Taxonomy for Social Network Data Types from the Viewpoint of Privacy and User Control

Christian Richthammer, Michael Netter, Moritz Riesner and Günther Pernul
Department of Information Systems
University of Regensburg
Regensburg, Germany
{*firstname.lastname*}*@wiwi.uni-regensburg.de*

***Abstract*—The growing relevance and usage intensity of Online Social Networks (OSNs) along with the accumulation of a large amount of user data has led to privacy concerns among researchers and end users. Despite a 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, hence leading to confusion in related discussions. This paper proposes a taxonomy for data types on OSNs based on a thorough literature analysis and a conceptualization of typical OSN user activities. It aims at clarifying discussions among researchers, benefiting comparisons of data types within and across OSNs and at educating the end user about characteristics and implications of OSN data types. The taxonomy is evaluated by applying it to four major OSNs.**

***Keywords*-Online Social Networks, Taxonomy, Privacy, Data Types, Social Identity Management, Classification**

I. 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 to 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].

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. It is possible to derive rich profiles of the 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. This leads to the 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, which is also referred to as *Social Identity Management* (SIdM) [6].

Prompted by these developments, privacy concerns have been voiced by researchers. Numerous studies have been conducted on privacy issues on OSNs in general [7], [8] as well as on people's awareness in this context [9], [10] and on potential hazards [11], [12]. Proposals for improving the user's understanding of disclosed information [13] and improving privacy protection on OSNs [14] 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 API of the popular site Facebook reveals as many as 83 distinct data elements that can be associated with a user [15]. Works in assessing the access control models in OSNs (e.g. [16], [17]) do not differentiate between different attributes of the user identity while others only focus on singular aspects such as the owner and creator of items [18]. Still, it is seldom, or only briefly [19] considered that attribute implementations on OSNs vary widely in implementation, semantics, applicable policies [20] and privacy controls [21] and thus carry far-reaching implications for the user.

This paper aims at tackling the possible results of the lack of a generally accepted terminology for describing and differentiating data types on OSNs by developing and proposing a detailed taxonomy. 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 implications of dealing with particular OSN data types.

The remainder of the paper is organized as follows. Work related to classifying OSN data types is discussed and compared to our contribution in the Section II. The scope and methodology of the research are defined in Section III. The proposed taxonomy is introduced in Section IV accompanied by an analysis of related literature and a conceptualization of fundamental OSN user activities involving user data. The taxonomy is evaluated in Section V by applying it to four major OSNs. Section VI concludes the paper.

II. Related Work

This section provides an overview of related work regarding the study of user activities on OSNs and the conceptualization of data types.

In [22] and [23], fundamental user activities on OSNs are described. The focus of the analysis in [22] is on user activities accustomed to a small community of OSN, which is why it does not allow to draw generic conclusions. The discussion in [23] 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.

In [7], two different approaches for categorizing data are mentioned. 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 in [7], 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 in [4] lacks a discussion of activity-related data types and solely focuses on the two dimensions mentioned before.

Beye et al. [24] 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, the one in [24] contains a well-founded explanation on the origin of the data types. However, their definition (e.g. the definition of the data type *Messages*) is 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 the other authors who needed their classifications only as a basis for further examinations, Schneier [19] 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. IP address) can be arranged. Årnes et al. [25] pick up the ideas proposed in [19]. Although being a meaningful extension to [19], 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.

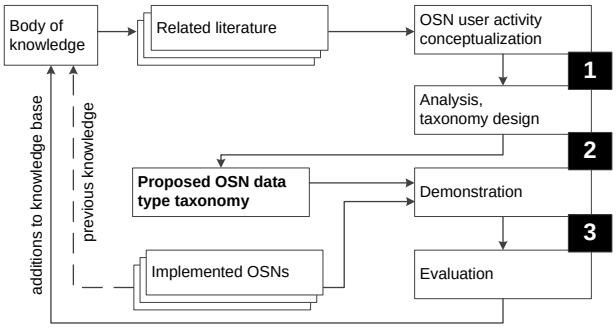

Figure 1. Research model

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.

III. Research Scope and Approach

A. Problem Scope

This work aims at developing a taxonomy for describing and classifying data types on OSNs, thereby benefiting three areas. A 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 in 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 seperately 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.

