

What Your Friends Tell Others About You: Low Cost Linkability of Social Network Profiles

Sebastian Labitzke Irina Taranu Hannes Hartenstein

Steinbuch Centre for Computing & Institute of Telematics

Karlsruhe Institute of Technology (KIT), Germany

sebastian.labitzke@kit.edu, irina.taranu@student.kit.edu, hannes.hartenstein@kit.edu

ABSTRACT

Due to the amount of personally identifiable information shared by users of online social networks (OSNs) and the often not adequately adjusted privacy settings, it is possible to identify a user's several OSN profiles. In this paper, we illustrate that third parties just have to take a small step to link OSN profiles of a user and, consequently, to aggregate various pieces of information shared in several OSNs. Particularly, based on statistical results we illustrate how the often publicly available friends lists can be exploited to link several OSN profiles of a single natural person. The results presented in this paper show that profiles can be linked even at low cost, i.e., without complex correlation techniques or high computational power. To assess the risk of privacy leakage by profile linking, we, additionally, report how often specific pieces of information are made publicly available by users in four of the most popular OSNs. We show that users tend to publish different pieces of information in different OSNs and, thereby, demonstrate that by linking friends lists more information about a user can be gained than the user shared in a single OSN. For the study we analyzed more than 180,000 user profiles and compared more than seven million pairs of profiles to investigate profile linkability.

Keywords

Social Network Analysis, Linkability, Friends Lists

1. INTRODUCTION

Today, online social networks (OSNs) are commonly used to keep in touch with people from all areas of life. Sharing personal data, i.e., personally identifiable information (PII) seems to be so attractive for many OSN members that the recognition of risk regarding privacy often fades into the background. Facebook claims that, on average, every single one of their more than 600 million users provides more than 90 pieces of content per month¹. The behavior of OSN users

¹<https://www.facebook.com/press/info.php?statistics>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

The 5th SNA-KDD Workshop '11 (SNA-KDD'11), August 21, 2011, San Diego CA USA

Copyright 2011 ACM 978-1-4503-0225-8 ...\$5.00.

and their generosity regarding the amount of shared data has been, for instance, investigated in [5], [14], and [11]. If nowadays third parties know just a few details about a person, the probability to find this person inside one of the mass of OSNs is already high [16] and, thus, more information about this person is accessible. Possibilities to restrict access to published information in terms of privacy settings provided by OSNs are only used by a subset of all OSN members. In most OSNs, just a minority of members hide all of their data from members beyond their friends lists. In Section 3, we illustrate this fact on the basis of analyzed statistical data gathered by 180,000 investigated OSN profiles.

Additionally, publicly available user information is made available by OSN websites that can easily be parsed by software because of their standardized layouts [4]. Consequently, this constitutes an invitation for third parties to crawl OSNs in order to gather as much information about people as possible despite the fact that this might be illegal. If it is then possible to gather information from several OSNs by profile linking and to associate such information with a natural person, the resulting cumulative data set is of a greater value than unlinked and individual pieces of information. Hence, linking of OSN profiles might be profitable for third parties if users provide some information publicly in one OSN and some other in another OSN.

To turn the tables, linking of profiles could also be used to reveal and demonstrate users their virtual appearance across social networks. Bold and simple revealing of their linkability to users might help to motivate a more careful and adequate adjustment of privacy settings. The psychological point of view regarding such opportunity of motivation is investigated in our preceding study [13]. In this study, we also addressed restrictions for OSN analysis with respect to the German data protection act and countermeasures that are implemented by OSNs to disturb crawlers in gathering data via parsing of OSN profiles. However, the focus of the current work is to demonstrate users, who are not aware of the linkability of their OSN profiles, implications for privacy with respect to possible aggregations of information published in several OSNs. Such implications occur if those users do not entirely hide their shared data from strangers. In contrast, if users linked their profiles on their own by use of services such as <https://about.me/>, the linkability is implicitly given and intended by users.

In this paper, we start by demonstrating the average availability of every type of information published in user profiles of four popular OSNs. The objective of this part of the study is to assess the potential risk regarding profile linking,

or rather, to assess the additional information that can be gained about a user by linking his OSN profiles. Furthermore, we present a “low cost” concept to link OSN profiles of a single natural person. Low cost means that we focus on profile linking on the basis of simple string comparisons of provided data rather than on the use of complex correlation algorithms, such as, for instance, face recognition techniques [17]. In this manner, the analysis of the statistical data shows that especially the often publicly available list of friends can be easily exploited to effectively link OSN profiles of a single natural person. We further constitute that just a low overlap of two friends lists is a sufficient indication to identify a single user as the owner of two corresponding OSN profiles. The underlying linkability investigation is based on seven million comparisons of pairs of profiles. In conclusion, the contributions of this paper are:

- We show a current snapshot (based on an extensive set of measurements) of the availability of information in profiles of four popular OSNs.
- We analyze the diversity of information users share via OSNs with respect to the OSN itself.
- We demonstrate that information can be gathered efficiently by profile linking using only a manageable number of string comparisons on the basis of friends lists.

The paper is structured as follows. Section 2 analyses the related work and point out the unique characteristics of the presented study. Section 3 demonstrates the results according to the investigated availability of specific information inside OSN profiles and formulates associated hypothesis. Furthermore, the derived research questions are highlighted. Section 4 comprises methods, definitions, and fundamentals with respect to the analysis of statistical data regarding the profile linkability. Results of the linkability investigation are demonstrated in Section 5. Furthermore, we discuss the aforementioned results on a meta-level in Section 6, before Section 7 concludes the paper. In the Appendix A, we note essential information on the compliant concept regarding data sampling.

2. RELATED WORK

Publicly available PII and the linkability of OSN profiles have been investigated by other previous studies. In the following, we show the relation and differentiation of these studies and our approach.

