to tag his contacts in text posts/comments, on photos, and at locations. Although being limited to text posts, Twitter's tagging feature creates more extensive privacy issues than the ones provided by Facebook and Google+ because Twitter's users lack the ability to remove these tags on their own. LinkedIn also limits the tagging feature to referencing contacts in text posts/comments but here, these

references can be easily removed. Instagram supports tags on photos.

Private communication data can be found on all inspected OSNs except Instagram. Facebook and Google+ offer this feature via instant messaging and without any limitations concerning availability and text length. On Twitter, private messages are provided via *Direct Messages*, which resemble a private *Tweet* and therefore are limited to 140 characters. Apart from messages to contacts, LinkedIn provides a feature called *InMail*, which is advertised as a professional and credible way to reach anyone on LinkedIn without prior introduction but which is only available to premium users. In addition to text messages, Facebook and Google+ offer video chats as another type of private communication data. Facebook also has the poking feature mentioned in Section 4. The only way of privately contacting another user on Instagram is to use Instagram Direct, a feature that was introduced in December 2013, and disclose a photo or video item only for that particular person. Thus, the two users are able to chat via the comments under the item.

Significant discrepancies concerning the availability of the data types have been identified when posting items. Firstly, there are differences in the complexity of the items and secondly, the ability to post them may be restricted to the user's own domain. Facebook and Google+ enable their users to post text, photos, photo albums, videos, their current location, and other objects (e.g., questions, events). LinkedIn users are also able to post text and single photos but cannot create entire albums. However, they can enrich their status updates with other documents, such as MS Office files or PDF files. In contrast, Twitter limits its *Tweets* to text and single photos. Note that users have the possibility to enrich their text posts with their current location or a link to an uploaded video but not to disclose these elements outside of a Tweet. Posting on Instagram is - because of its specialized character - limited to single photos and videos, which can also be enriched with a location. Regardless of the complexity of the items, the user is always able to post them in his own domain (*disclosed data*) as well as to share the ones originally published by his contacts (*disseminated data*). On the contrary, posting in foreign domains without any content-wise limitations is only possible on Facebook (*entrusted data* and vice versa *incidental data*). However, Facebook's users can turn off this feature in their privacy settings and are able to control the visibility of incidental data on a fine-grained level. On Google+, LinkedIn, and Instagram, posting in foreign domains is limited to commenting on items disclosed by the domain's owner. These comments inherit the visibility of their corresponding items, giving the domain's owner full control over them. To be precise, LinkedIn users are able to post on organizations' pages. But what we understand by foreign domains are the domains of other human contacts. Publicly addressing contacts on Twitter is done by making a response (e.g., @johndoe), which does not appear in the contact's domain and therefore is not treated as entrusted data. With Instagram Direct, users are able to directly share photos and videos with 1 to up to 15 contacts. But similar to Twitter, media shared in this way stay in the user's domain and do not represent entrusted data. Moreover, Instagram Direct has to be seen as some kind of access control mechanism rather than posting in foreign domains. As incidental data is just the opposite of entrusted data, it is limitedly present on Google+, LinkedIn, and Instagram, and cannot be found on Twitter.

Summarizing the application of the taxonomy, most of its elements can be found on all of the five inspected OSNs demonstrating the suitability to describe the most important characteristics of OSNs. Furthermore, it demonstrated the taxonomy's capability of capturing different instantiations of a particular data type on different OSNs and the number of items contained in it. This is especially true for extended profile data where Twitter and Instagram provide only a few additional input fields because self-presentation is achieved by posting items and not by entering static profile information here, and where LinkedIn focuses more on work and science affiliated attributes because of the orientation toward professional networks. Another important observation is that Facebook has most features, especially concerning the distinctive data types, i.e., entrusted/incidental data and contextual data. Hence, there are more potential hazards for casual OSN users and more aspects that might be interesting for researchers in this area.

6.2 Mapping of Facebook data types into the taxonomy

In the previous section, we successfully applied the taxonomy to compare different OSNs, demonstrating its usefulness in one of its major usage scenarios. Subsequently, we evaluate its ability to cover all data types existing on OSNs by iterating overall OSN data types and mapping them into the taxonomy. Hereunto, Facebook as the currently largest and most popular OSN is used.

