A Study on the Unawareness of Shared Photos in Social Network Services

Benjamin Henne
Distributed Computing & Security Group
Leibniz Universität Hannover
Hannover, Germany
Email: henne@dcsec.uni-hannover.de

Marcel Linke
Distributed Computing & Security Group
Leibniz Universität Hannover
Hannover, Germany
Email: linke@dcsec.uni-hannover.de

Matthew Smith
Usable Security and Privacy Group
Universität Bonn
Bonn, Germany
Email: smith@cs.uni-bonn.de

***Abstract*—The habit of sharing huge amounts of photos on the social Web is creating many new privacy issues that have been addressed in a multitude of research efforts in the past few years. Most of this research focused on securing users' own photos and managing their online identity in terms of access control. However, implications for other people have not yet received sufficient attention. Photos affect the privacy of all depicted people and those mentioned in metadata. Therefore awareness about photos is a key issue for protecting privacy. While prior work studied only self-reported ratings, we quantify how aware people are of shared media. In the context of Facebook we show user-unawareness based on a study we conducted via an app combined with two supplementary questionnaires. We present data about 75 million photos and metadata as well as study results of 2254 participants.**

I. Introduction

The publication of photos has been a point of contention in privacy debates for over a century [1]. While this used to be a problem mainly for journalists and press agencies in the past, such issues have become omnipresent today. Taking a photo that depicts others was a far smaller problem when people collected photos in paper photo albums at home. However, due to the massive sharing of photos on the social Web, many more photos become public and raise severe privacy concerns. This issue was raised to the research community by Vint Cerf in his keynote at CHI 2013 when talking about photo sharing, privacy issues of third parties, and the sensible use of technology that we are developing. With the use of modern high-resolution cameras, additional threats are created even to people in the background of photos. The magnitude of smart devices that are able to immediately share photos online have driven this problem to the extreme and this trend is set to continue with the introduction of devices such as Google Glass. While research has found ever more ways of conveniently sharing and mashing up social media, there is very little research on systems that allow users to deal with media of others that might raise privacy concerns. We believe that one reason for this is that the extent of the problem is not well understood and that there is a false sense of complacency amongst users as well as researchers.

In the past few years, research addressed different issues arising from online photo sharing, finding that people feel that their privacy is threatened by photos taken by nearly any other person, no matter if they are from inside their social circle [2] or outside [3]. In social network services, the activities of photo sharing and tagging people have often become part of personal identity management [4], [5] instead of simply sharing visual impressions with friends. Although a rising number of people become more cautious about sharing their personal data in social network services, users still create threats to their own privacy by using inappropriate access control settings [6]. While most research focused on securing the users' own photos and managing their online identity in terms of access control [7], implications for other people have not yet received sufficient attention [8]. Shared photos affect the privacy of all people visible in a photo and mentioned in metadata like tags and comments.

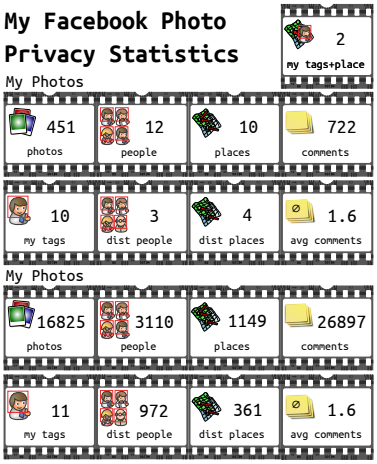
One key human factor of protecting one's privacy against threats that might arise from photos of other people is being aware of such photos and the associated threats. While privacy awareness has already been a topic of research [9], most of the work focused on the awareness of being tracked, data being visible to others or collected by services and apps [10]–[14]. The topic of awareness about shared media has only been explored in few studies [15]–[17] to date. In prior work [16], we showed that people report that they generally feel unaware about other peoples' photos on the Web that could affect them and that they would like to have tools to assist them. However, the initial problem that still needs to be addressed in the field is making people aware of potential privacy problems of other peoples' media. They must realize that they might have a problem. People must become aware of the problems' dimension to let them realize that their privacy may be endangered much more than they might expect and that they might need to address these issues by using appropriate privacy-enhancing tools that allow to cope with the huge amount of shared photos. To lay the foundation for tackling this issue, we developed a Facebook app that allows users to see the scope of shared photos in their circle of friends.

In this work, we present a study based on this app combined with a preceding and a post-questionnaire, in which we quantify how unaware people are of the scale of the problem. We present the data collected by the app, which contains an evaluation of 75.7 million photos visible to 2753 app users of which most were recruited via an online news site and a radio broadcast. By comparing the data with values estimated by the 2245 participants who also took part in the preceding questionnaire, we quantify the lack of knowledge of the amount of shared photos and metadata. As the participants proved a massive lack of knowledge, we must conclude that

they are unaware of (the dimension of) potential privacy issues. Furthermore, we present results from the post-questionnaire on the users' feelings about being aware of online photos in general and compare our results to results from prior work. Our work shows that people are currently not aware of the amount of potentially threatening photos shared in their circle of friends and thus may see little reason to act.

II. Study & App Design

During execution the app *Photo Privacy Statistics* presented users a short pre-questionnaire about their expectations about the amount of photos shared by their friends as well as corresponding tags, in particular how many times they believed they had been tagged. The app then calculated the statistics about the users' own photos and displayed them in the web browser. Since the calculations for friends' photos could take a significant amount of time, this was executed in the background and users were notified by email or Facebook notification on completion. An example result is shown in Figure 1. In the notification email and at the bottom of the results page the users were invited to fill in a short post-questionnaire. Answering the questionnaires was optional. Users were also offered a button to post their results on their Facebook Timeline with a catchy image and a textual explanation as shown in Figure 1. The app and questionnaires were available in English and German.



"My 295 friends share at least 16825 photos with 3110 person tags and 1149 place tags. 26897 comments have been made to those photos.
I was tagged 11 times, 2 times on photos with a location tag. 972 other people were tagged as well. 361 different places were tagged.
What about you? Try the Photo Privacy Statistics app!
18.6 % of my friends do not share photos or deny access to photos for apps others use."

Fig. 1. Image and generated text as used in the app's "Post to Timeline" feature for a random user with average friends count

A. Recruitment of Participants

To gather a broad range of participants we recruited users via online news, a radio broadcast and via Facebook Timeline posts. In the first phase users were mainly acquired from the peer circles of our research group by announcing the app and posting results on Facebook Timelines. We collected data from 113 people, consisting mainly of university affiliates and students, forming the *inital* group. Afterwards, the app was announced by a public press information of our university, resulting in a radio interview about the app broadcasted by the German public broadcasting radio station *DRadio Wissen* in the context of a special theme day called *private affair*, gaining us the *radio* group of 79 people. Following the radio broadcast, an online news article about the app was published on the IT news website of the biggest German yellow press paper (*Bild*), enabling us to recruit the *yellow press* group with 2561 users, of which at least 1698 people were from Germany, while we accumulated more international users as the app's user count rose—probably due to Facebook search.

In this work we present data from 2753 app users. 2275 of them answered at least one question of the optional pre-questionnaire; 2245 estimated the amount of shared photos or metadata. The optional post-questionnaire was answered by 269 people.