B. 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 [26], [27] is applicable to the problem scope.

Table I
MAPPING BETWEEN DATA TYPES OF RELATED CLASSIFICATIONS AND THIS WORK

Schneier [19]	Årnes et al. [25]	**Proposed Taxonomy**	Beye et al. [24]
		Login data	Login credentials
Service data	Mandatory profile data	**Mandatory data**	
	Extended profile data	**Extended profile data**	Profiles
	Personal network data	**Network data**	Connections
			Groups
		Ratings / interests	Preferences / ratings / interests
		Private communication data	
Disclosed data	Self published data at home	**Disclosed data**	Messages
Entrusted data	Self published data away	**Entrusted data**	Multimedia
Incidental data	Other users‘ data	**Incidental data**	
		Disseminated data	
	Metadata	**Contextual data**	Tags
Behavioral data	Behavioral data	**Application data**	Behavioral information
	Connection data	**Connection data**	
Derived data	Derived data		

Conception: ——— corresponding – – – – deviating

For conducting design-oriented research, a process model consisting of six steps – Problem identification, Elicitation of solution objectives, Solution design, Demonstration, Evaluation, Communication – has been proposed [27].

The research approach employed in this work adapts the process proposed in [27]. 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 related literature is analyzed in regards to possible elements of an OSN attribute taxonomy (Section IV). 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 [27].

Evaluation is deemed as a *central and essential activity* [28] and a *key element* [29] in design-oriented research. Correspondingly, the design research process [27] contains both a demonstration and a dedicated evaluation step. The taxonomy is demonstrated (step 2 in Figure 1) by applying it to four major OSNs and identifying actually implemented data types for each element of the taxonomy (Section V). On this basis, the taxonomy is evaluated (step 3 in Figure 1) in regards to the contribution to its three objectives that were stated above.

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

IV. 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 II, a thorough literature analysis reveals in essence the following three related approaches: Schneier [19], its refinement by Arnes et al. [25] and the classification by Beye et al. [24]. Table I 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 Table I leads to several observations: Firstly, it reveals that to some extent terminology is not consistently used, such as the different understanding of behavioral data [19], [25] and behavioral information [24]. Secondly, a general lack of granularity can be attributed to some existing data type definitions, as observable in the generic conceptualization of profiles in [24]. Consequently, it is difficult to precisely specify data elements as needed in scientific discussions. Lastly, some works either do not cover all available data types, such as the missing specification of data related to the connection with other users in [19] or focus on data elements whose existence is difficult to verify (e.g. the probability-based derived data in [19], [25] 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 2 illustrates OSN entities and possible activities. As can be seen, most activities are either initiated by the user or one