Facebook's Graph API reference [19] is employed in order to capture all data types. In particular, the focus is on the root node */user* including all of its edges and fields except for */user/home* and */user/notifications*, which are only used internally by Facebook for notification purposes related to other data types. Besides the */user* node, few other nodes contain data types related to the user. They are marked by an asterisk (*) in Figure 5, which provides an overview of the mapping of the aforementioned data types into our taxonomy. Note that the denomination of the root node */user* is omitted to improve readability. A prefixed slash (/) is used to distinguish edges from fields. In order to cover all user-related data types on Facebook, we

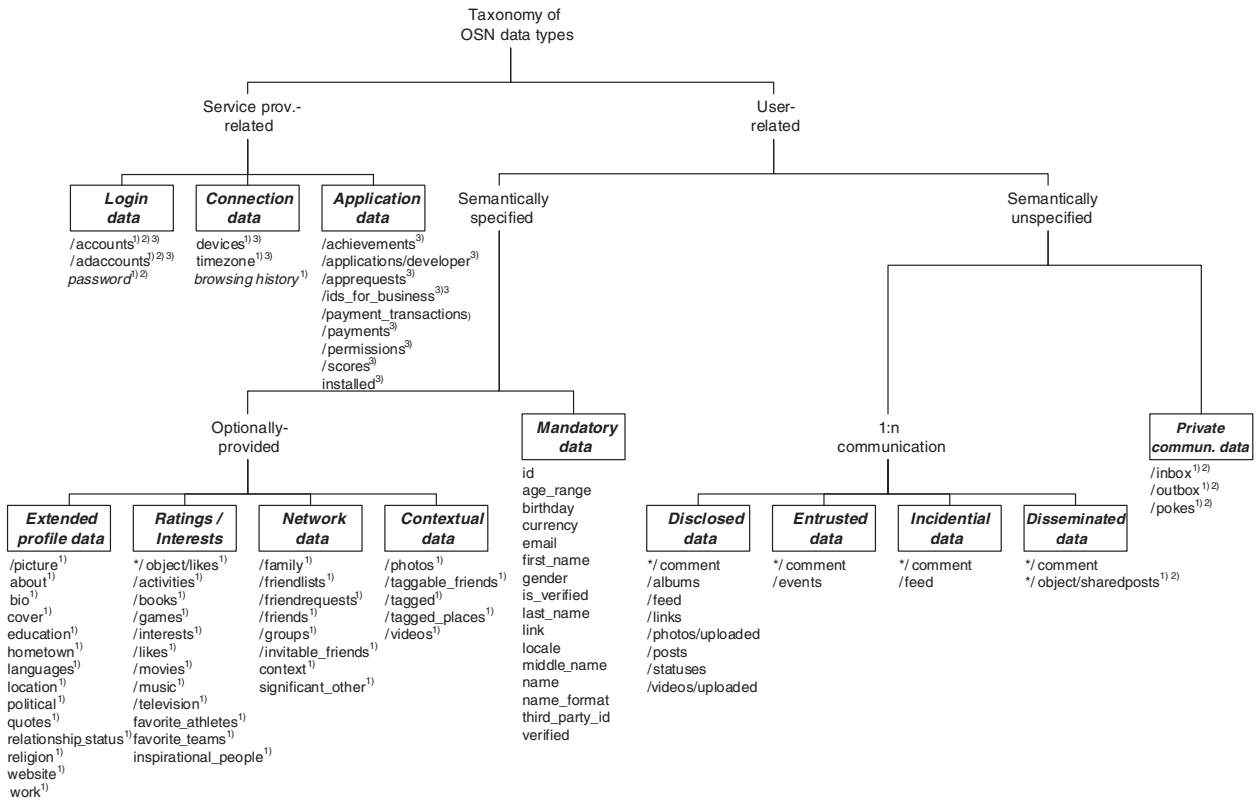

Figure 5 Mapping of Facebook data types into the proposed taxonomy.

further include additional data types (such as connection data and login data, e.g., the user's password and browsing history) into our analysis in addition to Facebook's Graph API reference.