B. Demographics

Of the 2753 app users, 84.5 % reported being male, 15.1 % being female, and 0.4 % did not offer gender information. In the pre-questionnaire 2139 users gave their ages, ranging from 13 to 77 years with mode = 26, median = 30 and mad = 10.4. We manually checked the validity of extreme values using public profile pictures for users whose access tokens were still valid. All data we stored was anonymized conforming to German data protection laws. Since most of our users were acquired via German online news, most of them originated in German-speaking countries. At least 64.4 % of the users were from Germany (Facebook profile: 27 %, geo-coding IP: 37.4 %); 1.2 % from Austria; 1% from Switzerland; and 0.4 % from the Netherlands. The United States were the top non-German-speaking user origin with 0.6 %. We could not obtain country data for 28.6 % of our users. The remaining 3.8 % of our users were from 40 other countries on all continents.

C. Statistical Methods

Normality testing (Shapiro-Wilk, D'Agostino), histograms and reported quantiles showed significant deviations from normal distributions for all participants' answers as well as all data collected via Facebook. Thus we employ non-parametric test measures to discuss our results. While we mainly use quantiles and median absolute deviation from the median (mad) to describe data distribution, we also report the mean and standard deviation, especially in cases of comparison to prior works' results that report those measures. As most p values are very small, all results reported as being significant are still significant under Bonferroni correction.

III. On App Users' and their Friends' photos

With respect to privacy-related data and metadata, the anonymized dataset created by our Facebook app can be summarized as follows: The dataset contains information about 75.7 million photos of 2753 app users and 572 K of their 817 K friends. About 30 % of the friends did not share photos or denied access to photos for *apps others use*. The portion of photos owned by non-users was 99.2 %. Our users had a mean of 296 friends. The distribution of friends count was described by quantiles: $Q_{0.25} = 116, Q_{0.5} = 221, Q_{0.75} = 383$, and $Q_{0.95} = 738$, with a maximum of 5405 friends.

11.3 % (8.5 M.) of all photos had a location tag and the 610 K different locations were tagged 14 times in the mean. 55.7 % of the different locations occurred only once, 13 % twice, and 11.4 % occurred 3 to 5 times. The frequency distribution had a long tail ($Q_{0.9} = 19, Q_{0.95} = 48, Q_{0.99} = 170$)

with the most frequent location (Germany's capital city Berlin) occurring 90837 times.

In 22.4 % (17 M.) of all photos at least one person was tagged with link to corresponding Facebook profiles. These person tags included photos a person was marked in as well as mentions like a profile reference in a photo's caption. Additional 0.5 % of all photos had at least one non-linked text-only tag but no linked tags, while 0.8 % of all photos contained both.

Altogether, 35 million tags (34 M. with links to profiles) were attached to those photos; each photo had a mean of 2 tags; the maximum was 205 tags. The tag counts per photo were distributed as follows: 56.2 % of the photos had only one tag; 25.2 % had two tags; 8.7 % had three tags; 4 % had four tags; 1.9 % had five tags; 2.7 % of them had six to ten tags; and 1.3 % more than ten tags.

6.3 million different people were tagged altogether. Frequencies of how often each person was tagged are characterized by quantiles: $Q_{0.25} = 1, Q_{0.5} = 2, Q_{0.75} = 4$, $Q_{0.95} = 199$ and the most frequently tagged person was tagged 6229 times. However, we must keep in mind that these numbers present lower bounds, since tags in friends' photos include tags of friends of friends, while these people were very probably tagged in others' photos as well, which were outside of scope of our Facebook app. If we restrict our data to app users, 2421 of them (87.9 %) were tagged with the following frequencies: 18.3 % up to three times; 21.4 % four to ten times; and 60.3 % of them were tagged more than ten times including a long tail with $Q_{0.5} = 16, Q_{0.75} = 38, Q_{0.9} = 88, Q_{0.95} = 136$. In photos with tags of our app users which additionally had a location tag, 1758 users (63.9 %) were tagged at least once. 22.5 % of those users were tagged in exactly one such photo; 14.3 % in two photos, and 10.2 % of them in three. Altogether, 74.6 % of them were tagged in up to ten of those photos, while 25.4 % were distributed in a long tail with $Q_{0.9} = 23, Q_{0.95} = 42$ and a maximum of 573 photos.

Comments and captions may include texts that raise privacy concerns. Like tags without profile links, they usually do not trigger notifications. 28.3 % of all photos had at least one comment. The mean comment count of those images was 4.5 with a maximum of 20244 (distribution of comments per photo: $Q_{0.25} = 1$, $Q_{0.5} = 3, Q_{0.75} = 5, Q_{0.95} = 14$). 29.4 % of all photos had a caption which was neither empty nor ended with an image file extension (gif/jpg/png), which we checked to exclude captions being file names as automatically set by some upload tools.

The app results presented to users contained 17 summarizing values from each user's personal point of view (cf. Figure 1). Figure 2 shows the frequency distributions of the 9 most important values: photos of the user and friends' photos visible to them (precisely: visible to our app), corresponding location and person tags, tags of themselves, and all photos with a location tag they were tagged in. Long tails of frequency distributions were cut for visualization reasons. Users' personal results also included counts of distinct locations and distinct tagged persons as well as comment counts and average comments per photo.

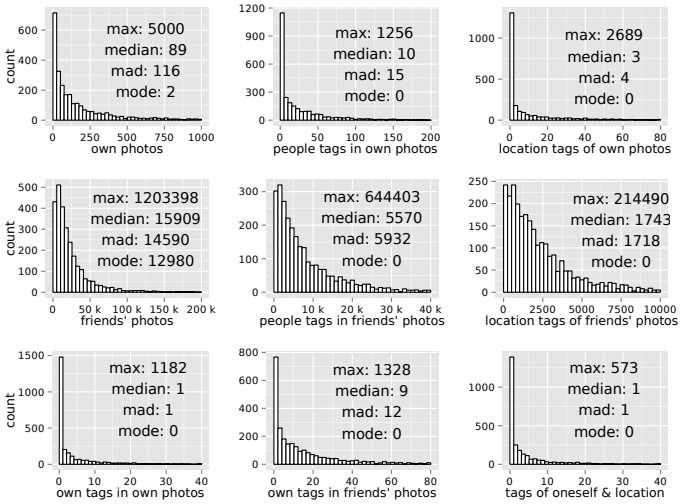


Fig. 2. Frequency distributions of photos, location tags, people tags and photos having both kinds of tag per app user in the presented dataset

A. Facebook Privacy Setting

Facebook users can setup a privacy option called *apps others use*, which allows users to opt out of allowing apps used by their friends access to their data like photos or profile information. From our app's perspective the Facebook API states that a user who opted out has no photos and the response is indistinguishable form a user who actually does not share any photos. However, sharing absolutely no photos on Facebook is hard, since photos posted to the Timeline and profile pictures are included as well. During initial testing, every time we found the API returned no photos, this privacy option was enabled. While there certainly are users who share no photos at all, the lion's share probably enabled this privacy feature. So we interpret no shared photos as a sign that this privacy option is set. Consequently, if we assume this value as a measure for opt-outs via *apps other use*, the number of photos is even higher than presented before as about 30% of our users' friends did not share photos with our app.