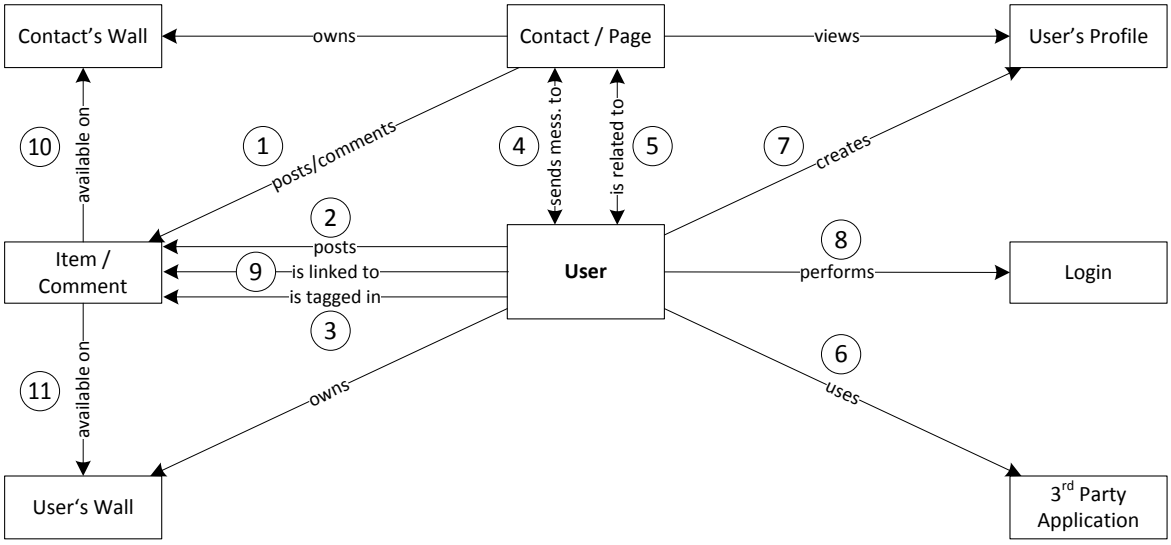

Figure 2. Fundamental user activities on OSNs

of his contacts. The subsequent elicitation of data types will refer to the numbered steps in Figure 2 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 [30]: *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 stakeholder, personal data is used for the purpose of SIdM. In the following, accruing data types for each stakeholder are discussed in detail.

A. 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 service usage. Drawing on user activities identified in Figure 2, several service provider-related activities can be identified. In the following, data emerging from these activities is classified into three separate data types.

Login data. OSN service usage requires prior user authentication to prevent identity theft, which is represented by activity 8 in Figure 2 and is consistent with the respective data type in [24] (cf. Table I). 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, which eventually leads to the compilation of a more comprehensive profile.

Connection data. While not OSN specific, requesting – i.e. connecting to and using – Internet-based services (activity 8 in Figure 2) leads to a variety of digital traces created by protocols on several layers of the OSI model. Table I shows that the definition is consistent with [25], while a broader conceptualization is used in [24]. 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.

Application data. Besides OSN platform usage, data originating from the use of third party services (activity 6 in Figure 2) running within the boundaries of the OSN platform or having API access can be differentiated. None of the related works 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. Depending on the data instance, privacy implications may range from none to serious.

B. 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:

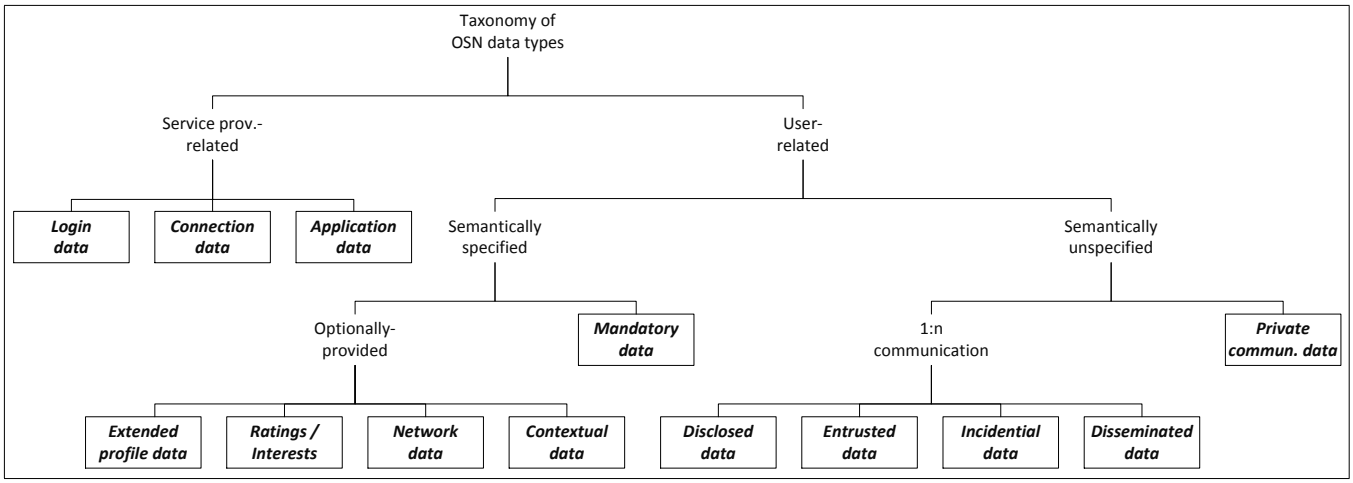

Figure 3. Proposed taxonomy of OSN data types

Semantically specified and *semantically unspecified* data. The first category refers to data instances that have a clearly defined meaning and its 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.

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.

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 2), which – except for the term – corresponds with [19] and [25] (cf. Table I). 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.

