How I and others can link my various social network profiles as a basis to reveal my virtual appearance

Sebastian Labitzke, Jochen Dinger, Hannes Hartenstein
sebastian.labitzke@kit.edu, jochen.dinger@fiducia.de, hannes.hartenstein@kit.edu

Abstract: Today, many Internet users take part in online social networks (OSNs) like Facebook, Xing, LinkedIn, MySpace, studiVZ, and others. However, the motives for participation are multifaceted and stand in contrast to the motivation not to join OSNs because of the potential danger with respect to privacy aspects. This risk is already perceived by some OSN users but still underestimated. The naive handling with personal data inside such networks combined with many possibilities to link several OSN profiles of a single user paves the way for third-parties to gather the comprehensive virtual appearance of a person. Therefore, to turn the tables, such correlations of information might be useful to demonstrate to a user a measure of how linkable his personal data is, given the current exposure of his public data in OSN profiles. In this paper we point out some astonishing simple ways to automatically link data from different OSNs. As we want to and have to act compliant to the German data protection laws, we analyze corresponding requirements and sketch our concept for compliant statistical sampling. We further show preliminary results regarding successful profile correlations based on, even under legal and technical constraints, extracted friends lists.

1 Introduction

The online social network (OSN) Facebook is actively used by over 500 million people (logged in within the last 30 days) and grew by 500 percent within the last two years\(^1\). The number of members of many other social networks is growing as well (e.g. [MKG+08]), users are typically member of more than a single OSN, and in every OSN people publish a lot of personal data. For instance, Facebook claims that an average user uploads and shares 90 pieces of content each month. The main reasons for joining OSNs and sharing such amount of data are the possibility to meet so called “friends” (and strangers as well), the chance to keep in touch with those people, and not least the opportunity to express oneself through uploaded photos, videos, comments, or other small contributions. Also the asynchronous means of communications make OSNs attractive for many people.

Quite clearly, thoughtless sharing of personal data bears certain risks. The obvious question is: why do many users not think about these risks while providing personal data? Answers can be found in behavioral psychology with respect to rewards and punishments of human actions. The exposure of personal data ensures direct rewards in terms of enabling the user to participate and take part in the OSN. The more data, and consequently content, a user shares, the better he can be located and identified inside the network and

\(^1\)http://www.facebook.com/press/info.php?timeline
his profile will arise more interest. Reactions in terms of comments on the user’s new content intensify these rewards. Furthermore, by sharing content about oneself an individual distinctive urge for self-representation may be satisfied, which can also be a desired reward for users. Thus, the motivation to publish content is increased by the expectation of benefits. Psychologists refer to this as “positive reinforcement” of the motivation to do something [Zim04].

In order to develop users’ risk awareness regarding privacy aspects, a negative reinforcement, i.e., an increased motivation not to share personal data in the sense of conditioning of users, is necessary. Referring to this, it is known that the time gap between an action and a consecutive set of rewards and punishments, as well as their nature and severity, shapes future behavior. This time gap is known as “deferred gratification” ([MM83], [Mis74]). In the context of OSNs, rewards occur on a short and punishments, if any at all, on an often longer time scale. As a consequence, punishments and therefore sustainable recognitions of the potential of danger are less effective on the motivation to share personal data in OSNs than the above discussed rewards a user will get promptly. Such recognitions or rather a conditioning can be achieved by different approaches. In general, promising options are the intensification of the consequences of an action and/or the reduction of the time gap between the action (here: providing personal data) and the occurrence of punishments. Psychologists refer to the conditioning by regulated rewards and punishments as “operant conditioning”. However, it is not clear which punishments should be provoked to achieve an effect on users’ behavior in OSNs. Furthermore, punishments would unnecessarily cause detriments to users as well as OSNs would not endorse any sanctions on publishing data as this stands obviously in contrast to their business model.