The median percentages of friends who did not share photos differed across the user groups: the *initial* group had a median of 35.1 %, the *radio* group 32.7 %, and the *yellow press* group 26.2 %. Differences between the yellow press group and the others were significant (Kruskal-Wallis test $\chi^2_2 = 68, p < 0.001$; pairwise Wilcoxon-Mann-Whitney tests: initial ~ radio $Z = 0.7495, p = 0.45$; initial ~ press $Z = -6.7088, p < 0.001$; radio ~ press $Z = 5.0354, p < 0.001$).

B. Removal of App Permissions

On our app's landing page as well as in the notification email, we told users that they could remove the app (permissions) after receiving the notification, since the result visualization was independent of the Facebook app itself. When checking access token validity of app users more than two weeks after app usage by exemplarily re-requesting their friends lists, 89.1 % of the users still allowed access to their data. Interestingly, we found a difference in behavior of the three user groups: 92.9 % of the *initial* group (mainly students) removed permissions after usage, 16.5 % of the *radio* group removed permissions and only 7.1 % of the *yellow press* group

removed permissions after app usage. These differences were significant (Pearson's Chi-squared test $\chi^2_2 = 820, p < 0.001$; pairwise $\chi^2_1 = \{111, 8.4, 819\}, p < 0.004$).

IV. Expectations versus Reality

When comparing expectations of users to real values, we focus on the 2245 users who gave estimations on the count of friends' photos or corresponding tags of locations or people. The demographics of the subset users who answered the survey were virtually identical to that of all app users. In the following, we report important values as portion of all these participants (abbr: oap) besides relative percentages.

A. Friends' Photos

In the pre-questionnaire we asked our participants to estimate how many photos shared by all their friends they can view on Facebook. Possible predefined answers were: *no answer* (default), *no idea*, *less than 50*, *50*, numbers described by $d \cdot 10^k, d \in \{1, \ldots, 9\}, k \in \{2, \ldots, 5\}$, *1000000*, and *more than 1 M*. In their answers 77.1 % of the 2245 participants entered an estimation of the number of photos shared by their friends; 22.4 % of them answered with *no idea* and 0.5 % of them gave no answer about the count of shared photos. The participants mostly estimated the number of photos shared by their friend to be 1000, which was also the median of all estimations (other quartiles: $Q_{0.25} = 400, Q_{0.75} = 8000$). The median of the real numbers of shared photos was 15909 ($Q_{0.25} = 7722, Q_{0.75} = 30687$) with a maximum of 1203398. The frequency distributions of both are shown as overlaid histograms in Figure 3.

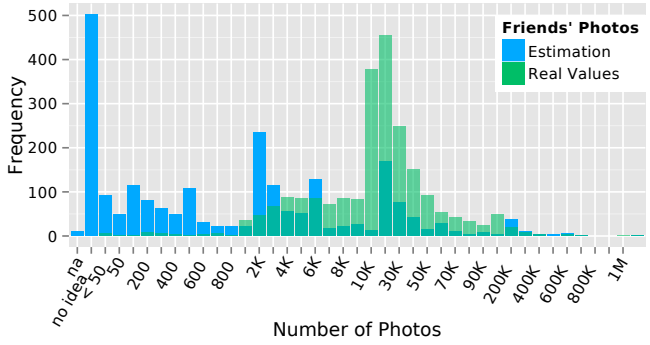


Fig. 3. Users' estimations and measured counts of photos

The first question we need to answer is how many participants estimated the number of photos shared by their friends correctly. We judge a user to have estimated correctly if the real value is closer to the estimated number than to its predecessor or successor (e. g. 1000 matching $(950 : 1500]$ real photos). When applying this measure, we achieve the following partition of user estimations: 3.9 % of the estimations were considered as correct, while 96.1 % of them were incorrect. If we relax the measure of correctness, so that intervals range from an answer's predecessor to its successor (i. e. 1000 matching $[900 : 2000]$ photos), 8.2 % of the estimations were correct in terms of our definition (6.3 % oap.), while 91.8 % were not. The latter one includes predefined answers as having been interpreted by participants as *at least* and *up to* relative to its neighbors. We further rely on that measure, since we did not specify how predefined values should have been considered by participants with the aim to ease their estimation.

The next question is how strongly the remaining 70.8 % of all participants (1590 people) misestimated the numbers. Since the initial comparison of frequency distributions does not allow any conclusion about the error of a single estimation, we compare the estimated values and the count of shared photos captured by the Facebook app by calculating the *misestimation* as the difference of both values. 8.6 % of the people who misestimated overestimated the number of shared photos (6.1 % oap.), while 91.4 % of them underestimated the number of photos that their friends share with them (64.7 % oap.). The frequency distribution of all misestimations is shown in Figure 4. To exemplarily visualize absolute misestimations, Figure 5 shows a random uniform subsample of 100 estimations of participants with misestimation errors.

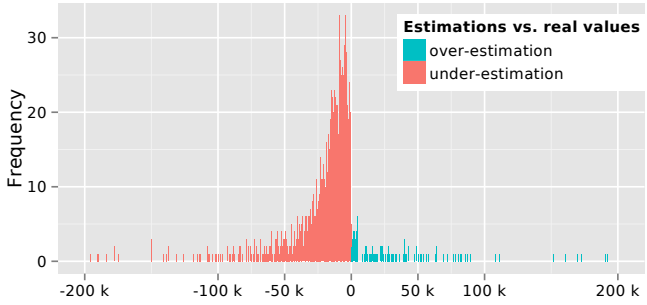


Fig. 4. Histogram of misestimation: friends' photos

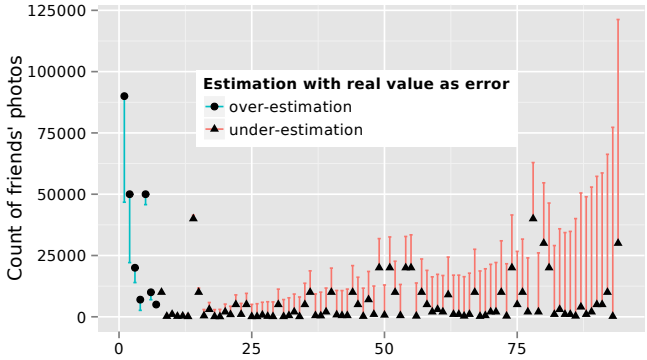


Fig. 5. Random uniform subsample of 1732 estimations of friends' photos (points) compared to real values (error bars) ordered by error

We further investigate the *factor of misestimation* for considering its degree by dividing both values. For this, we always divide the bigger by the smaller value, while a negative value indicates underestimation and a positive one means overestimation. The answer *less than 50* was recoded to 25 as the average of possible values and the value 1000001 was used for calculations of *more than 1M*. The factor of misestimation was distributed in the range from -38989 to 258 ($Q_{0.05} = -312, Q_{0.95} = 2$) with the median factor being an underestimation of factor -11 ($Q_{0.25} = -38, Q_{0.75} = -3$). Finally, we classify misestimations by the decimal *magnitude of misestimation* factors. The frequency distribution of magnitudes is shown in Figure 6. 91.9 % of overestimations were made with the magnitude of 1 (5.6 % oap.), while 8.1 % of overestimations had a higher magnitude (0.5 % oap.).