Sections 2 and 4 already contain some exemplary remarks concerning the shortcomings of the previously published taxonomies. In order to illustrate their limitations in detail, in Figure 5, we show their inability to map some of Facebook's data types (i.e., no suitable data type exists in their taxonomies). Data types that cannot be mapped into the taxonomy introduced by Schneier [23] are marked with [1)], while those that cannot be mapped into the taxonomies introduced by Årnes et al.[29] and Beye et al. [28] are marked with [2)] and [3)], respectively. For instance, Schneier's taxonomy [23] can be criticized for not being able to cover login data, connection data, extended profile data, ratings/interests, network data, contextual data, disseminated data, and private communication data. The work presented by Årnes et al. [29] improves on this but still does not feature any equivalents for login data, disseminated data, and private communication data. The taxonomy of Beye et al. [28] has shortcomings related to its insufficient definitions of login credentials as well as behavioral information and, additionally, is missing an equivalent for application data. Although it is possible to map the data types related to $1:n$ communication and private communication into the categories proposed by Beye et al. [28], a lot of privacy-relevant information is lost when treating all of them simply under the general term *Messages*.

6.3 Evaluation of the privacy relevance metric

As pointed out, privacy relevance depends on multiple factors and varies for different OSNs and user profiles. To demonstrate our metric, we created a fictive user profile on Facebook that has 150 friends (n_f = 150). We mainly used Facebook's default privacy settings. However, we changed the visibility of extended profile data to 'friends' in order to demonstrate the differences. For login data, connection data, and application data, a value of 0 was defined since the service provider is assumed to be trusted. As weighting parameter we chose $\gamma_1 = 2$, $\gamma_2 = 1$, and $\gamma_3 = 2$. The reference context set is defined to entail no privacy risk ($c_1 = 0$, $c_2 =$ user, $c_3 = \{\text{user,user}\}$). For the friends of our fictive user, we assumed the same privacy settings, leading to the following values as depicted in Table 3.

The privacy relevance metric introduced could clarify that the different data types of OSN data classified in our taxonomy greatly vary in the privacy risks they carry.

Table 3 Proposed PRM for classes of the taxonomy by the example of a Facebook profile

Data types	PRM-value
Login data	0
Connection data	0
Application data	0
Mandatory data	0.4
Extended profile data	0.2
Ratings/ interests	0.6
Network data	0.4
Contextual data	0.6
Private commun. data	≈ 0
Disclosed data	0.2
Entrusted data	0.6
Incidental data	0.2
Disseminated data	0.8

These findings can help end users to increase their privacy awareness when dealing with OSN data, on the one hand. Particularly, the fact that the loss of control over own data carries high risk is an essential and valuable insight gained. The service providers, on the other hand, can benefit as well, since this metric may suggest which concepts should be revised in order to foster user privacy and improve the quality of their services with regard to privacy.

This quite generic approach also brings some limitations that need to be discussed. Currently, the metric does not consider different expectations of the end users toward privacy. To cover this issue, the context dimensions could be extended by user sensitivity. In this paper, however, our aim is to emphasize the difference of privacy risks attributed to distinct data types. Thus, we left varying end user perceptions out of consideration for the moment. The weight parameters were furthermore only exemplary chosen to demonstrate the functioning. To create a more realistic setting, empirical tests to find appropriate values should be conducted in the future.

7 Conclusions

Despite the growing body of research addressing OSN privacy issues, currently, *data* as one of the fundamental building blocks of OSN is not well understood. The lack of a generally accepted terminology and classification for existing data elements as well as the small number of publications considering implications of differing semantics of data types for social identity management further substantiates the argument.

Yet, data is at the core of any discussion of privacy issues on OSNs. Without a precise terminology and classification of all types of data on OSNs, it is difficult to unambiguously specify privacy-related problems which ultimately impedes the development of appropriate solutions.

To address these shortcomings, a taxonomy for OSN data types was developed in this paper. Based on a design-oriented methodology, first, the body of literature was analyzed to identify possible data elements and terminological inconsistencies. Subsequently, a hierarchically structured taxonomy was derived by studying fundamental user activities on OSNs and step-wise classifying the identified data types into non-redundant partitions. The discussion of data types revealed that privacy mainly depends on the interplay of a data element's content, the extent and granularity of user control, and its concrete implementation. Based on the understanding of privacy implications of different data types, a privacy relevance metric was proposed which allows to quantitatively assess privacy threats for a given context. The subsequent evaluation of applying the taxonomy to five major OSNs demonstrates its applicability to existing OSNs and reveals implementation-specific differences in privacy settings of various data types. A detailed evaluation using all of Facebook's data types further shows that it is possible to map all these data types into the taxonomy.

Endnote

[a]Note that this article is an extended version of the paper by Richthammer et al. presented at the 2013 ARES conference [41].