Meanwhile, discussions and reports about privacy with respect to OSNs in the media are urging more users to make use of privacy settings that are provided by OSNs to hide parts of a profile from strangers. This is also confirmed by a German survey, in which 1,208 teenagers between 12 and 19 years were asked about their behavior in OSNs [jim10]. In the recent survey two thirds confirmed that they hide parts of their profiles from strangers, while in an earlier study (2009) less than a half of those interviewed stated that they make use of privacy settings. In 2008 only 25 percent of Facebook users made use of restricted privacy settings [KW08]. While more and more OSN members have adjusted their privacy settings, the fact that a large number of them are still unduly generous in sharing personal data applies to most OSNs. In particular, many users still do not hide parts of their OSN profiles or do not hide every published information from strangers (cf. Section 2). Moreover, people are often member of more than a single OSN, even though Facebook acquires a remarkable market share in recent time. Users’ multiplicity of profiles are comprehensible due to the different objectives of current OSNs (business and private networks, networks for fellow students etc.). Furthermore, Torkjazi et al. show that OSNs usually get into a downturn of interest after a major increase of members [TRW09] which prompt users to register in additional OSNs. According to that, if a person has more than one OSN profile, it is often easy to merge these profiles on the basis of provided information. Such linking of profiles can be used to gather and associate various shared information [MV09] by third-parties, but also to illustrate to a user how his comprehensive virtual appearance looks like despite possibly adjusted privacy settings. In contrast to inapplicable operant
conditioning, such an illustration might be a far better conditioning approach to encourage users to hide their personal data in OSNs. The aim is to demonstrate the potential of danger due to the disclosure of measures of how easy information of OSN profiles can be associated.

Furthermore, such possibility to reveal users’ own virtual appearance by linking OSN profiles might constitute a basis for a novel identity management system with which users will be able to monitor the proliferation of their published information on their own in the future. On the contrary, in enterprise identity management systems the “flow” and processing of identity related information is typically monitored and governed by various oversight boards. While the publishing of personal information in OSNs is essentially at the users discretion, the flow of such information and their appearance in several systems can not be monitored or further managed by users, today.

In this paper, we take a first step in the direction of such an identity (self-)management component. Therefore, we assess which kind of personal information is published by users and discuss possibilities to correlate users’ profiles from different OSNs based on this data. More precisely, we unfold the availability of personal attributes in different OSNs that can be used to link OSN profiles. Our investigations demonstrate that many profiles still provide sufficient information to establish such links between OSN profiles. This is particularly critical if a user makes use of restrictive privacy settings in one OSN and non-restrictive settings in another. While the restrictive settings shall help to protect provided data, the settings might be ineffective due to the linkability of both profiles. The still existing risk of an attack on privacy can apparently be overlooked by users. Additionally, we show how crucial friends lists are in order to link OSN profiles. Our statistical analysis on the linkability of profiles on the basis of friends list reveals the “fingerprint” character of such individual lists of names. To get such specific knowledge we performed statistical analysis on the users’ data without archiving any reference to OSN profiles or sufficient information to identify a natural person. According to our analysis of German Data Protection Act and to the best of our knowledge, the study (presented in Section 5) is compliant to the German law and ensures users’ privacy. Hence, in this paper a “lower bound” is made explicit that indicates what is possible by correlation of OSN profiles without breaking the law. However, it is easy to imagine the potentialities for people who do not care about legal aspects. Such people are merely faced with some technical measures implemented by OSNs to avoid software-based extractions of personal data by third-parties. In the course of this paper we additionally point out the deficiency of such measures that in turn should not prevent automatic data extraction.

Section 2 presents an overview of related work that addresses the amount of voluntarily published data in OSNs and the linking of profiles. Section 3 discusses restrictions for studies on personal data of OSN profiles on the basis of the German Data Protection Act and introduces requirements to act compliant. Section 4 points out technical measures of OSNs to prevent extractions of personal data and assesses their effectiveness. We then look at the concept for an implementation to carry out the study, at statistical results concerning the concrete data users provide in OSNs, and at a preliminary estimated quality of profile correlations on the basis of friends lists in Section 5. Section 6 concludes the paper.
2 Related work