In Comparison, underestimations were mostly made with higher magnitudes: 43.7 % of the underestimations were made with the magnitude of 10 (28.3 % oap.). Likewise 42 % of underestimations were made with the magnitude of 1 (27.2 % oap.). Additional 13.1 % of underestimations were made with the magnitude of 100 (8.5 % oap.) and 1.2 % with magnitudes 1000 and 10 k (0.7 % oap.).

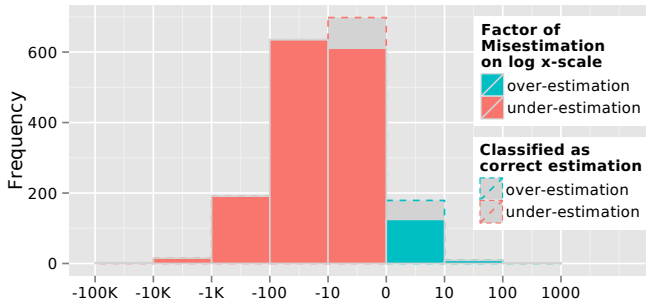


Fig. 6. Histogram of magnitude of misestimation of photos

When we count participants who estimated correctly in terms of our definition, only the frequency of misestimations with magnitude of 1 differs which was predetermined by interval design. Classifying the correct estimations, 62 % of them were underestimations and 38 % were overestimations. These additional values are shown as grey top on bars in Figure 6. Summing up, only a limited subset of participants were able to estimate the number of photos shared by friends correctly. Most participants underestimated the number.

In the pre-questionnaire about 6 % of all participants estimated the count of friends' photos to be in the same dimension as the number of photos they share themselves. We cannot verify if this based on mistake or misunderstanding, or if this mirrored their real belief. Comparing data of participant subgroups (*initial*, *radio*, *yellow press*), we could not find any significant differences neither in the app data nor in misestimations.

B. Location Tags

Next, we asked our participants to estimate how many photos shared by their friends have a location tag. Possible answers were: *no answer* (default), *no idea*, *none*, *few (< 10 %)*, *10 / 20 / 33 / 50 %*, *more (> 50 %)*, and *all*. Numeric answers were supplemented by more imaginable texts like *every* i^{th} *photo*, which also were the reason for choosing the non-equally distanced values. In their answers 81.2 % of the 2245 participants entered an estimation about the portion of photos with location tags; 16.6 % of them answered with *no idea*; and 2.2 % of them gave no answer.

The ratio of participants' answers is shown in Table I. Their median estimation was that 20 % of their friends' photos have a location tag ($Q_{0.25} = 10\,\%, Q_{0.75} = 33\,\%$), while the median of real percentages was 10.8 % (mad = 4 %, $Q_{0.25} = 8.4\,\%, Q_{0.75} = 13.9\,\%$). Grouping participants by their estimation, the medians of real values were nearly identical (10.5 to 11.7) with values in ranges between 0 and 41.4 (cf. Table I). Cutting values into intervals of multiples of 5, the mode interval of all subgroups and that of the complete data was $[10 : 15)$.

TABLE I. CORRECTNESS OF ESTIMATIONS OF LOCATION TAGS

Est.	Ratio	$p \in$ range	median	correct if	$^{\text{correct}}/_{\text{all}}$	*md*
0	0.4 %	(6.6 : 12.3)	11.3	[0 : 5]	0/7	-6.2
< 10	15.2 %	(1 : 26.6)	10.5	[0 : 10]	138/277	-3.8
10	23.4 %	(0 : 35.8*)	10.5	[5 : 15]	343/426	-2.5
20	23.9 %	(1 : 40.3)	10.9	[15 : 26.5]	84/436	5.0
33	19.1 %	(1 : 41.4)	11.0	[26.5 : 41.5]	3/349	15.5
50	9.8 %	(2 : 38.1)	10.8	[41.5 : 75]	0/178	30.7
> 50	7.7 %	(3 : 26)	11.7	[50 : 100]	0/140	38.3
100	0.5 %	(2.3 : 26)	10.6	[75 : 100]	0/9	64.4
all	100 %	(0 : 41.4*)	10.8	31.2 % correct estimations		

p = participant's real value * excluded one outlier of 74.1

As above, we have to select a measure for deciding if a participant estimated correctly. We value an estimation as correct if the rounded real percentage was closer to the estimated value than to its neighbors, or, respectively, if the rounded real value was in the interval in the case of the predefined interval answers. Table I contains details about the matching intervals, the amount of correct estimations as well as the median distance (*md*) between the real values and intervals. When applying the described measure, we find 31.2 % of the estimations to be correct (25.3 % oap.).

To examine the amount of misestimation of the remaining 55.9 % of all participants (1253 people), we compute the *misestimation* as the difference between estimated and real values. Only 17.8 % of all misestimations (10 % oap.) were underestimations of the portion of photos having a location tag, while 82.2 % of them (45.9 % oap.) were overestimations of the amount of tagged photos. Figure 7 shows the frequency distribution of misestimations. It additionally shows the participants we regarded as having estimated correctly (58 % underestimation, 42 % overestimation). Compared to the number of photos, participants seem to be more aware about the portion of photos having location tags. Most of them overestimated the portion.

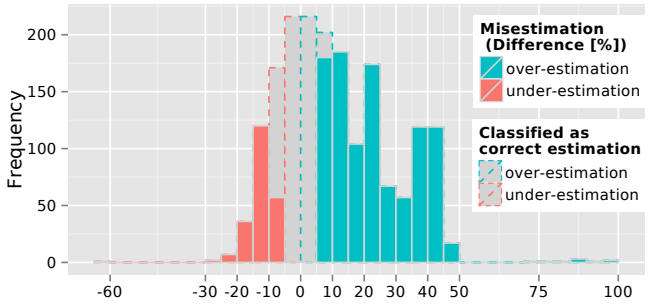


Fig. 7. Histogram of misestimation: photos with location tag

C. Person Tags

We also asked our participants to estimate how many photos shared by their friends have person tags. The set of predefined answers was the same as for location tags. 83.3 % of the 2245 participants entered an estimation about the portion of photos with people tagged in it; 13.9 % of them answered with *no idea*; and 2.8 % of them gave no answer.

Table II shows the ratio of participants' answers. The median estimation of the participants was that 33 % / every

3rd of their friends' photos have person tagged in it ($Q_{0.25} = 10\,\%, Q_{0.75} = 50\,\%$), while the median of real percentages was 17.8 % (mad = 6.9 %, $Q_{0.25} = 13.5\,\%, Q_{0.75} = 22.7\,\%$). Grouping participants by their estimation, real values ranged between 0 and 47.5 (cf. Table II). The median of real percentages of the groups steadily rose with the estimated value from 11.2 to 22.6. Cutting values into intervals of multiples of 5, the mode increased with estimations, in fact $[5 : 10)$ once; $[10 : 15)$ twice; $[15 : 20)$ three times; $[20 : 25)$ twice.