As emphasized in the introduction, users reveal a lot of PII via OSN profiles. The “amount of information shared on a user’s profile as well as in the process of communication with others” is called self-disclosure [8]. In the past few years, several studies were performed to describe which pieces of PII users reveal (especially via Facebook) and which preventive measures they take to hide their data from third parties. Since the results of such studies changed over time, the evolution of user behavior in Facebook can be derived with respect to the results of related studies and the current status on the availability of users’ information presented in this paper. Gross and Acquisti examined the self-disclosure of OSN users based on empirical data of more than 4,000 Facebook profiles of Carnegie Mellon University students in 2005 [5]. At this point in time, only 1.2% of the users prevented others outside the network from finding their profiles

and only 0.06% entirely restricted the visibility of their profile. In 2007, Lampe et al. collected about 38,000 Facebook profiles [14]. 19% were set up with access restrictions and 59% of the profile fields contained information. A further study performed by Brown et al. investigated 7,919 profiles from the Facebook network of the Michigan University in 2008 [3]. They found that 68% of users let their profiles visible for everyone from the network. Of these, 86% had a publicly available friends list. In contrast to such related studies, we present a current snapshot of the situation of self-disclosure in Facebook and, additionally, three other popular OSNs based on a more extensive set of measurements. Furthermore, we compare the availability of information within Facebook to these other OSNs. According to this data, it can be seen that the awareness of users regarding the adjustment of privacy settings has grown in the past. Nevertheless, we show that a majority of users is still generous with respect to publicly available PII.

To stay in control of shared data, users have to adjust their privacy settings. In 2008, Krishnamurthy et al. investigated privacy settings of OSNs and showed that most users do not change the default configuration [10]. Only 1% of Twitter users, 25% of Facebook users and 21% of MySpace users changed their privacy settings. Studies such as [11] and [12] also examined the availability of information about users in different OSNs, trying to find out what information OSNs contain and how much information users can protect from strangers or attackers. The authors showed that not every piece of information could be made unavailable for others. In [11] is demonstrated that the friends lists can be accessed in ten of the twelve analyzed OSNs if the users maintain the default privacy settings. We demonstrate in this paper that nowadays in three of four analyzed OSNs still only 7% to 22% of users hide their profile completely from strangers, except information that can not be hidden.

If users do not adjust, or rather, are not able to adjust their privacy settings appropriately, one of the consequences is the linkability of OSN profiles. The authors of [20] demonstrated a strategy to link profiles by revealing user’s several names within various web communities, such as OSNs. In 2005 researchers estimated that there is an overlap of profiles in two OSNs of 15% [15]. A study conducted by compete.com in 2007 showed a more major member overlap in OSNs². There was an overlap of Facebook profiles with profiles of MySpace of 64% and, vice versa, only 20%. Besides profile linking, analysis on and comparisons of OSN profiles can be abused for further attacks such as identity clone attacks, as was investigated in [2] or the re-identification of OSN profiles if corresponding users visit a website [19].

However, Motoyama et al. developed a system for matching profiles in OSNs [16]. The goal was to figure out sufficient criteria to find a person in an OSN, so that the search is as effective as possible and the correct profile is found. This has allowed the authors to match user profiles of Facebook with user profiles in MySpace and vice versa. Thereby, they showed that 43% of users with overlapping profiles in both OSNs have similar privacy settings, the rest take different security measures. Most likely MySpace profiles are private while Facebook profiles remain open. This shows that for at least 57% of users new insights can be gained by finding corresponding profiles. This underpins the hypothesis

²<http://blog.compete.com/2007/11/12/connecting-the-social-graph-member-overlap-at-opensocial-and-facebook>

that profile linking is profitable for third parties in light of gathering as much information about a user as possible.

Additionally, Motoyama et al. show only a little overlap of friends lists [16]. We demonstrate in this paper that such merely low overlaps are sufficient to link OSN profiles of a single natural person. Therefore, we disprove the hypothesis that the rare overlap of friends lists can not be (ab-)used to link OSN profiles of a natural person to gain more information about this person.

3. OBSERVATIONS AND RESEARCH QUESTIONS

At first, we assess which and how often pieces of information are shared by users within four popular OSNs. In doing so, we focus on information whose access is not restricted by privacy settings. Table 1 lists attributes published within either of the four evaluated OSNs. For each attribute, we state the ratio of users who did not restrict the attribute's visibility, i.e., attributes can be seen by any logged in user. Hence, the results reveal which information is extractable out of OSN profiles in any case.

The presented statistical data was collected between January and March 2011. The data was gathered by use of a software that we specifically implemented for this purpose. This software is designed to comply to the German data protection act. In the appendix, we note information on the underlying compliant concept of data sampling. By use of this software, we analyzed more than 110.000 Facebook profiles, more than 43.000 profiles of StudiVZ, more than 25.000 MySpace profiles and more than 10.000 profiles of Xing.

The first column of Table 1 classifies the types of information a user can publish via an OSN profile. Information which is shared by using chats or walls also referred to as pinboards have not been considered in this study. The category *general* subsumes the availability of names of users and their friends lists which form the basis of the linkability part of the study presented in Section 5. Moreover, this category indicates how many profiles are completely hidden from strangers by adjusted privacy settings. The category *number of...* illustrates how many profiles we found in a single search request on average and at the maximum. This category also states the number of friends a user has on average, including the corresponding standard deviations, and the size of the largest detected friends lists. Whereas the category *personal* summarizes almost never changing attributes like date of birth or gender, the category *contact* lists all attributes that can be used to locate a user physically. The categories *job* and *higher education* give insights into the occupations of users. Users provide information about their family and relationship background, current frame of mind and physical shape via attributes of the category *oneself and relations*, whereas the attributes of the category *views and attitudes* contain the personal views regarding politics, religion, etc. The category *hobbies* represents attributes to tell others something about further activities whereas *favorites* summarizes favored music, literature, and so on.

Remarkably, the statistical values of Xing differ from others (see Table 1). Some attributes are mandatory to create a Xing profile, such as the type of job, the job description, the company, the area of business, the country and the city. These attributes are available to every Xing user and the access can not be restricted by adjusting privacy settings.

In contrast, any contact information can only be seen by direct friends, so that we mention the availability of these attributes at 0%. Currently, further pieces of information are accessible by any Xing user if the user simply published this piece of information.

However, one of the most remarkable observations within the presented statistical data is the diversity of availability of information with respect to the specific OSN. For instance, the date of birth is made available by 64% of all evaluated StudiVZ users. In contrast, only 0.84% of Facebook users share this particular information publicly. A similar difference regarding the availability applies to the university, hometown, current residence, cv, relationship status, and more. In light of the fact that many users participate at more than one OSN, such diversity of the availability of information results in the following worthwhile opportunity for third parties: if it is possible to link several OSN profiles of a single person, a comprehensive set of information can be gathered and correlated by third parties. Therefore, the hypotheses are as follows:

- The different behavior of users regarding specific information shared in different OSNs constitutes a privacy threat for users if profiles are linkable.
- The lower the costs (e.g. computational power, and implementation effort) required for linking a profile are, the broader is the group of third parties that are able to link profiles.