In the past, several researchers investigated the publicly available amount of information inside OSNs. As early as 2005 the use of privacy settings of 4,000 Carnegie Mellon University students were determined by Gross and Acquisti [GA05]. They discovered that 89% of analyzed users provided their full real names and 90% a profile image. Furthermore, 80% of analyzed profiles contained information that can be used to identify a person. Merely 1.2% of users configured that they cannot be found by using the OSN search and only 0.06% restricted the visibility of their profile. In 2007 30,000 Facebook profiles were analyzed by Lampe et al. [LES07]. At that time, already 19% made use of privacy settings but in average 59% of provided fields inside a profile were filled with personal information. These studies combined with parts of our results presented in Section 5 show the evolution of risk awareness and sensitivity of OSN members.

In [KW08] and [KW10a] Krishnamurthy and Wills analyzed the default configurations of privacy settings, which information can be hidden from strangers, and how many users adjusted the privacy settings in 12 OSNs. They pointed out that in most OSNs personal data is classified in categories. Users are able to adjust the visibility of these categories instead of more fine grained access configurations. Furthermore, in [KW08] it is shown that only 1% of Twitter users, 25% of Facebook users, and 21% of MySpace users have adjusted their privacy settings whereas in [KW10a] the authors disclose the technical possibilities for third-parties to gather personal information of OSN users. In contrast to our investigation, they analyzed a way to get information about OSN profiles of a visitor of a website by tracking him surfing through different websites and OSNs. For a follow-up study and on the basis of the knowledge of their previous studies, the same authors compared 13 mobile OSNs in 2010 [KW10b]. They demonstrated that friends lists are available in six mobile variants of OSN websites, in other six by default and in one network friends lists are never publicly available. These three papers show that a lot of information (especially friends lists) is still accessible for third-parties. This in turn underpins our thesis that average OSN profiles still provide enough information to link several profiles of a single person.

Besides users behavior in OSNs with respect to the amount of publicly available personal data, the linkability of OSN profiles has been examined. The authors of [MV09] demonstrated which criteria are suitable to effectively identify a user in an OSN. Therefore, they extracted several attributes of OSN profiles and, afterwards, they used these attributes to search a profile inside another network. With this study the authors pointed out that 25.2% of all analyzed Facebook profiles overlapped with a single MySpace profile. In contrast to our study, the authors used as much information as needed to identify a person inside a network, whereas we quantify the potential of single attributes regarding their suitability to link one OSN profile to another profile in a different OSN. Thereby, our aim is to identify information in OSN profiles that is more worthy of protection than others. The aim of the authors work was to provide a system to locate individuals in Facebook and MySpace as well as to figure out how many users are member of both of these OSNs. Furthermore, the authors stated that friends lists of two profiles of the same natural person just rarely overlap. We show in this paper that this more or less rare overlap is yet sufficiently large to distinguish it from overlaps of profiles from different people.
3 Compliance requirements on statistical analysis

For a tool that should enhance users’ privacy as well as for an underlying study, compliance with privacy regulation is a must. Therefore, this section discusses legal requirements which have to be met for statistical sampling that ensures privacy of OSN members.

Derived from German constitution (Art.2x1 in conjunction with Art.1x1, GG) and German Data Protection Act (BDSG) the right to privacy and the right to self-determination with respect to personally identifiable information are established. According to the BDSG §4.3 and §4a.1 personal information may only be used for a desired purpose and only an owner of information, respectively the OSN user himself, can approve data processing by third-parties. Users’ data inside OSNs are publicly available to take part of the network and activities inside the network. Such data is not allowed to be used by third-parties for further processing. Some OSNs explicitly state in their terms and conditions that automated extractions of data are prohibited to protect OSN members.

According to our analysis of BDSG, it is not possible to legitimize any extraction, processing, and archiving of users’ profiles and contained personal data without a previously stated agreement of any affected user (permission facts). Neither §40 (BDSG) is applicable, which regularizes the processing and use of personal data by research institutions, because it is required that the research institution gathers the data from users on its own, nor the investigation of extracted OSN data is included in the category of market and public opinion research purposes, which are separately regulated in §28 (BDSG).