TABLE II. CORRECTNESS OF ESTIMATIONS OF PERSON TAGS

Est.	Ratio	$p \in$ range	median	correct if in	correct/all	*md*
0	0.5 %	(4.4 : 21.3)	11.2	[0 : 5]	1/10	-8.1
< 10	11.0 %	(0 : 42.7)	14.3	[0 : 10]	42/206	-5.8
10	14.0 %	(3.2 : 32.2)	15.2	[5 : 15]	131/262	-4.7
20	19.5 %	(2.2 : 36.9)	16.5	[15 : 26.5]	215/363	2.6
33	20.3 %	(4.2 : 36.8)	18.6	[26.5 : 41.5]	44/380	8.9
50	16.8 %	(3.4 : 47.5]	19.4	[41.5 : 75]	3/313	22.2
> 50	17.0 %	(8.7 : 44.2)	21.8	[50 : 100]	0/318	28.2
100	0.9 %	(8.3 : 28.9)	22.6	[75 : 100]	0/17	52.4
all	100 %	(0 : 47.5)	17.9	23.3 % correct estimations		

p = participant's real value

Since this question had the same set of predefined answers as the question on location tags, we applied the same measure for determining correct estimations. Details about the matching intervals, the amount of correct estimations and the median distance (*md*) between the real values and intervals are shown in Table II. In the case of person tags, only 23.3 % of the estimations were determined to be correct (19.4 % oap.). However, our results suggest that participants' answers about the amount of person tags were more informed compared to answers about location tags. There is at least a trend that real values correspond more strongly to estimations in this case when considering the median and mode values of the described subgroups, or comparing ranked values of estimation and reality (Spearman's $\rho = 0.316, p < 0.01$). This is potentially an interesting finding and we will follow up in future work why this estimation is more accurate than the previous in the hopes of finding a way to make users' awareness of different kinds of metadata more accurate.

Again, the amount of *misestimation* of the 63.9 % of all participants (1433 people) who did not estimate correctly was calculated as the difference between estimations and real values. Only 22.5 % of the misestimations (14.4 % oap.) were underestimations of the portion of photos having person tags, while 77.5 % of them (49.5 % oap.) were overestimations. Figure 8 shows the frequency distribution of misestimations. It also visualizes the portion of those participants that were considered as having estimated correctly (54.1 % underestimation, 45.9 % overestimation). As for location tags, participants tended to overestimate the portion of photos with person tags.

1) Average Person Tags in a Photo: Finally, we asked the participants to estimate how many people are tagged in a photo with person tags on average. Possible answers were *no answer* (default), *no idea*, *1 . . . 10*, or *more*. 78 % of the 2245 participants entered an estimation about the average number of person tags per photo; 18 % of them answered with *no idea* and 4 % of them gave no answer.

The median answer was that each photo contains 3 person tags (mean = 2.8). Answers were distributed like this: 7.5 % of the estimations stated that there was one tag per photo on average; 39.3 % stated two tags; 35.5 % said three tags; 8 % said four tags; 4.9 % stated five tags; 3.6 % stated six to 10 tags; and 1.2 % of estimations were that more than 10 person tags were in a photo on average. The median of real values of person tags per photo was 2 (mean = 2). The mean numbers rounded to integers were distributed as follows: 1.1 % of participants had access to friends' photos with one person tagged on average; for 93.4 % of them the average person tag count per photo was two; 4.6 % had an average count of three; 0.5 % had an average tag count of four; 0.2 % had an average count of five; an 0.2 % of participants had more than five to ten tags per photo on the average.

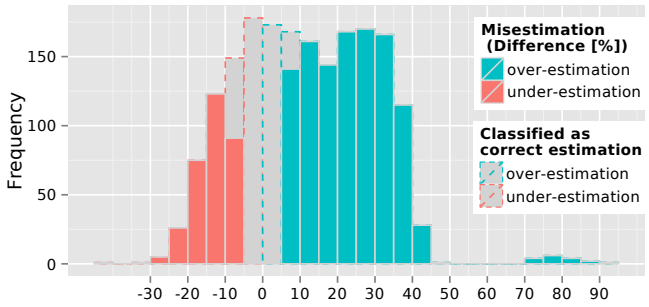


Fig. 8. Histogram of misestimation: photos with person tags

None of the participants who estimated *more than 10* tags per photo estimated correctly: They all overestimated. The maximum mean number of person tags in their friends' photos was 4.6. Focusing on other answers, when comparing each participant's answer with the rounded mean number of person tags in his friends' photos, 39.6 % of the estimations were correct (30.6 % oap.). In case of 43.2 % of the estimations, the estimated values only differed by one tag from the rounded mean of real values (33.4 % oap.). Classifying the *misestimation* as the difference between estimations (options *1* to *10*) and rounded real mean values, 15.5 % of misestimations were underestimations (7.2 % oap.), while 84.5 % of them were overestimations (39.3 % oap.). Figure 9 shows the frequency distribution of the misestimation for all answers. The answer *more than 10* was recoded with the value 11 for calculations.

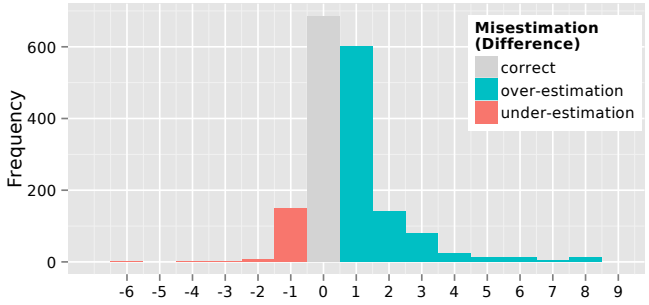


Fig. 9. Histogram of misestimation of average number of person tags in photos with person tags

V. RE-EVALUATION OF ESTIMATIONS

In the post-questionnaire we re-revaluated participants' estimations about the amount of shared photos and their

metadata. 78.8 % of the participants stated to be male; 20.8 % female; and 0.4 % did not specify their gender. Age ranged from 13 to 76 with median 31 and mode 26. At least 67.3 % of them were from Germany; 6.3 % from other countries; and the remaining did not reveal their origin. The median time between sending notifications about completed results and participation was about 5 hours ($Q_{0.75} \approx 14\,\text{h}, Q_{0.95} \approx 2\,\text{d}$) with a maximum of 20 days. Participants' initial estimation values were not shown again.

We asked the participants of the post-questionnaire how the numbers presented by the app compared to their expectations. Differences in mean ratings of the items were significant (Friedman test, $\chi^2_3 = 31.88, p < 0.001$). Participants' answers are visualized in Figure 10.

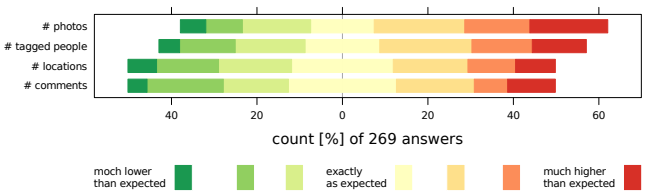


Fig. 10. How did the numbers of photos that your friends share compare to your expectations?