We showed that linking of profiles might be worthwhile for third parties because users share different information in different OSNs. In light of the aforementioned hypotheses, the following research questions are derived:

- Are typically shared attributes sufficient to successfully link profiles?
- Are friends lists, as the most available information despite the specific OSN, sufficient to link profiles?
- Can profiles be linked at low costs, so that a high number of third parties are able to link profiles?

In general, any type of information of OSN profiles might be sufficient to link profiles if these attributes can be compared among each other, such as comparing profile images via face recognition, status messages via text mining, shared PII (see Section 6.1), and so on. However, we are looking for a low cost strategy to link OSN profiles, i.e., we try to establish links between OSN profiles with less computational power and less complex correlation techniques than, for instance, required for face recognitions as well as without complex mining algorithms. Friends lists are publicly available in many OSN profiles (between 40% and 67% of all analyzed profiles) and comparisons of two friends lists can be done via simple string comparisons. Therefore, we focus on the investigation of the friends list based linkability and show that string comparisons are sufficient for profile linking.

4. METHODOLOGY AND DEFINITIONS

In this section, we describe the methodology for the investigation of friends list based profile linking. Furthermore, we define terms and metrics necessary to analyze and interpret the statistical data gathered by comparing seven million pairs of profiles during our study.

Category	Attribute	StudiVZ	Facebook	MySpace	Xing
General	name	100.00%	100.00%	100.00%	100.00%
	<i>no further information available</i>	6.97%	21.53%	20.34%	n.a.
	friends lists	48.13%	59.45%	67.04%	40.98%
Number of...	...profiles per search request (avg.)	11	20	12	6
	...profiles per search request (max.)	288	545	491	1554
	...friends (avg.)	67	141	21	44
	...friends (std. deviation)	74.01	216.32	75.47	85.39
	...friends (max.)	918	5499	2058	2878
Personal	graduation/title	5.51%	- ¹	- ¹	69.03%
	date of birth (dob)/age	64.00% (dob)	0.84% (dob)	32.06% (age)	0%
	zodiac sign	10.51%	- ¹	20.72%	- ¹
	gender	71.06%	49.92%	32.06%	- ¹
Contact	hometown	23.74%	8.77%	6.49%	- ¹
	current residence/region	48.46%	10.32%	32.04%	mandatory (100%)
	homeland or current country	18.69%	n.a.	28.78%	mandatory (100%)
	address	0%	0.11%	- ¹	0%
	email	- ¹	0.62%	- ¹	0%
	mobile phone number	- ¹	1.19%	- ¹	0%
Job	company/occupation	5.75%	9.63%	4.91%	mandatory (100%)
	type of job	7.85%	n.a.	n.a.	mandatory (100%)
	income	- ¹	- ¹	2.21%	- ¹
Higher education	university	51.67%	8.83%	n.a.	n.a.
	field of study or study path	11.21%	n.a.	n.a.	n.a.
	languages	10.04%	- ¹	- ¹	n.a.
	general education/cv	23.71%	2.81%	6.35%	n.a.
	current school	- ¹	16.00%	n.a.	- ¹
Oneself and relations	about myself	16.33%	n.a.	6.22%	- ¹
	relationship status	26.51%	13.12%	n.a.	- ¹
	status message	49.85%	n.a.	20.72%	- ¹
	physique	- ¹	- ¹	6.93%	- ¹
	parentage	- ¹	n.a.	6.02%	- ¹
	children	- ¹	- ¹	7.61%	- ¹
Views and attitudes	sexual orientation	- ¹	8.16%	7.77%	- ¹
	interests/looking for...	31.41%	8.89%	8.92%	- ¹
	political direction	18.42%	0.33%	- ¹	- ¹
	religious views	- ¹	0.46%	4.96%	- ¹
	smoking and imbibing	- ¹	- ¹	5.52%	- ¹
Hobbies	interests/details	25.44%	8.50%	20.72%	n.a.
	clubs/activities/groups	16.57%	14.29%	0.77%	n.a.
Favorites	favorite citation	21.00%	4.78%	- ¹	- ¹
	favorite music	26.02%	16.75%	6.84%	- ¹
	favorite books	19.32%	5.88%	5.20%	- ¹
	favorite movies	22.04%	10.47%	5.90%	- ¹
	favorite TV shows	- ¹	12.73%	5.51%	- ¹

¹ This attribute does not appear in such OSN profile by default

Table 1: Publicly available personally identifiable information in OSNs (unrestricted access) [%]

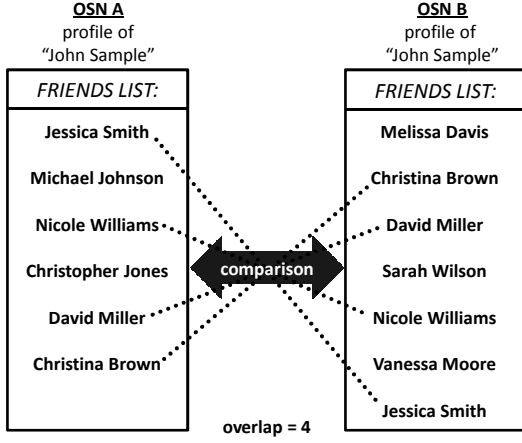


Figure 1: Exemplary comparison of two friends lists

4.1 Overlap metric

By comparing profiles from different OSNs, the objective is to identify users who own a profile in two or more OSNs. In order to compare two given profiles found in different OSNs, or rather, their friends lists, we proceed as follows. First, our analysis software considers each friends list as a set of entries, where each entry is given by the name of a single friend. Then, the intersection of these two sets S_1 and S_2 is determined ($S_1 \cap S_2$). We refer to the number of entries within this intersection as *overlap*, i.e., the overlap comprises the number of names that appeared in both friends lists ($|S_1 \cap S_2|$). A more sophisticated determination of the overlap by calculating, for instance, the Jaccard index [6] is not necessary because the whole number of friends of compared profiles is insignificant for the results presented in Section 5. We refer to the defined “simple” determination of the overlap as a single *comparison* (compare Figure 1). Later on, the overlap is used as one of two correlation metrics that have to be taken into account to decide whether two compared OSN profiles belong to the same natural person.