Optionally-provided Data:

Besides 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: 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 2) is guided by a form that contains input fields for attribute types like address, education, favorite music, favorite films, hobbies, interests, etc. The profile picture, which is a common feature of OSNs, is also arranged in this category. According to Table I, this conceptualization is in line with [19] and [25], while the profiles category in [24] 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 extended profile data that allows for a rich description, the study of user activities (activities 5 and 9 in Figure 2) reveals that binary or predefined multi-value attributes related to existing entities such as pages and shared items are used to refine how one is seen by others (e.g. by liking favorite bands). Corresponding with [24] in Table I, 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, whereas privacy implications depend on default or available visibility settings.

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 2). The collection of all connections of a particular user is often referred to

as his social graph [24] and describes data concerning the network the user has built around himself on the OSN, which conforms to the definition of [25] as presented in Table I. 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. As knowledge of a user's social graph allows to draw inferences about his identity, access to network data significantly impacts privacy.

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 2). 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 Table I shows that while corresponding with [24], this type of data is only partly covered in [19] and [25]. Contextual information poses a serious privacy risk as information that was previously not machine-processable (e.g. searchable) is made explicitly known to the system and consequently impacts a user's representation on the social network. Yet, an OSN's implementation of this data type needs to be thoroughly examined to estimate its concrete privacy implications.

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. 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.

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 2), which is only partly covered in [24] as illustrated in Table I. 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.

Table II
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

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 II). As can be seen from Table I, the first three data types subsequently discussed are based on [19] and [25].

Disclosed data. A frequent user activity on OSNs is to post information on one's wall (activities 2 and 11 in Figure 2). In conceptual terms, the data is generated and published by a user in his own domain. From a privacy perspective, the user has full control over the visibility within the limits of the concrete implementation as no other user is affected with the shared item.

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 2), 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 2), i.e. the contact is both 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 II, user-generated data elements are considered that are further disseminated by a contact within his own domain (activities 1 and 10 in Figure 2). 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 in [18], 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.

C. Summary

Figure 3 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 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.

In the subsequent section, the application of the taxonomy to four major OSNs is demonstrated and their usefulness to address the objectives stated in Section I is evaluated.

V. Evaluation

A. Evaluation Approach

One approach to provide a suitable evaluation is to focus on demonstration, which is described as a light-weight evaluation in [28]. In [27], the task of demonstration is defined as showing the use of the artifact to solve one or more instances of the problem. Transferred to the given context, applying the taxonomy to different OSNs seems appropriate.

In the following, four major OSNs – Facebook, Google+, Twitter, LinkedIn – 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.

B. Application of the Taxonomy to OSNs

Table III gives an overview on data types on OSNs as available on February 15, 2013.

1) Service Provider-related Data Types:

Login data can be found on all OSNs. The four inspected ones 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.

Connection data is collected by all OSNs. In order to inspect the items arranged in this category, the privacy policies of the four 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 three categories (device information, log information, location information) and then mentions the most important examples. 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 or do not define any broad categories at all. Moreover, the four inspected OSNs differ in the examples they list and their level of detail.

Application data is available on all four 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 is bigger by far as there are a lot of providers for games. As mentioned in Section IV, games may process credit card information because of In-App purchases whereas website connectors and smartphone apps do not collect additional data 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.

2) User-related Data:

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+. In contrast, LinkedIn forces the user to indicate his job status, 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 (indicated by * in Table III), giving the user the ability to alleviate certain threats (e.g. social engineering attacks with the help of personalized emails).

Which **extended profile data** is ultimately present in addition to the profile photo and the cover photo – each inspected OSN uses these concepts – 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

Table III
DEMONSTRATION OF THE TAXONOMY ON FACEBOOK, GOOGLE+, TWITTER AND LINKEDIN

Data types	Facebook	Google+	Twitter	LinkedIn
Login data	Email, phone, password	Email, password	Email, username, password	Email, password
Connection data	Device information, log information, location information, cookies	Device information, log information, location information, cookies	Device information, log information, location information, cookies	Device information, log information, location information, cookies
Application data	Usage statistics, credit card information	Usage statistics, credit card information	Usage statistics	Usage statistics
Mandatory data	Name, email*, birthday*, gender*	Name, email*, birthday*, gender*	Name, email*	Name, email, job status
Extended profile data	Several general-purpose input fields	Several general-purpose input fields	Three single input fields (location, website, bio)	Several professionally-related input fields
Ratings/interests	Page, status/photo/video	Page, status/photo/video	Verified account, Tweet	Company, status
Network data	Unidirectional, bidirectional	Unidirectional	Unidirectional	Bidirectional
Contextual data	Tag in status/comment, on photo, at location	Tag in status/comment, on photo, at location	Mention in Tweet	n/a
Private communication data	Private message, video chat, poke	Private message, video chat	Private message	Private message
Disclosed data	Text post, photo (album), video, check-in	Text post, photo (album), video, check-in	Text post, single photo	Text post
Entrusted data	*See disclosed data*	*Restricted to comments on disclosed data*	n/a	*Restricted to comments on disclosed data*
Incidental data	*See disclosed data*	*Restricted to comments on disclosed data*	n/a	*Restricted to comments on disclosed data*
Disseminated data	*See disclosed data*	*See disclosed data*	*See disclosed data*	*See disclosed data*