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The authors would like to thank the anonymous reviewers for their helpful comments. This research is partly funded by the European Union within the PADGETS project (no. 248920), the European Regional Development Funds (ERDF) within the SECBIT project (http://www.secbit.de/) and by the 'Bavarian State Ministry of Education, Science and the Arts' as part of the FORSEC research association (http://www.bayforsec.de/).

Received: 6 March 2014 Accepted: 10 July 2014
Published: 6 August 2014

References

1. D Boyd, N Ellison, Social network sites – definition, history, and scholarship. J. Comput.-Mediated Commun. **13**, 210–230 (2007)
2. MM Skeels, J Grudin, When social networks cross boundaries: a case study of workplace use of Facebook and LinkedIn, in *Proceedings of the International SIGGROUP Conference on Supporting Group Work* (ACM New York, 2009)
3. K Riemer, A Richter, Tweet inside: microblogging in a corporate context, in *Proceedings of the 23rd Bled eConference eTrust: Implications for the Individual, Enterprises and Society*, (2010)
4. J Park, S Kim, C Kamhoua, K Kwiat, Optimal state management of data sharing in online social network (OSN) services, in *Proceedings of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)* (IEEE Computer Society Washington, DC, 2012)
5. D Irani, S Webb, K Li, C Pu, Large online social footprints – an emerging threat, in *Proceedings of the 12th IEEE International Conference on*