In case of the number of photos shared by their friends, 53.6 % of the 248 participants stated that the number was higher than expected, while 30.7 % stated that it was lower and 15.7 % of them answered that is was exactly as expected. Of the 214 participants who had initially made an estimation about the number of friends' photos, 59.3 % gave consistent answers towards their estimation and app results (less than expected while overestimating; more than expected while underestimation). In contrast, 25.7 % answered inconsistently and the remaining 15 % answered with *as expected*. 12.5 % of the latter were regarded as having estimated correctly and 59.4 % misestimated by a factor up to 10, while 28.1 % made a higher misestimation up to a factor of 850. In case of the four answers from *much higher than expected* to *exactly as expected*, the median misestimation (underestimation) of participants was -9, -6, -4, -2. For the other options the answers were mostly inconsistent. The reported percentages also include people who did not make an estimation before app usage. For comparison Table III shows the percentages of subgroups that did not answer initially or answered with *no idea*.

TABLE III. EXPECTATIONS COMPARED TO REALITY OF POST-QUESTIONNAIRE PARTICIPANTS WHO DID NOT ESTIMATE VALUES IN THE PRE-QUESTIONNAIRE

estimation answer	photos *no idea*	photos *na*	locations *no idea*	locations *na*	person tags *no idea*	person tags *na*
count	34	21	25	22	18	25
lower [%]	32.4	33.4	40.0	31.8	72.2	36.0
expected [%]	20.6	19.0	32.0	45.5	11.1	12.0
higher [%]	47.0	47.6	28.0	22.7	16.7	52.0

na = *no answer* to optional question

In case of the number of location tags in friends' photos, 39.1 % of the post-questionnaire participants stated that the number was higher than expected, while 39.1 % stated that it was lower and 21.8 % answered it was exactly as expected. Of the 226 participants who initially gave estimations, 46 % answered consistently. In contrast, 33.3 % gave inconsistent answers concerning their initial estimation. The remaining 20.7 % answered with *as expected*. 30.4 % of the latter estimated correctly in terms of our definition; 28.3 % of them had real numbers that differed up to 10 % from their estimation's interval; and for the remaining, the difference was higher. In case of the number of person tags in friends' photos, 47.6 % of the participants stated that the number was higher than expected, while 34.3 % stated it having been lower and 18.1 % of them said that is was exactly as expected. Of the 222 participants who initially made estimations, 41.1 % gave consistent answers. In contrast, 41.6 % answered inconsistently concerning their initial guess. The remaining 17.3 % answered with *as expected*. For person tags, comparing estimation results and answers to this question is difficult, since we asked two different questions on person tags in the pre-questionnaire and only one in the post-questionnaire. Overall, only 6 participants gave inconsistent answers towards all initial estimations, but only 27 participants gave completely consistent answers. Time between estimations and answering the post-questionnaire may be a reason.

In the post-questionnaire, we also asked the participants how they felt about the real numbers shown by the app on a 7-point scale from (1) *very upset* to (7) *very enthusiastic* with (4) as *neutral*. As shown in Figure 11, more than half of the participants (56.9 %) stated to feel neutral about the app results. In comparison, 27.4 % stated to be upset (4 % very upset) about results and 15.7 % stated to be enthusiastic about them.

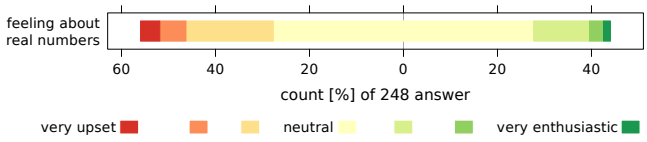


Fig. 11. Please describe your feeling about the real numbers on a scale from *very upset* to *very enthusiastic*.

VI. USERS' FEELINGS ABOUT AWARENESS

In the second part of the post-questionnaire, we wanted to assess our participants' feelings about the awareness of photos in general. Prior work [16] already evaluated people's feelings concerning awareness about photos and metadata in an online survey of 414 participants from a German university. We selected questions of that study to re-evaluate their results in a broader audience and a more realistic setting. Additional to re-evaluation, we investigated a further assumption about the privacy benefit of person tags.

A. Feeling of Being Aware about Online Photos

We asked the participants to rate how well they feel informed about different types (nice/decent, bad/unwanted) of photos of themselves on the Web, using the 7-point scale from (1) *completely sufficient* to (7) *completely insufficient* (cf. Figure 12). The differences in the answers were significant (Wilcoxon test, $Z = -6.64, p < 0.001$). Participants' perceived level of awareness tended to be slightly insufficient in the mean in the case of decent photos (mean = 4.6, sd = 1.9,

median = 5, mode = 7). 56 % of them chose a level worse than neutral, while only 6 % stated that their level of information is *completely sufficient*. In comparison, their average perception was clearly insufficient in case of objectionable photos (mean = 5.2, sd = 1.8, median = 6, mode = 7). While only 4 % stated a level of *completely sufficient*, 70 % of the participants asserted that their level of information about bad photos was worse than neutral to *completely insufficient*. Compared to prior work's results, our participants indicated to feel far worse informed about photos of themselves on the Web. Less than half ($^6/_{22}, ^4/_{11}$) of them stated to be *completely sufficient* informed, while up to twice as many as before ($^{56}/_{25}, ^{70}/_{39}$) stated to be insufficiently informed.

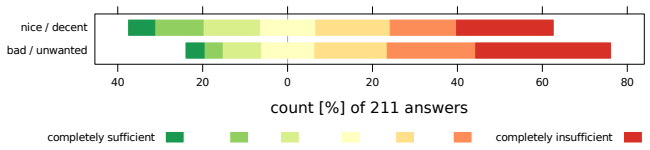


Fig. 12. How well do you feel informed about photos of you anywhere on the Web?

B. Perception of Being Tagged in Photos

Prior work [16] reported that being tagged in shared photos was the most common way how participants found out about photos they were depicted in. 75 % of their participants (94 % indicated being Facebook users) stated to be notified via email when being tagged in a photo. In comparison, 52 % stated to get to know about photos by chance, while less than 40 % got to know about photos by a personal message and 18 % were actively looking for photos of themselves. This leads us to the hypothesis that being tagged can also be perceived as a privacy benefit rather than a reason for privacy concerns. To test this, we asked our participants in the pre-questionnaire whether they perceive being tagged in photos as beneficial for their privacy. On a 7-point scale from (1) *no; it is a big threat to privacy* to (7) *yes; it bears large privacy benefits* with (4) as *neutral* only 15.4 % of the 2013 participants indicated that they perceive being tagged as a beneficial for privacy (cf. Figure 13). While 28.9 % answered with neutral, 55.6 % indicated to perceive it as threat to their privacy (median = 3, mad = 1.5) with 43.2 % in the worst two elements. Surprisingly, although automatic notification about being tagged seems to be the best way of getting to know about photos of oneself, only 15.4 % of the participants agreed to our hypothesis. The reasons behind this are worthy of further study.

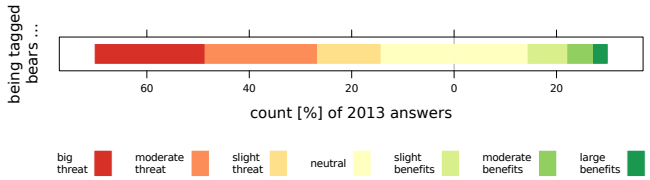


Fig. 13. How you perceive being tagged in a photo?