data elements to refine one's profile but with a professional focus (such as experience and skills). 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. 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 and LinkedIn, 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 companies, respectively. Moreover, the focus lies more on staying informed about these pages rather than publicly demonstrating certain interests. 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, at least the 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). As shown in Table III, 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 always reveals 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 does not support this feature at all, which can be traced back to the previously mentioned motivation of establishing a network of professionals who have little use for this.

Private communication data can also be found on all OSNs because sending private messages is always possible. 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. LinkedIn calls their private messages *InMail* and does not offer them to users

with basic accounts. 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 IV.

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 can 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). 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*. On LinkedIn, it is only possible to post text which, however, can also be enriched with a link to other objects. 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+ and LinkedIn, 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. 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. As incidental data is just the opposite of entrusted data, it is limited on Google+ and LinkedIn, and cannot be found on Twitter.

C. Evaluation Summary

Summarizing the application of the taxonomy, most of its elements can be found on all of the four inspected OSNs demonstrating the suitability to describe the most important characteristics of OSNs. Furthermore, the evaluation 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 provides only three additional input fields because of the nonexistent desire for self-presentation and where LinkedIn focuses more on work and science affiliated attributes because of the orientation towards 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.

VI. Conclusion

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 on these sites further substantiate the argument.

Yet, data is at the core of any discussion of privacy issues on OSN. Without a precise terminology and classification of all types of data on OSN 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 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. The subsequent evaluation of applying the taxonomy to four major OSNs demonstrates its applicability to existing OSNs and reveals implementation-specific differences in privacy settings of various data types.

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) and the European Regional Development Funds (ERDF) within the SECBIT project.

References

[1] d. boyd and N. Ellison, "Social Network Sites – Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, vol. 13, pp. 210–230, 2007.

[2] M. M. Skeels and J. Grudin, "When Social Networks Cross Boundaries: a Case Study of Workplace Use of Facebook and Linkedin," in *Proc. of the International SIGGROUP Conference on Supporting Group Work*. ACM, 2009.

[3] K. Riemer and A. Richter, "Tweet Inside: Microblogging in a Corporate Context," in *Proc. of the 23rd Bled eConference eTrust: Implications for the Individual, Enterprises and Society*, 2010.

[4] J. Park, S. Kim, C. Kamhoua, and K. Kwiat, "Optimal State Management of Data Sharing in Online Social Network (OSN) Services," in *Proc. of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. IEEE Computer Society, 2012.

[5] D. Irani, S. Webb, K. Li, and C. Pu, "Large Online Social Footprints – An Emerging Threat," in *Proc. of the 12th IEEE International Conference on Computational Science and Engineering (CSE)*. IEEE Computer Society, 2009.

[6] M. Netter, M. Riesner, and G. Pernul, "Assisted Social Identity Management," in *Proc. of the 10th international conference on Wirtschaftsinformatik*, 2011.

[7] A. Ho, A. Maiga, and E. Aimeur, "Privacy Protection Issues in Social Networking Sites," in *Proc. of the 2009 ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, 2009.

[8] M. Madejski, M. Johnson, and S. Bellovin, "The Failure of Online Social Network Privacy Settings," Columbia University, Tech. Rep., 2011.

[9] C. Ngeno, P. Zavarsky, D. Lindskog, and R. Ruhl, "User's Perspective: Privacy and Security of Information on Social Networks," in *Proc. of the 2nd IEEE International Conference on Social Computing (SocialCom)*. IEEE Computer Society, 2010.