In the following, we refer to profiles that are identified as profiles of a single user as a *match*. OSNs usually provide a feature to search user profiles and other content of the network by generic search strings. In order to detect a profile that belongs to a user who also owns a specific profile p of OSN A in OSN B, we compare profile p to every possibly matching profile of OSN B. In this context, the characteristic “possibly matching” means any profile of OSN B that can be found with the same search string used for finding p in OSN A. In the appendix, we give more information regarding the selection of search strings.

Since we did a multitude of single comparisons of friends list pairs, we structured these comparisons as follows. A *single* profile which is found by using a specific search string s is compared to every profile that can be found inside another OSN with the same search string. If n constitutes the number of profiles found in OSN B with s , we compare n profiles with profile p (1:n). We refer to such a set of corresponding comparisons as a *comparison set* (cs). Figure 2 illustrates the concept of comparison sets. In Figure 2(a), a single cs is shown. Figure 2(b) depicts several comparison sets assembled with the same search string. Therein, an arrow represents a single comparison of two friends lists and

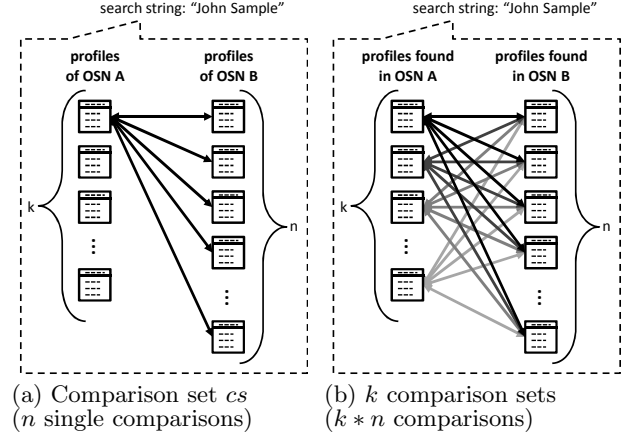


Figure 2: Illustration of the comparison set concept

arrows of the same grey-level illustrate a comparison set.

For example, suppose we search for a person named “John Sample” in two OSNs, OSN A and OSN B. Assume that the search returns 200 profiles in OSN A and 100 profiles in OSN B with the name “John Sample”. To identify the same user of a specific profile p of OSN A within the 100 profiles of OSN B, we firstly compare the friends list of the profile p with all 100 friends lists of users named “John Sample” from OSN B. These 100 comparisons form a comparison set (cs) consisting of 100 single comparisons (1:100, compare Figure 2(a)). To check each of the 200 profiles found in OSN A, the comparison procedure has to be executed for each of the 200 found profiles, resulting in 200 cs of 100 comparisons each (compare Figure 2(b)).

4.2 Maximum overlap and target comparison

Assuming that comparisons of friends lists are sufficient to identify profiles of a single natural person in several OSNs, the comparison of two friends lists that constitutes the largest overlap inside a cs has the highest likelihood to indicate a “match”. In general, resulting overlaps of a single cs , i.e., overlaps that are detected due to n comparisons between a single profile p and each profile found by the same search string within another OSN, are defined by the function:

$$o(cs), cs \in \{comparison_i | 1 \leq i \leq n\}$$

Therefore, we define the largest detected overlap of a cs as $max(o(cs))$, i.e., this function provides the maximum number of equal names detected in a specific comparison of a single cs . A $max(o(cs))$ indicates the overlap of the comparison with the largest overlap inside a cs .

Generally, even more than one comparison of a single cs can result in the same determined overlap. If the maximum overlap is only detected in a single comparison within the cs , we, additionally, denote the corresponding comparison as a *target comparison*. Target comparisons stand out from the others because no other comparison within the corresponding cs resulted in such a high overlap. The function $f(max(o(cs)))$ represents the number of comparisons (of a single cs) that resulted in the maximum overlap.

Figure 3 shows a histogram of an exemplary cs . Therein, the detected overlaps of an exemplary comparison set with twelve single comparisons (1:12) are shown, i.e., twelve profiles of OSN B (found with a search string s) were compared

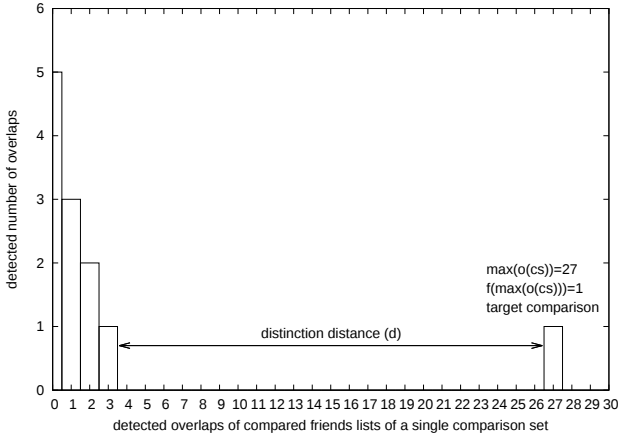


Figure 3: Histogram of detected comparison overlaps for an exemplary (1:12) comparison set cs

to a single profile of OSN A (found also with s). The maximum detected overlap of compared friends lists is at 27 ($\max(o(cs)) = 27$), which is a target comparison because it is only detected in a single of the corresponding twelve comparisons ($f(\max(o(cs))) = 1$).

4.3 Distinction distance metric

The gap between the overlap value of a target comparison and the next lower detected overlap inside a cs is called the distinction distance d and quantifies how distinct the target comparison stands from the other comparisons (see Figure 3). Therefore, the distinction distance quantifies the discriminative power of a maximum overlap with respect to other comparisons of the corresponding comparison set.

Distinction distances constitute the second correlation metric to decide whether two compared profiles belong to the same natural person. The assumption is: the higher the distinction distance the higher the probability that a match is detected. The distinction distance d of the overlap of the exemplary target comparison is at 24, calculated by 27 minus three (the next lower detected overlap), i.e., by means of d the target comparison distinguishes significantly from other comparisons inside this exemplary cs . Particularly, all other comparisons resulted in an overlap that is significantly lower (here: a maximum of 3 equal names).