To further assess the perception of person tags, we asked the participants of our post-questionnaire to rate how much they like different uses/effects of person tags on the 7-point scale from (1) *like it very much* to (7) *dislike it very much* with (4) as *neutral*. The answers of 194 Facebook users (cf. Figure 14) confirmed prior results. Becoming aware of photos of oneself is not the most important effect of tagging. Most participants stated that they significantly prefer (Wilcoxon test, $Z = -3.56, p < 0.001$) finding photos of others (mean = 3.5, sd = 1.6) to finding photos of themselves (mean = 4.0, sd = 2.0). Feelings about finding photos of oneself tend to neutral. This backs the result that people do not see much awareness benefit in being tagged. Participants indicated that they rather dislike that others can find photos of themselves (mean = 4.7, sd = 1.7). On average the participants who tended to dislike photos being found, were also tagged less often in their own and friends' images. However, no clear correlation could be shown due to the relatively low number of answers. In case of own photos found by themselves (72 disliked): Spearman's $\rho_1 = \rho_2 = -0.3, p < 0.01$; and in case of own photos found by others (100 disliked): $\rho_1 = -0.22, \rho_2 = -0.38, p < 0.01$.

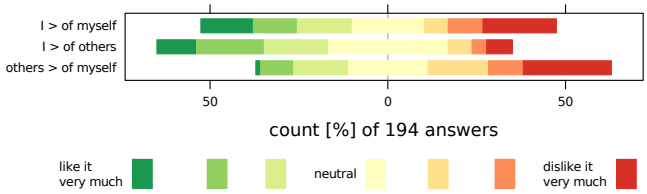


Fig. 14. How much do you like these uses of person tags? (Who finds photos showing whom?)

C. Reasons for Tagging People

We asked our participants to state the frequencies for specific reasons for tagging other people in photos using the 7-point scale from (1) *not at all* to (7) *very often* (cf. Figure 15). Answers on reasons differed significantly (Wilcoxon test, $Z = -3.11, p < 0.01$). 46 % of the 214 participants stated that they never tag people only to raise their awareness about a photo. The remaining 54 % rated this item with a mean value of 4 (sd = 1.8), while 11 % were in top two. Likewise, 54 % of the participants stated that they never tag people only to spread a photo someone is depicted in. The remaining 46 % rated this reason with a mean value of 3.6 (sd = 1.8), while 5 % were in top two. Our results about spreading are virtually identical to prior work, while considerably less participants indicated to tag for awareness reasons: 46 % of our participants answered with *not at all*, while only 30 % did so in prior work. The mean value of the remaining answers was 4 instead of 5.34.

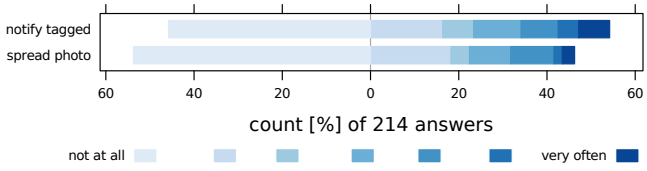


Fig. 15. How frequently do you tag for these reasons?

VII. Discussion

This section discusses the main findings from our study.

A. Demographics

The skewed gender distribution of our participants probably was caused by how we gathered participants. 93 % of them

were from the yellow press group. The readers of the yellow press paper "Bild" online news have a 60:40 male/female distribution. We assume that the readership was additionally skewed towards male readers, because the paper published our article in the IT news section. We could not find any significant difference between male and female participants for app results and estimations. Additionally, we could not identify differences when grouping participants by age or by the most often specified origin countries. One reason could be a potentially small influence of a person's gender, age or country on what that person's friends share or tag.

B. Amount of Photos and Metadata per User

Reports like "more than 350 million photos are uploaded every day on average" [18] regularly outline the amount of shared media. However, users of the social Web are not able to get an overview of the amount of photos in their personal environment. Our Facebook app presents a first step towards this need by raising awareness about the amount of photos posted by the circle of friends. Contrary to general reports about social services, such numbers might be more tangible for single users. The number of photos was often already enormous when counting only photos in the circle of direct friends, while services of the social Web allow access to personal data with an even higher degree of separation, like friends of friends or even public.

Figure 2 gives a short overview of the amount of data visible to single users. The average user of our app had access to 16 k photos; the photos had a median of over 5.5 k person tags, 1.7 k location tags and 21.9 k comments. This amount of data is already much higher than what most users could manually review.

When considering a possible privacy invasion by apps, the summary of our app dataset additionally illustrates how fast the amount of data given to apps grows. For instance, the relatively small set of 2753 app users allowed the app access to at least 75 million photos in which 6.3 million different people were tagged. Consequently, even such a small scale app distribution can invade the privacy of over 6 million people. Those statistics could be very valuable for privacy education.

Less than one third of all friends disallowed *apps others use* the access to their photos. The lion's share does not regard this restriction as necessary or might not know about this privacy setting at all. Considering the scale of the problem this is something worth working on.

C. Estimation and Unawareness

Many participants of our study were not aware of the amount of shared photos that might raise concerns to them as shown by the estimations summarized in Table IV. Most of our 2245 participants were not able to correctly estimate the amount of photos shared by their direct friends. 22.4 % of the participants stated that they have *no idea* about the amount of shared photos. 64.7 % underestimated the number of photos shared by their friends, whereas more than half of the underestimations (37.5 % oap.) were made with a factor higher than 10. Participants underestimated the amount of shared photos by notable factors. Even when allowing misestimations of the magnitude of 1, only 39 % of all participants did not make a substantial misestimation. We must conclude that most people are not aware of the amount of photos that might raise concerns to them. Additionally, we must keep in mind that our app only addressed photos of direct Facebook friends, and hence presented numbers form only lower bounds.

TABLE IV. SUMMARY: ESTIMATIONS VERSUS REAL VALUES

data	*na*	*no idea*	correct	underestimation	overestimation
photo count	0.5	22.4	6.3	64.7 (+3.9)	6.1 (+2.4)
w/locations	2.2	16.6	25.3	10 (+14.7)	45.9 (+10.6)
w/tags	2.8	13.9	19.4	14.4 (+10.5)	49.5 (+8.9)
avg. tags	4.0	18.0	30.6	7.2	40.2

na = *no answer*, misestimations in brackets are those regarded as correct

In case of tags more people were able to estimate correctly: 25.3 % of all participants estimated the portion of photos with location tag correctly and 19.4 % estimated correctly concerning the percentage of photos having person tags in it. Although percentages were lower, participants' answers seemed to be more informed in case of person tags when comparing medians of real values grouped by estimations. Similarly, 30.6 % of the participants did an exact estimation of the mean count of person tags in tagged photos. In cases of metadata, the main part of participants overestimated. While up to half of them overestimated, less than 15 % of all participants underestimated the amount of different metadata. Facebook users seem to have a better feeling for the amount of tags in photos shared by their friends than for the number of photos. Maybe the sample of photos shared on Timelines is sufficient to get a more representative notion of the amount of metadata usage. In comparison, users might not browse all friends' photo albums and even if their Timeline indicates that some friends are sharing more than others, this does not give a hint about the quantity of shared photos. From the privacy perspective, overestimation is better than underestimation, but it also testifies unawareness.