[10] M. Netter, M. Riesner, M. Weber, and G. Pernul, "Privacy Settings in Online Social Networks – Preferences, Perception, and Reality," in *Proc. of the 46th Hawaii International International Conference on Systems Science*, 2013.

[11] D. Rosenblum, "What Anyone Can Know: The Privacy Risks of Social Networking Sites," *IEEE Security & Privacy*, vol. 5, pp. 40–49, 2007.

[12] D. Michalopoulos and I. Mavridis, "Surveying Privacy Leaks Through Online Social Networks," in *Proc. of the 14th Panhellenic Conference on Informatics (PCI)*. IEEE Computer Society, 2010.

[13] H. Lipford, A. Besmer, and J. Watson, "Understanding Privacy Settings in Facebook with an Audience View," in *Proc. of the 1st Conference on Usability, Psychology, and Security (UPSEC)*. USENIX Association, 2008.

[14] W. Luo, Q. Xie, and U. Hengartner, "FaceCloak: An Architecture for User Privacy on Social Networking Sites," in *Proc. of the 12th IEEE International Conference on Computational Science and Engineering (CSE)*. IEEE Computer Society, 2009.

[15] "Facebook Graph API Reference," https://developers.facebook.com/docs/reference/api/, accessed on February 4th, 2013, 2013.

[16] P. Fong, M. Anwar, and Z. Zhao, "A Privacy Preservation Model for Facebook-Style Social Network Systems," in *Proc. of the 14th European Conference on Research in Computer Security (ESORICS)*. Springer, 2009.

[17] J. Park, R. Sandhu, and Y. Cheng, "ACON : Activity-Centric Access Control for Social Computing," in *Proc. of the 6th International Conference on Availability, Reliability and Security (ARES)*. IEEE Computer Society, 2011.

[18] H. Hu, G.-J. Ahn, and J. Jorgensen, "Detecting and resolving privacy conflicts for collaborative data sharing in online social networks," in *Proc. of the 27th Annual Computer Security Applications Conference (ACSAC)*. ACM, 2011.

[19] B. Schneier, "A Taxonomy of Social Networking Data," *IEEE Security & Privacy*, vol. 8, pp. 88–88, 2010.

[20] M. Riesner and G. Pernul, "Maintaining a Consistent Representation of Self across Multiple Social Networking Sites – A Data-centric Perspective," in *Proc. 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, 2012.

[21] M. Riesner, M. Netter, and G. Pernul, "An Analysis of Implemented and Desirable Settings for Identity Management on Social Networking Sites," in *Proc. of the 7th International Conference on Availability, Reliability and Security (ARES)*, 2012.

[22] J. Surma and A. Furmanek, "Improving Marketing Response by Data Mining in Social Network," in *Proc. of the 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE Computer Society, 2010.

[23] J. Zhang, J. Tang, B. Liang, Z. Yang, S. Wang, J. Zuo, and J. Li, "Recommendation over a Heterogeneous Social Network," in *Proc. of the 9th International Conference on Web-Age Information Management (WAIM)*. IEEE Computer Society, 2008.

[24] M. Beye, A. Jeckmans, Z. Erkin, P. Hartel, R. Lagendijk, and Q. Tang, "Privacy in Online Social Networks," in *Computational Social Networks: Security and Privacy*. Springer, 2012.

[25] A. Årnes, J. Skorstad, and L. Michelsen, "Social Network Services and Privacy," Datatilsynet, Tech. Rep., 2011. [Online]. Available: http://www.datatilsynet.no/global/english/11_00643_5_parti_rapport_facebook_2011.pdf

[26] A. Hevner, S. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Quarterly*, vol. 28, pp. 75–105, 2004.

[27] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee, "A Design Science Research Methodology for Information Systems Research," *Journal of Management Information Systems*, vol. 24, pp. 45–77, 2007.

[28] J. Venable, J. Pries-Heje, and R. Baskerville, "A Comprehensive Framework for Evaluation in Design Science Research," in *Proc. of the 7th International Conference on Design Science Research in Information Systems: Advances in Theory and Practice*. Springer, 2012.

[29] A. Hevner and S. Chatterjee, *Design Research in Information Systems: Theory and Practice*. Springer, 2010.

[30] M. Ziegele and O. Quiring, "Privacy in Social Network Sites," in *Privacy Online. Perspectives on Privacy and Self-Disclosure in the Social Web*. Springer, 2011.