4.4 Analysis of statistical data

As motivated before, we concentrate on comparisons that result in a maximal overlap $\max(o(cs))$ of friends inside a cs to analyze the linkability of profiles on the basis of friends lists. Thereby, we try to figure out whether these maximum overlaps are overlaps of target comparisons, i.e., whether no other comparison of the same cs result in the same maximum overlap ($f(\max(o(cs))) = 1$). Moreover, we investigate the distinction distance of such target comparisons to quantify the significance of the assumption that a target comparison indicates two profiles of the same natural person.

4.4.1 Aggregated graphs

To give an overview of the values and occurrences of detected maximum overlaps, we show graphs for which we consider only the values of $\max(o(cs))$ of every cs . Thereby, all $\max(o(cs))$ that correspond to comparisons of profiles of a

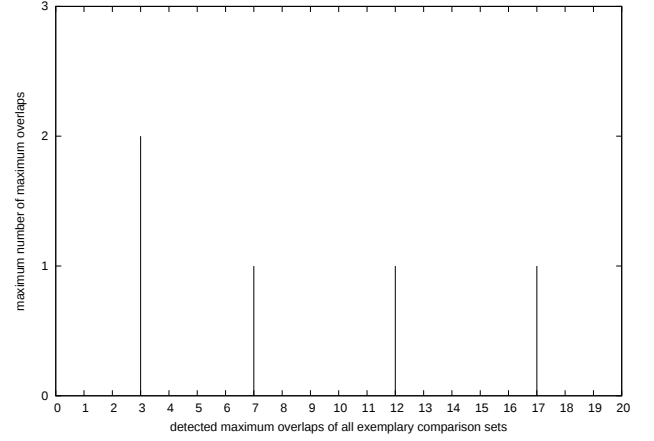


Figure 4: Exemplary aggregated max-overlaps graph (maximum number of maximum overlaps)

specific OSN A and profiles of a specific OSN B are aggregated in a single graph. The y-value of such a graph indicates the *maximal* number of comparisons that resulted in a $\max(o(cs))$ within every cs . For instance, if this y-value equals one, in every cs the corresponding maximum overlap is not detected more often than once. In turn, the corresponding comparisons are *target comparison*. If the y-value is greater than one, the maximum overlap is at least found for two comparisons of a single cs and, therefore, this maximum overlap does not correspond to a target comparison.

An exemplary plot of such a graph is shown in Figure 4. For this diagram assume five comparison sets cs_i , $i=\{1, 2, 3, 4, 5\}$ for which the maximum overlaps are as follows: $\max(o(cs_1)) = 3$, $\max(o(cs_2)) = 7$, $\max(o(cs_3)) = 12$, $\max(o(cs_4)) = 12$ as well, and $\max(o(cs_5)) = 17$. Assume that the maximum overlap of three is detected in two single comparisons in the first comparison set cs_1 . The corresponding aggregated graph shows four peaks, as seen in Figure 4. Three of these peaks indicate a maximum overlap with a maximum number of one because in each of the corresponding four cs these values are not detected or just detected in one single comparison. Note that the peaks indicate the *maximum* number of a maximum overlap and do not indicate the number of cs with a specific determined maximum overlap. In this example, only one comparison within each of two corresponding comparison sets existed whose overlap matched the maximum overlap of 12. In contrast, the peak at three has a y-value of two because we assumed that this maximum overlap is found two times in cs_1 so that the maximum number of such specific maximum overlap three is two. We refer to such graphs as *aggregated max-overlaps graphs*, i.e., graphs in which the maximum overlap of every cs is plotted against its maximum number within every cs .

4.4.2 Illustration of distinction distances

As mentioned above, the expressiveness of the maximum overlap can be estimated by means of its distinction distance d , i.e., the higher the value of d , the more the maximum overlap stands apart from all other detected overlaps of a cs . In turn, a high value of d indicates a high probability that two compared profiles belong to a single natural person because of the associated outlier characteristic of the

maximum overlap. In Section 5, we, therefore, plot the average distinction distances d for every detected maximum overlap and the corresponding standard deviations into the aggregated max-overlaps graphs.

If the maximum overlap stands considerably apart from all other detected overlaps in most analyzed cs , the average value of d converges to the value of the overlap. An average d of the same value as a corresponding overlap means that the next lower detected overlap in every cs is zero because the distinction distance is at its maximum. Distances which converge to the maximum show that other overlaps are much lower and can be neglected. In contrast, if the average d is much lower than the corresponding overlap or its standard deviation is unusually high, other overlaps were detected that are not much lower than the maximum overlap. If that would be the case, the metrics *maximum overlap* and *distinction distance* are not sufficiently significant to reach conclusions about whether two compared friends lists belong to OSN profiles owned by the same natural person.

5. RESULTS

In this section, we analyze gathered statistical data and demonstrate friends list driven profile linkability for four popular OSNs. The objective is to compare a friends list from a specific user profile of OSN A with the friends lists of profiles from another OSN B. Thereby, we try to find a profile from OSN B which belongs to the same user who owns the profile of OSN A. Firstly, we assume that the user is registered with the same name in both OSNs. A discussion whether the results are transferable to profiles that are not registered with the same name follows in Section 6.2.

5.1 Friends list driven profile linking

In the following, we show aggregated max-overlaps graphs including the illustration of average distinction distances, such as it is introduced in Section 4.4. We compared friends lists of each of four analyzed OSNs with friends lists taken from the other three OSNs and vice versa. Thus, we could present twelve plots in which the maximum overlaps and their average distinction distances are shown. In the following we depict four of these graphs because they already show the gained scientific knowledge. Figure 5(a) shows comparisons between friends lists of profiles of Facebook and StudiVZ, Figure 5(b) visualizes the results of comparisons between StudiVZ and XING, whereas Figure 5(c) and Figure 5(d) present the results of comparisons between Facebook and Xing or MySpace, respectively.

The objective of Figure 5 is to illustrate a phenomenon we observed: beginning at a certain maximum overlap the corresponding maximum number of every larger maximum overlap is one. Rather, low maximum overlaps around one up to three indicate a higher maximum number. Therefore, maximum overlaps higher than three are invariably associated to target comparisons, i.e., every maximum overlap higher than three is detected in none or just a single comparison within every comparison set.

The average distinction distances of maximum overlaps (also shown in Figure 5) strengthen this observation. Considering these distances, two distinctive phenomena are obvious. Firstly, every average distinction distance d is just a little bit lower than its corresponding maximum overlap. Secondly, it can be noticed that the standard deviations of d are so small that they can hardly be recognized.