In contrast to work with personal attributes itself, we are merely interested in whether a user has published a single information or not because of the objective to point out attributes that could be used to link OSN profiles. It is not of interest for our investigation to analyze which concrete value of an attribute a user has published. Therefore, we do not need to permanently store a user’s name, his published attributes, and his OSN profile itself. This concrete data is only needed during the automated analysis of profiles. Keeping it transiently in main memory is sufficient for our purpose. Such short-term processing of personal data separates the handling of personal information by OSNs, which has been authorized by the user, and the processing of merely statistical data that do not allow any conclusions regarding personal identities. Such shortest possible time gaps in the sense of separation of two domains of jurisdiction are referred to as a “logical second” [Win00]. The time span during which data is temporarily kept in main memory can only be defined as a logical second if it can be ensured that it is not possible to access the raw data of OSN profiles by a natural person. This applies to console outputs as well as to permanently stored data. As a consequence, working with a software which extracts only statistical data out of OSN profiles is compliant on the basis of German law only if it can be proven that nobody is able to recover personal attributes or profile information with a distinct link to a natural person or an OSN profile at runtime of the analysis software as well as afterwards. Hence, all processing steps in the analysis software have to be executed automatically.
4 Assessment of technical measures for privacy enhancement

As it is emphasized in Section 1, the publication of personal data bears a lot of risks such as the possibility for third-parties to implement software that crawls through OSNs and gathers information by emulation of users’ behavior, respectively the technical behavior of a browser. Such software-based crawlers are often hard to distinguish from human users by OSNs. Before Section 5 presents the results of our statistical analysis on users’ extractable data inside OSNs and the linkability of OSN profiles, we quickly discuss countermeasures of OSNs that are implemented to avoid automated extractions of personal data.

It is obvious, the better a crawler emulates the behavior of a human being and a browser, the more difficult it is to detect such intruding software. Hence, crawlers use previously registered accounts and determine parameters such as initialization vectors and valid session identifiers to login into OSNs. Additionally, to stay logged in crawlers have implemented a handling for cookies. The data will be extracted via parsing of HTML pages and identifying the information by means of tags or keywords inside the HTML code. Such keywords are commonly static and self-explanatory.

OSNs implemented several countermeasures to thwart and prevent crawling. If an alleged browser tries to get responses more and faster than ordinary users would try, CAPTCHAs are presented before the requested content is replied. CAPTCHAs are pictures with distorted letters that a user has to type in to be able to carry on surfing. CAPTCHAs are a common possibility to disturb crawling attacks, but an OSN has to weigh the safeness of its users and the potentially decreasing usability with respect to the amount of occurrences of CAPTCHAs. Into the bargain, today’s CAPTCHA-challenges are breakable by software, too [ZYL+10]. Apart from that, a fix delay between sent HTTP-requests is often sufficient sophistication to prevent the occurrence of CAPTCHAs. Since only a valid e-mail address and some not further validated personal information is needed to create an OSN account, any time a profile is yet blocked by a CAPTCHA the software is able to register a new account and re-login into the OSN. This is also effective if OSNs are blocking further browsing for 24 hours after a fix amount of HTTP-gets within a fix amount of time. However, as long as these measures can be circumvented by re-login with another account, it is only a low extra effort for attackers to implement a bypass inside their crawlers.

Beyond these countermeasures, switches between standard websites (www...) and sites for mobile devices support crawlers to avoid Java Script (and its variations like AJAX) that are more difficult to emulate. Usually, the mobile pages that any common OSN provide are solely based on native HTML code and if the software is logged in into one of these alternatives the access to the other one is also open. It is obvious that such and other countermeasures are not sufficient to completely avoid automated data extraction.