D. Inconsistencies

In the post-questionnaire we found some users answering consistently to the results gained from comparing estimations and real values, but we also had a notable portion of participants who answered inconsistently; for instance underestimating but stating result to be lower than expected. One reason for this might be time that passed between different parts of the study. The inconsistency could also well be another indicator for the absence of awareness. Since all three parts of our study were without any incentive but personal interest, we could dismiss deliberate misbehavior or purely random answers for most of our participants.

E. Users' Feelings

Our post-questionnaire results confirmed the general results of prior work [16] and exhibited an even worse situation in some cases. The users of our app stated to be worse informed about photos of themselves on the Web. The average perception of the effect of being tagged was virtually identical. Participants stated to use person tags less for awareness reasons and only few participants saw a benefit to privacy in being tagged. Therefore, our hypothesis about such benefit does not hold for our participants. Interestingly, while most participants

stated not to be upset about the amount of photos shown to them in their Facebook app results, nearly three quarter of them stated to be insufficiently informed about photos.

VIII. Conclusion

In this work we presented the results of a study we built to measure user awareness concerning photos shared on Facebook as well as corresponding metadata beyond basic self-reporting. The results of our study show that many participants were extremely unaware of the amount of shared photos and additional data. Their inability to estimate provides evidence for their lack of knowledge about the dimension of shared data and the potential threat to privacy. Our study proved a high level of user unawareness about shared media and analyzed the users' feelings towards this.

The results offer two valuable insights: Firstly, users are not aware of the extent of data shared by their online friends. Our results emphasize the need for privacy-enhancing technology that enables users to become aware of photos relevant to them. This goes beyond the current privacy settings and access control mechanisms since this effect is created by other peoples' media. We also believe that users need to become more aware of the scale of this issue, to create a basis for informed decision making. While most people have heard of the huge amounts of data shared globally, our Facebook app makes these numbers more personal.

Secondly, it is likely that users do not understand the impact of their own sharing habits either. While we cannot easily make people re-think their habits, we can support them with information and tools to realize the scope of their personal shared media, their habits, and consequences. This leads to significant challenges for privacy research. We must enable users to manage their own media, but also to become aware of and be able to influence media of other people that affects them, and also to become aware how their media affects others.

IX. Future Work

Motivated by the results of this work, we plan further studies on awareness and research on how technical facilities can be created to support user awareness about data and easing the handling of the vast amounts of media contained in each group of friends domain. The app we created for this study will be extended as a user-privacy centered browser for friends' photos. While we used location and person tags, distinct persons and places, captions and comments for dataset analysis, this data could also be used to browse and sort through photos. This will facilitate the browsing of friends' photos for manual privacy-related screening. Users could for instance easily access photos with cumulative metadata matches like simultaneous people and location tags, many comments plus being tagged, or best friends being tagged. After having obtained a base dataset of users, we will extend the presented statistics app to enable users to assess their own privacy in comparison to other users.

References

[1] S. D. Warren and L. D. Brandeis, "The right to privacy," *Harward Law Review*, vol. 4, no. 5, pp. 193–220, December 1890. [Online]. Available: http://groups.csail.mit.edu/mac/classes/6.805/articles/privacy/Privacy_brand_warr2.html

[2] A. Besmer and H. R. Lipford, "Privacy Perceptions of Photo Sharing in Facebook," in *Proceedings of the 4th Symposium on Usable Privacy and Security*, ser. SOUPS '08. ACM, 2008.

[3] S. Ahern, D. Eckles, N. Good, S. King, M. Naaman, and R. Nair, "Over-exposed? Privacy patterns and considerations in online and mobile photo sharing," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '07. ACM, 2007, pp. 357–366.

[4] A. Besmer and H. R. Lipford, "Moving beyond untagging: photo privacy in a tagged world," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. ACM, 2010, pp. 1563–1572.

[5] M. Johnson, S. Egelman, and S. M. Bellovin, "Facebook and privacy: it's complicated," in *Proceedings of the Eighth Symposium on Usable Privacy and Security*, ser. SOUPS '12. ACM, 2012, pp. 9:1–9:15.

[6] Y. Liu, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, "Analyzing Facebook privacy settings: User expectations vs. reality," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, ser. IMC '11. ACM, 2011, pp. 61–70.

[7] P. Klemperer, Y. Liang, M. Mazurek, M. Sleeper, B. Ur, L. Bauer, L. F. Cranor, N. Gupta, and M. Reiter, "Tag, you can see it!: using tags for access control in photo sharing," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12. ACM, 2012, pp. 377–386.

[8] A. Lampinen, V. Lehtinen, A. Lehmuskallio, and S. Tamminen, "We're in it together: interpersonal management of disclosure in social network services," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. ACM, 2011, pp. 3217–3226. [Online]. Available: http://doi.acm.org/10.1145/1978942.1979420

[9] S. Pötzsch, "Privacy awareness: A means to solve the privacy paradox?" in *The Future of Identity in the Information Society*, ser. IFIP Advances in Information and Communication Technology. Springer, 2009, vol. 298, pp. 226–236. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-03315-5_17

[10] A. Acquisti and R. Gross, "Imagined communities: Awareness, information sharing, and privacy on the facebook," in *Privacy Enhancing Technologies*, ser. LNCS. Springer, 2006, vol. 4258, pp. 36–58. [Online]. Available: http://dx.doi.org/10.1007/11957454_3

[11] D. Malandrino, A. Petta, V. Scarano, L. Serra, R. Spinelli, and B. Krishnamurthy, "Privacy awareness about information leakage: Who knows what about me?" in *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society*, ser. WPES '13. ACM, 2013, pp. 279–284. [Online]. Available: http://doi.acm.org/10.1145/2517840.2517868

[12] C. E. Wills and M. Zeljkovic, "A personalized approach to web privacy - awareness, attitudes and actions." *Information Management & Computer Security*, vol. 19, no. 1, pp. 53–73, 2011.

[13] J. Lin, S. Amini, J. I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang, "Expectation and purpose: Understanding users' mental models of mobile app privacy through crowdsourcing," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, ser. UbiComp '12. ACM, 2012, pp. 501–510. [Online]. Available: http://doi.acm.org/10.1145/2370216.2370290

[14] B. Konings, F. Schaub, and M. Weber, "Who, how, and why? enhancing privacy awareness in ubiquitous computing," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, 2013, pp. 364–367.

[15] T. Burghardt, A. Walter, E. Buchmann, and K. Bohm, "Primo - towards privacy aware image sharing," in *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, ser. WI-IAT '08. IEEE, 2008, pp. 21–24.

[16] B. Henne and M. Smith, "Awareness about photos on the web and if privacy-privacy-tradeoffs may help," in *Financial Cryptography and Data Security*, ser. LNCS, no. 7862. Springer, 2013.

[17] B. Henne, C. Szongott, and M. Smith, "Snapme if you can: Privacy threats of other peoples' geo-tagged media and what we can do about it," in *Proceedings of the Sixth ACM Conference on Security and Privacy in Wireless and Mobile Networks*, ser. WiSec '13. ACM, 2013.

[18] Facebook, Ericsson and Qualcomm, "A focus on efficiency," Internet.org, Whitepaper, September 2013.