Therefore, the interpretation of these effects is as follows: an overlap that is a maximum in a cs always stands considerably apart from all other detected overlaps. That means that an associated target comparison of a maximum overlap is in all cases isolated from the rest of detected overlaps. In particular, we determined that maximum overlaps are isolated from other overlaps detected in comparisons of a cs if these maximum overlaps are higher than three. In short, the graphs show that each comparison set almost exclusively resulted in very small overlaps of friends lists (see average distinction distance), except a single higher overlap which can be detected in many comparison sets and which stands considerably apart from the others.

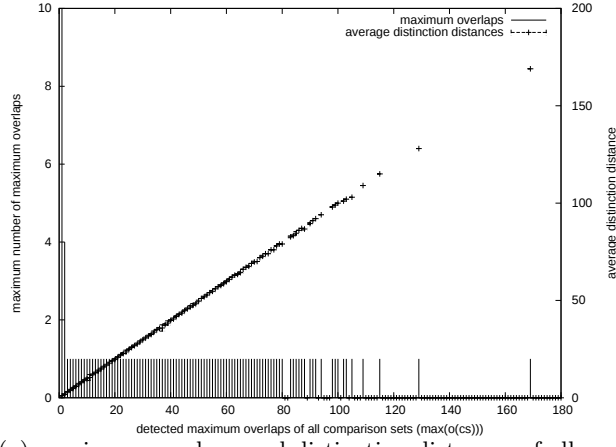
Therefore, we assume that OSN profiles of the same natural person are identified if the comparison of two friends lists results in a higher overlap than all other comparisons of friends lists of “possibly matching” OSN profiles of the same comparison set. The results imply that maximum overlaps of three or lower are not sufficient to decide whether two profiles belong to a single natural person or not, because such maximum overlaps occur more than a single time in some cs . In contrast, maximum overlaps higher than three indicate a “match”. A discussion with respect to error rates regarding such indication is presented in the following section.

5.2 Evaluation of linkability results

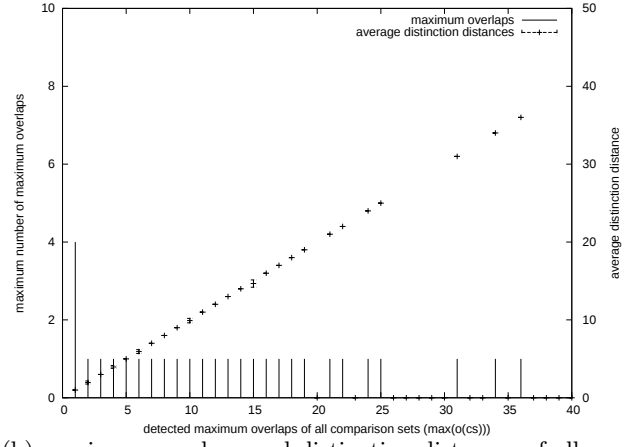
The results indicated that two profiles with a friends lists overlap of more than three belong to the same natural person if this person is registered with the same name in both profiles. This interpretation cannot be proven without talking to the owners of the profiles themselves. However, for being compliant to the German data protection act and for preserving privacy of OSN users, we are unable to contact profile owners. In order to confirm the aforementioned interpretation of the results, we implemented a module that enabled the analysis software to compare additional available information of OSN profiles, such as the given hometown, region, university, date of birth, etc. With this module activated we did another short run with which the software found about 300 comparisons that resulted in an overlap greater than three.

In almost all of these cases the software found a minimum of one other information that appeared in both compared profiles equally. In more than half of all cases the profile image was exactly the same in both profiles. Other attributes that were often available in both profiles were the user’s hometown, region, university or the date of birth. In non of the analyzed comparisons the software found two mutually exclusive pieces of information. This fact confirms the hypothesis that two profiles with the same user name and an overlap of friends lists greater than three are owned by the same natural person.

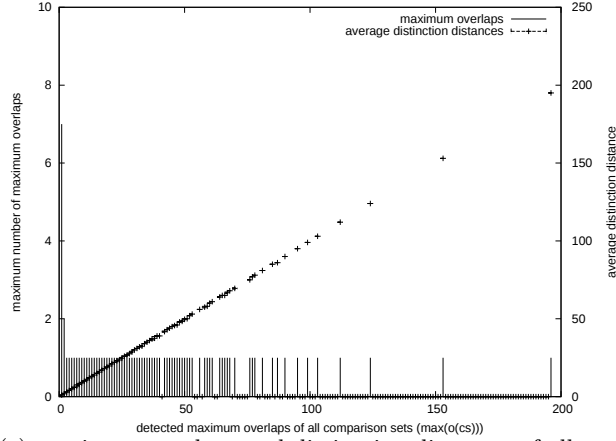
In Section 3 we showed that profile linking might be worthwhile for third parties for gathering more information about natural persons. By use of the presented strategy for profile linking, occurrences of false positives, i.e., accidentally linked profiles owned by different natural persons, cannot be entirely excluded. But, dedicated error rates could not be worked out because of restrictions regarding privacy demanded by the German data protection act. However, it is questionable whether third parties care about any false positives at all. The results of our study hint that the number of false positives is remarkably low. Therefore, potentially



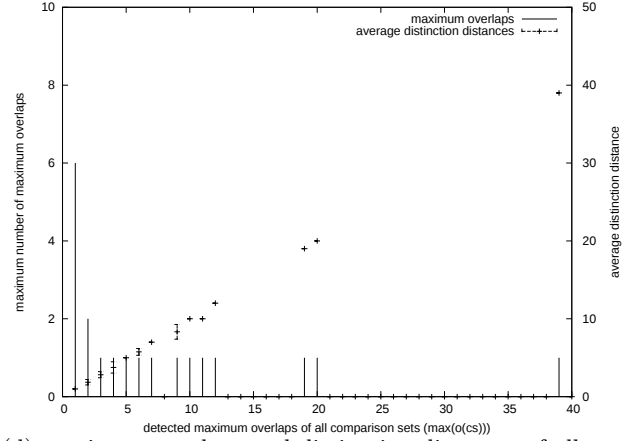
(a) maximum overlaps and distinction distances of all cs of comparisons between Facebook and StudiVZ



(b) maximum overlaps and distinction distances of all cs of comparisons between StudiVZ and Xing



(c) maximum overlaps and distinction distances of all cs of comparisons between Facebook and Xing



(d) maximum overlaps and distinction distances of all cs of comparisons between Facebook and MySpace

Figure 5: Aggregated max-overlaps graphs and average distinction distances. Left y-axis: maximum number of each of detected maximum overlaps inside every cs with respect to two specific OSNs. Right y-axis: average distinction distances. For further information, see Section 4.4

existing false positives are negligible compared to the number of profiles that third parties are able to link correctly. Furthermore, the expected gain of information (as shown in Section 3) countervails less probable occurrences of false positives from the perspective of third parties.