5 Profile correlations based on friends lists

In this section we present a statistical evaluation on numerous profile pages of two of the most popular social networks in Germany and also worldwide. We refer to these OSNs
as OSN 1 and OSN 2 to not motivate attackers to implement crawlers in the future. We analyzed about 15,000 profiles of OSN 1 and about 50,000 profiles of OSN 2 to identify attributes that are suitable to be used to link OSN profiles of a single person from different OSNs (for instance by third-parties or to show a user how linkable his OSN profiles are). This analysis is done on the basis of the requirements stated in Section 3 and ensures, therefore, compliance to the German law as well as do not jeopardize users’ privacy. Before this section presents the results of the study, the analysis software as well as its measures to ensure a lawful investigation are highlighted in the following.

The analysis software searches for a string (first name and last name) in different OSNs, gets lists of profiles, analyzes the publicly available information and compares each detected profile of one OSN with each detected one of the other OSNs. Afterwards, statistical data is extracted and stored permanently. Since the specific user name the software is searching for is not decisive for the results and in light of the presented restrictions concerning the fact that nobody should have access to a searched name, these input parameters have to be automatically generated. We implemented a fully automated generation of search parameters which in turn randomly picks and concatenates first names and last names out of large lists of popular names. A further characteristic of the software is that these input parameters as well as all extracted profile data are discarded subsequent to the analysis such that the analyzed profiles are unidentifiable.

To summarize, our analysis software autonomously generates input parameters for search queries, extracts publicly available information out of OSN profiles that are found with the searching term in different OSNs, calculates statistical data, tries to link profiles and discards all information that could identify the analyzed profiles. For sure, the requirement that the network will not be disturbed in terms of network load regarding numerous HTTP-requests by the developed software has additionally been taken into consideration.

In OSN 1 we figured out that 62.85% of profiles contain a publicly available birth date, whereas in OSN 2 only 1.24% of all analyzed profiles provide this information. In general, a users name and his birth date are applicable to identify a person out of a large circle of persons. Otherwise, it cannot be guaranteed that solely these two attributes appear uniquely in a group of people. Moreover, the probability that a combination of name and birth date is unique inside a large social network is obviously low and not sufficient to identify people. However, if it is an aim to link profiles of a specific person, in many cases these attributes provide enough information to establish such a link, depending on the amount of people who use the same name inside the networks.

In analogy to the birth date-attribute, some other attributes can be used to identify a user in an OSN. For instance, the name of the attended college or university is available in 53.88% of all analyzed profiles of OSN 1, in OSN 2 this information can be seen on 11.75% of profiles. A relationship status is provided by 25.56% (OSN 1), respectively 17.67% (OSN2) users. The hometown is available in 23.79% (OSN 1) or 13.36% (OSN 2) of the profiles. Less suitable to link OSN profiles are attributes like the gender that is publicly available at 69.64% (OSN 1) or 71.31% (OSN 2) of the profiles. Since in the majority of cases the name implies the gender, this information has no additional value for the aim of user identification. However, 46.13% of profiles at OSN 1 provide a fully accessible friends
list and even 58.76% at OSN 2. In the following, we discuss the thesis that friends lists provides sufficient information to link OSN profiles.

This thesis is based on the assumption that a single combination of names that a friends list contain is to the greatest possible extent unique inside the network and, therefore, like a fingerprint of a single user. Furthermore, if such a single user is member of different OSNs, we assume that his friends lists overlap. Of course, if OSNs serve a particular purpose, i.e. for instance, business networking, the overlap of friends lists of OSNs that are dedicated for connecting to friends and family members are accordingly less. However, if this overlap is significantly higher than what we can expect for two OSN profiles with the same name but from different people, friends lists are a suitable type of information that can be easily (mis-)used to link profiles.

![Figure 1: Histogram that shows the probability of friends lists overlap of two profiles from different OSNs with respect to the extent of the overlap and a corresponding CDF](image)

Figure 1 shows the overlap of friends lists of profiles from two different OSNs. Therefore, we did more than 350,000 comparisons of friends lists, whereby in each comparison both analyzed profiles are registered under the same user name in the respective OSNs. The diagram shows the comparisons that result in overlaps between 0.1% and 100%. The x-axis shows the proportion of friends lists of OSN 1 that overlap with compared friends lists of OSN 2 and is divided into bins of 5%. On the y-axis the relative amount of comparisons within the different bins are drawn. The grey curve constitutes the corresponding cumulative distribution function with its y-axis on the right. It can be seen that the slope of the curve increases at about 7% overlap and has a local maximum at about 17% overlap.