Computational Science and Engineering (CSE) (IEEE Computer Society Washington, 2009)
6. T Burns, *Erving Goffman*. (Taylor & Francis, New York, 1991)
7. E Goffman, *The Presentation of Self in Everyday Life*. (Anchor, New York, 1959)
8. H Nissenbaum, *Privacy in Context - Technology, Policy, and the Integrity of Social Life*. (Stanford University Press, Stanford, 2010)
9. M Netter, M Riesner, G Pernul, Assisted social identity management, in *Proceedings of the 10th International Conference on Wirtschaftsinformatik* (AIS Electronic Library Zurich, 2011)
10. A Ho, A Maiga, E Aimeur, Privacy protection issues in social networking sites, in *Proceedings of the 2009 ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)* (IEEE Computer Society Los Alamitos, 2009)
11. M Madejski, M Johnson, S Bellovin, The failure of online social network privacy settings. Technical report, (Columbia University 2011). http://academiccommons.columbia.edu/download/fedora_content/download/ac:135407/CONTENT/cucs-010-11.pdf. Accessed 19 July 2014
12. C Ngeno, P Zavarsky, D Lindskog, R Ruhl, User's perspective: privacy and security of information on social networks, in *Proceedings of the 2nd IEEE International Conference on Social Computing (SocialCom)* (IEEE Computer Society Washington, DC, 2010)
13. M Netter, M Riesner, M Weber, G Pernul, Privacy settings in online social networks – preferences, perception, and reality, in *Proceedings of the 46th Hawaii International Conference on Systems Science* (IEEE Computer Society Washington, DC, 2013)
14. D Rosenblum, What anyone can know: the privacy risks of social networking sites. IEEE Secur. & Privacy. **5**, 40–49 (2007)
15. D Michalopoulos, I Mavridis, Surveying privacy leaks through online social networks, in *Proceedings of the 14th Panhellenic Conference on Informatics (PCI)* (IEEE Computer Society Washington, DC, 2010)
16. H Lipford, A Besmer, J Watson, Understanding privacy settings in Facebook with an audience view, in *Proceedings of the 1st Conference on Usability, Psychology, and Security (UPSEC)* (USENIX Association Berkeley, 2008)
17. B Carminati, E Ferrari, R Heatherly, M Kantarcioglu, B Thuraisingham, Semantic web-based social network access control. Comput. & Secur. **30**(2–3), 108–115 (2011)
18. W Luo, Q Xie, U Hengartner, FaceCloak: an architecture for user privacy on social networking sites, in *Proceedings of the 12th IEEE International Conference on Computational Science and Engineering (CSE)* (IEEE Computer Society Washington, DC, 2009)
19. Facebook Graph API Reference. https://developers.facebook.com/docs/graph-api. Accessed 4 July 2014
20. P Fong, M Anwar, Z Zhao, A privacy preservation model for Facebook-style social network systems, in *Proceedings of the 14th European Conference on Research in Computer Security (ESORICS)* (Springer Berlin, Heidelberg, 2009)
21. J Park, R Sandhu, Y Cheng, ACON: activity-centric access control for social computing, in *Proceedings of the 6th International Conference on Availability, Reliability and Security (ARES)* (IEEE Computer Society Washington, DC, 2011)
22. H Hu, G-J Ahn, J Jorgensen, Detecting and resolving privacy conflicts for collaborative data sharing in online social networks, in *Proceedings of the 27th Annual Computer Security Applications Conference (ACSAC)* (ACM New York, 2011)
23. B Schneier, A taxonomy of social networking data. IEEE Security & Privacy. **8**, 88–88 (2010)
24. M Riesner, G Pernul, Maintaining a consistent representation of self across multiple social networking sites – a data-centric perspective, in *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust* (IEEE Computer Society Washington, DC, 2012)
25. M Riesner, M Netter, G Pernul, An analysis of implemented and desirable settings for identity management on social networking sites, in *Proceedings of the 7th International Conference on Availability, Reliability and Security (ARES)* (IEEE Computer Society Washington, DC, 2012)
26. J Surma, A Furmanek, Improving marketing response by data mining in social network, in *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (IEEE Computer Society Los Alamitos, 2010)
27. J Zhang, J Tang, B Liang, Z Yang, S Wang, J Zuo, J Li, Recommendation over a heterogeneous social network, in *Proceedings of the 9th International Conference on Web-Age Information Management (WAIM)* (IEEE Computer Society Washington, DC, 2008)
28. M Beye, A Jeckmans, Z Erkin, P Hartel, R Lagendijk, Q Tang, Privacy in online social networks, in *Computational Social Networks: Security and Privacy* (Springer London, 2012)
29. A Årnes, J Skorstad, L Michelsen, Social network services and privacy. Technical report, Datatilsynet (2011). http://www.datatilsynet.no/global/english/11_00643_5_parti_rapport_facebook_2011.pdf. Accessed 19 July 2014
30. K Liu, E Terzi, A framework for computing the privacy scores of users in online social networks. ACM Trans. Knowl. Discov. Data. **5**(1), 1–30 (2010)
31. LA Cutillo, R Molva, M Önen, Analysis of privacy in online social networks from the graph theory perspective, in *Proceedings of the 2011 IEEE Global Telecommunications Conference (GLOBECOM)*, (2011), pp. 1–5
32. J Becker, H Chen, Measuring privacy risk in online social networks, in *Proceedings of W2SP 2009: Web 2.0 Security and Privacy*, (2009). http://w2spconf.com/2009/papers/s2p2.pdf. Accessed 19 July 2014
33. A Hevner, S March, J Park, S Ram, Design science in information systems research. MIS Q. **28**, 75–105 (2004)
34. K Peffers, T Tuunanen, M Rothenberger, S Chatterjee, A design science research methodology for information systems research. J. Manag. Inform. Syst. **24**, 45–77 (2007)
35. J Venable, J Pries-Heje, R Baskerville, A comprehensive framework for evaluation in design science research, in *Proceedings of the 7th International Conference on Design Science Research in Information Systems: Advances in Theory and Practice* (Springer Berlin, Heidelberg, 2012)
36. A Hevner, S Chatterjee, *Design Research in Information Systems: Theory and Practice*. (Springer, New York, 2010)
37. M Ziegele, O Quiring, Privacy Online. Perspectives on Privacy and Self-Disclosure in the Social Web (Springer Heidelberg, 2011)
38. P Eckersley, How unique is your web browser?, in *Proceedings of the 10th International Conference on Privacy Enhancing Technologies* PETS'10 (Springer Berlin, Heidelberg, 2010), pp. 1–18
39. B Carminati, E Ferrari, M Viviani, Security and Trust in Online Social Networks. Synthesis Lectures on Information Security, Privacy, and Trust. **4**(3), 1–120 (2013)
40. G Wondracek, T Holz, E Kirda, C Kruegel, A practical attack to de-anonymize social network users, in *Proceedings of the 2010 IEEE Symposium on Security and Privacy. S & P 2010* (IEEE Computer Society Washington, DC, 2010), pp. 223–238
41. C Richthammer, M Netter, M Riesner, G Pernul, Taxonomy for social network data types from the viewpoint of privacy and user control, in *Proceedings of the 8th International Conference on Availability, Reliability and Security (ARES)* (IEEE Computer Society Washington, DC, 2013), pp. 141–150

doi:10.1186/s13635-014-0011-7
Cite this article as: Richthammer *et al.*: **Taxonomy of social network data types.** *EURASIP Journal on Information Security* 2014 **2014**:11.