6. DISCUSSION

In the following, we discuss further options to link profiles as well as the linkability of profiles registered with different names. Moreover, we discuss the current risk awareness of users regarding the linkability on the basis of the presented results and selected statements from authors of related work.

6.1 Linking via other available information

As shown in Table 1 many profiles provide information in addition to friends lists. In the following we shortly discuss which information can also be (ab-)used to link OSN profiles. Because of space restrictions, we do not show details on the implementation of the investigation on the linkability based on user's shared attributes.

Generally, the more information a user shares in OSNs the higher the probability to find a possibly existing profile belonging to the same user in another OSN. As the user's date of birth is made available by most users in StudiVZ and MySpace, it is possible to link profiles based on this attribute. Furthermore, this applies to attributes, such as the current residence, hobbies and maybe information extractable from status messages. These attributes are made publicly available in more than 20% of the analyzed profiles of both mentioned OSNs. Less applicable for profile linking are attributes describing a user's university. This information is made accessible by 51.67% of StudiVZ users but only by 8.83% of Facebook users and it is never directly and publicly available in MySpace.

6.2 Profile linkability with different names

We showed that friends list based profile linking is possible if users are registered with the same name at the OSNs. This assumption is reasonable in light of the fact that, for instance, one OSN recently implemented an algorithm that examines the credibility of user's given name and does not

permit incongruous names. In general, a user’s OSN names need not necessarily be the same to be able to link his profiles. It is sufficient to extract a sample of profiles of OSN B that includes the profile of the owner of the investigated profile of OSN A. In this context, Zafarani and Liu demonstrated that a user name from one OSN can be used to identify other user names of the same person in various web communities, such as other OSNs [20]. In this section, we further discuss the linkability of OSN profiles which are configured with different names. Furthermore, we contrast the associated privacy threat to the threat to users registered with the same name in several OSNs.

Certainly, profiles of two differently named users who have the same city or the same hometown in common might have a higher overlap than two profiles that are just set up with the same name. The probability that a user with exactly the same name as another user exists who, additionally, has same friends is probably very low. However, the probability that in an evaluated population of users living in the same city two users with different names can be found who have a high number of overlapping OSN friends is obviously higher because of overlapping groups of friends in “real life”. Nevertheless, we are convinced that the metrics *maximum overlap* and *distinction distance* could also be sufficient to link profiles registered with different names.

Referring to table 1, in three of four OSNs the maximum number of OSN profiles returned by a single search request was between 288 and 545 independent of the given search string. These maximal sizes of query results indicate that not always every profile that matches the search string is returned when searching within an OSN. If, for instance, a user searches for the term “New York” (if such search function is provided by the OSN), not every member living in New York is returned but only about 300 to 550 of them. In order to find a person registered with a known name this number of query results is probably large enough to find a sought-after profile. However, the probability to find a specific profile, from which only a few attributes are known, is still low. In contrast, sampling of every member with a given attribute would be necessary to link profiles registered with different names.

However, it is not possible to evaluate high numbers of user profiles (in which a specific attribute is published) by using typical search features provided by today’s OSNs. In case OSNs would provide more sophisticated interfaces (such as SQL interfaces) to search for profiles, comparisons of friends lists could be used to link OSN profiles. The threat reduction regarding profile linking by use of different names would nullify in light of such interfaces. This also applies to user names that can be determined by a third party by use of the methods presented in [20].

6.3 Risk awareness of users

Krishnamurthy describes in [9] the significance of preserving privacy in the Internet and especially in OSNs. Krishnamurthy assumes that a possible reason for the complexity of preserving privacy is the ignorance of users regarding the protection of PII. Users share information through OSNs without thinking about possible consequences. Even if they have the possibility to secure their shared information by adjusting privacy settings, they are reluctant, maybe because of the urge for “satisfaction of the needs for belongingness and the esteem needs through self-presentation” [7].

Consequently, the amount of publicly available PII and, thus, options to link several OSN profiles of single natural persons paves the way for third parties to create comprehensive virtual appearances of users. Even though Facebook gains more and more market share, the linking of profiles might still be profitable for third parties due to the facts that other OSNs still exist, many of them are growing as well, and some OSNs are dedicated to a specific purpose (business contacts, dating, etc.). In general, Torkjazi et al. show that social networks tend to experience a phase in which users are moving to another OSN after a phase of an extensive growth [18]. Thereby, most users probably do not delete old and no longer used profiles and some OSNs do not even provide a possibility for deletion.

As early as 2004, Acquisti wrote that only the combination of the aspects *technology* and *risk awareness* has the potential of successfully solving the privacy problem whereas any of these aspects alone will most probably fail [1]. Recently, Krishnamurthy wrote, “From an awareness point of view, the situation is pretty bad” [9]. As we showed in this study (see Section 3), offering the option to adjust privacy settings is not sufficient to encourage people to use such settings adequately. Meanwhile, the media attention regarding privacy risks in OSNs is urging the majority of users to make use of privacy settings for some of the most critical attributes. However, we showed that in most profiles more than enough information remains publicly available to link OSN profiles of the same natural person. This study might increase the awareness of users and might motivate users to adjust privacy setting via corresponding technology provided by OSNs.

7. CONCLUSION

Online social networks allow users to publish personally identifiable information and people are using this opportunity extensively. However, users expect their data to be visible to users of their respective OSN only, e.g., private information being accessible to their Facebook friends only and not to their Xing contacts. In this paper, we demonstrated how several OSN profiles of a user can be linked at low cost to gather pieces of information. Particularly, we showed that by just comparing the names found within an account’s friends list profiles can be reliably linked. Additionally, we discussed that by aggregating OSN profiles and combining user specific information found within different OSNs third parties are able to create much more valuable data sets than possible by just extracting information provided in a single OSN profile.