The changed course at 7% overlap can be interpreted as a threshold which indicates that an identified overlap less than 7% could be a comparison of profiles of the same person or not. In general, the higher the overlap the higher the probability that the same person is owner of two compared profiles. Especially, this probability increases significantly if the comparison results in an overlap of more than 7%. With this interpretation, the local maximum constitutes an average overlap of friends lists of two profiles of a single person. However, the graph shows that a comparison of two friends lists of profiles of different OSNs results in many cases in no or just a minimal overlap. But, in some cases a higher overlap is obviously detectable and the results suggest that the overlap indicates a single person related to both compared profiles. Because of the requirements regarding compliance, it is not possible to validate such interpretation of the graphs. However, Table 1
Table 1: Examples of overlaps of single friends lists with friends lists from a different OSN

<table>
<thead>
<tr>
<th>Description</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of profiles in OSN 2 with identical name</td>
<td>111</td>
<td>193</td>
<td>97</td>
</tr>
<tr>
<td>Number of these profiles in OSN 2 with a friends list that overlap less than 2%</td>
<td>110</td>
<td>192</td>
<td>96</td>
</tr>
<tr>
<td>Number of these profiles in OSN 2 with a friends list that overlap more than 20%</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

shows exemplarily the results of comparisons of three single profiles of OSN 1 with all profiles of OSN 2 wherein users registered with an identical name. It can be seen that in these examples exist a gap between overlaps less than 2% and overlaps higher than 20% which in turn underpins the interpretation of Figure 1.

6 Conclusions and future work

In OSNs users are faced with a thin line between publishing data to be part of the OSN as well as getting social recognitions such as it is shown in Section 1 and otherwise the potential of danger regarding privacy. In fact, third-parties are able to extract users’ publicly available personal data without a permission of users and they are able to link information from different OSNs. In contrast to mandatory access control ensured by controlled and clear processes, e.g. as they are implemented in an identity management system, unintended data flows from OSNs to third-parties cannot be interrupted by users nor the OSN providers apart from adjusting privacy settings. The objective of the presented study is to demonstrate the linkability of OSN profiles if personal information is not hidden from strangers and therefore to show the possibilities of gathering data by third-parties.

In this paper, we firstly introduced a concept for statistical sampling on the basis of an analysis of German Data Protection Act that ensures compliance. Furthermore, we showed that the countermeasures OSNs have implemented to avoid unwanted data extraction by third-parties are not sufficient. Neither the integration of CAPTCHAs in page flows nor the blocking of further surfing after a fix amount of HTTP get-requests in a fix amount of time are effective to suppressing extractions entirely. In the following, we demonstrated that several attributes can be used to link information from different OSNs that belong to a single user. Moreover, we reinforced the thesis that friends lists, which are publicly accessible at about a half of all OSN profiles, can serve as fingerprints in OSNs and provide an opportunity to associate information. We substantiated the observation that an overlap of two compared friends lists of more than 7% is an indicator that both friends lists are part of profiles of a single natural person.

For future work, the presented study can be applied to constitute a basis for an application that shows users their concrete virtual appearance in a comprehensive manner. Particularly, the gained knowledge can be used to demonstrate a user the linkability of his OSN profiles and, thus, inadvertently exposed personal information can be unfolded. Such feedback
consequently results in an increased transparency that might motivate more users to adjust privacy settings in the future. An open issue to design such an application belongs to the question whether comparisons of a user’s OSN profile with other profiles (matching the name of the profile holder) to determine overlaps are compliant to data protection acts because a processing of personal information of third OSN members cannot be avoided.

Acknowledgement
We like to thank Dr. iur. Oliver Raabe (KIT, ZAR) for his help in the analysis of German Data Protection Act and our students Irina Taranu, Florian Werling, and Michael Biebl for their contributions and help in the implementation of our analysis software.

References