To get an idea of the potential threat regarding profile linking, we reported how often specific pieces of information are made publicly available by users in four of the most popular OSNs. We showed that friends lists are made publicly accessible in between 40% and 67% of all analyzed OSN profiles. Moreover, we found that the availability of attributes differs with respect to the OSN. In turn, this confirms the statement that profile linking is profitable for third parties because of the increased amount of information they can gather and associate to a single natural person.

We further showed that even a small overlap of two compared friends lists is sufficient to assume that associated profiles correspond to the same natural person. We assumed that the registered names of the profile owners are the same

in both profiles³. In this case, we identified that more than three friends have to appear in both compared friends lists to decide whether two compared friends lists belong to profiles of the same user. If an overlap higher than three is detected, the information shared in two OSN profiles can easily be linked by third parties. If users registered their OSN profiles with different names, profile linking might also be possible on the basis of comparisons of friends lists. Only the characteristics regarding the introduced distinction distance and the amount of false positives might differ from comparisons under the aforementioned assumption.

It is desirable that evolution of user behavior regarding the participation in OSNs will be closely monitored in the coming years. However, the presented study might motivate to adjust privacy settings more thoughtfully and should motivate more users to not share as much PII publicly.

8. REFERENCES

- [1] A. Acquisti. Privacy in electronic commerce and the economics of immediate gratification. In *Proc. of the 5th ACM Conf. on Electronic Commerce*, EC '04, pages 21–29, New York, NY, USA, 2004. ACM.
- [2] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your contacts are belong to us: automated identity theft attacks on social networks. In *Proc. of the 18th Int'l Conf. on World Wide Web*, WWW '09, pages 551–560, New York, NY, USA, 2009. ACM.
- [3] G. Brown, T. Howe, M. Ihbe, A. Prakash, and K. Borders. Social networks and context-aware spam. In *Proc. of the 2008 ACM Conf. on Computer Supported Cooperative Work*, pages 403–412, New York, NY, USA, 2008.
- [4] D. H. Chau, S. Pandit, S. Wang, and C. Faloutsos. Parallel crawling for online social networks. In *Proc. of the 16th Int'l Conf. on World Wide Web*, WWW '07, pages 1283–1284, New York, NY, USA, 2007. ACM.
- [5] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proc. of the 2005 ACM Wksp. on Privacy in the Electronic Soc.*, WPES, pages 71–80, New York, NY, USA, 2005. ACM.
- [6] P. Jaccard. The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11(2):37–50, 1912.
- [7] H. Krasnova, T. Hildebrand, O. Günther, A. Kovrigin, and A. Nowobilska. Why participate in an online social network: An empirical analysis. In *Proc. of the 16th European Conf. on Information Systems*, 2008.
- [8] H. Krasnova and N. F. Veltri. Privacy calculus on social networking sites: Explorative evidence from germany and USA. In *Proc. of the 2010 43rd Hawaii Int'l Conference on System Sciences*, HICSS '10, pages 1–10, Washington, DC, USA, 2010. IEEE.
- [9] B. Krishnamurthy. I know what you will do next summer. *SIGCOMM Comput. Commun. Rev.*, 40:65–70, Oct. 2010.
- [10] B. Krishnamurthy and C. Wills. Characterizing privacy in online social networks. In *Proc. of the 1st Wksp. on Online Social Networks*, WOSP '08, pages 37–42, New York, NY, USA, 2008. ACM.
- [11] B. Krishnamurthy and C. Wills. On the leakage of personally identifiable information via online social networks. *SIGCOMM Comput. Com. Rev.*, 40:112–117, Jan. 2010.
- [12] B. Krishnamurthy and C. Wills. Privacy leakage in mobile online social networks. In *Proc. of the 3rd Conf. on Online Social Networks*, WOSN'10, pages 4–4, Berkeley, CA, USA, 2010. USENIX Association.
- [13] S. Labitzke, J. Dinger, and H. Hartenstein. How I and others can link my various social network profiles as a basis to reveal my virtual appearance. In *LNI - Proc. of the 4th DFN Forum Com. Techn., GI-Edition*, 2011.
- [14] C. A. C. Lampe, N. Ellison, and C. Steinfield. A familiar face(book): profile elements as signals in an online social network. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '07, pages 435–444, New York, NY, USA, 2007. ACM.
- [15] H. Liu and P. Maes. Interestmap: Harvesting social network profiles for recommendations. In *Proc. of the Beyond Personalization Wksp.*, 2005.
- [16] M. Motoyama and G. Varghese. I seek you: searching and matching individuals in social networks. In *Proc. of the 11th Int'l Wksp. on Web Information and Data Management*, WIDM '09, pages 67–75, New York, NY, USA, 2009. ACM.
- [17] Z. Stone, T. Zickler, and T. Darrell. Autotagging facebook: Social network context improves photo annotation. In *Proc. of CVPR Wksp. on Internet Vision*, 2008.
- [18] M. Torkjazi, R. Rejaie, and W. Willinger. Hot today, gone tomorrow: on the migration of MySpace users. In *Proc. of the 2nd ACM Wksp. on Online Social Networks*, WOSN '09, pages 43–48, New York, NY, USA, 2009. ACM.
- [19] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel. A practical attack to de-anonymize social network users. In *Proc. of the 2010 IEEE Symp. on Security and Privacy*, SP '10, pages 223–238, Washington, DC, USA, 2010. IEEE Computer Society.
- [20] R. Zafarani and H. Liu. Connecting corresponding identities across communities. In *Proc. of the 3rd Int'l Conf. on Weblogs and Social Media*, ICWSM, 2009.

APPENDIX

A. NOTE ON COMPLIANT SAMPLING

As we want to and have to act compliant to the German data protection laws, the designed software used for the presented study preserves a maximum of privacy. The underlying compliant concept is published in a previous paper [13] in which we describe preliminary work for the study presented in this current paper. For instance, to be compliant to the German law, the software generates random name pairs on the basis of large lists of popular German first and last names and uses this unrevealed name pairs as search strings for sampling of OSN profiles. Furthermore, the software discards information about users provided by OSN profiles subsequent to statistical calculation, so that conclusions regarding specific analyzed profiles are not possible on the basis of stored statistical data. Of course, the network load produced by the software has also been taken into account. The analysis software delimits the number of outgoing packets and prevents accidental flooding.

³Recently, one OSN implemented an algorithm to check whether user's given name seems to be his full name.