Comparison-based Privacy: Nudging Privacy in Social Media (Position Paper)

Jan Henrik Ziegeldorf, Martin Henze, René Hummen, and Klaus Wehrle

Communication and Distributed Systems, RWTH Aachen University, Germany
{ziegeldorf, henze, hummen, wehrle}@comsys.rwth-aachen.de

Abstract. Social media continues to lead imprudent users into *over-sharing*, exposing them to various privacy threats. Recent research thus focusses on *nudging* the user into the 'right' direction. In this paper, we propose *Comparison-based Privacy (CbP)*, a design paradigm for privacy nudges that overcomes the limitations and challenges of existing approaches. *CbP* is based on the observation that comparison is a natural human behavior. With *CbP*, we transfer this observation to decision-making processes in the digital world by enabling the user to compare herself along privacy-relevant metrics to user-selected comparison groups. In doing so, our approach provides a framework for the integration of existing nudges under a self-adaptive, user-centric norm of privacy. Thus, we expect *CbP* not only to provide technical improvements, but to also increase user acceptance of privacy nudges. We also show how *CbP* can be implemented and present preliminary results.

Keywords: behavioral nudge; privacy; social media

1 Introduction

Over-sharing of personal information on social media has led to several privacy incidents: i) People have missed career opportunities [6], ii) embarrassed themselves [17], or iii) become victims of crimes [5]. In response, websites have implemented access and privacy controls. However, these protection mechanisms usually come with very lenient defaults and users often fail or simply neglect to set up individual settings [12,17]. Recent research explores the use of behavioral nudges to raise awareness about privacy risks and lead users to informed decisions about their social media privacy. These *privacy nudges* try to detect privacy sensitive contexts and warn users, e.g., a-priori to critical posts on Facebook [18].

The proposed nudges face two challenges inherent to their design. First, they require ground truth to detect sensitive content which is not available per se. Consequently, proposed systems focus only on very specific privacy threats, e.g., a nudge that warns about the disclosure of vacation plans based on laboriously hand-labelled data [13]. Second, the proposed systems convey only the subjective privacy norm of the person designing and training the system, i.e., privacy is defined in a *one-for-all* manner. Considering the importance of individual and

social aspects in privacy, it is not surprising that many users disagree and reject advise from nudges that dictate a one-for-all definition of privacy [10].

In this position paper, we propose *Comparison-based Privacy (CbP)*, a new best-effort approach for nudging privacy. *CbP* is motivated by the observation that comparisons are widely used by humans in their every-day lives to assess their own status, behavior, and decisions, and that such comparisons are also effective in influencing a person's behavior [7]. Therefore, we propose to support a user's decision making in privacy contexts by comparing her sharing behavior along different metrics (e.g., amount of shared content or usage patterns) to different comparison groups, which she can intuitively relate to (e.g., family, friends and colleagues, users with the same profession or same age). Because of its inherently relative nature, *CbP* neither assumes nor requires any fixed privacy norm or ground truth. Instead, a user is nudged completely based on the behavior of her peer groups. This also allows *CbP* to harmonize individual and social factors of privacy. Individual aspects are covered by the user's choice of comparison metrics, while social aspects are captured in the aggregated behavior of a specific comparison group. With this, *CbP* overcomes the restrictions of other privacy nudges and promises increased user acceptance, easier deployment and maintenance, and a certain degree of adaptivity to changing notions of privacy.

2 Problem Analysis and Related Work

The problem of over-sharing fundamentally stems from users' inability to responsibly decide how often to share which content with whom. Recently proposed privacy nudges tackle this problem by raising awareness about specific consequences of over-sharing. `PleaseRobMe`[1] addresses geo-location information and `FireMe!`[2] abusive language related to work. These systems are based on manually configured filter rules that allow to detect only very specific privacy risks. As improvement, [13] employs supervised machine learning to detect sensitive tweets and [9] automatically annotates text-based social media content with privacy labels. However, due to the apparent lack of ground truth to train these systems, only a small set of less than 1 000 hand-labeled tweets [13] or synthetic data [9] is used. [11] proposes a privacy score based on the sensitivity of profile items, which was exemplarily determined through a user study. These approaches show that providing ground truth on the sensitivity of social media content currently requires substantial manual effort. *CbP* avoids these efforts by basing nudging decisions solely on comparisons between a user and her peer groups.

A second challenge common to related work [10, 13, 18] is that the norm of privacy is dictated during system development and is immutable from there on. However, privacy is both an individual and social concept that cannot be defined in a one-for-all manner. First, individual factors such as a user's demographics, profession, or personal preferences play an important role. Second, privacy decisions are also shaped by the perception and appreciation of privacy in the user's

[1] `http://pleaserobme.com/`
[2] `http://fireme.l3s.uni-hannover.de/`

social environment. Negligence of these factors leads to non-acceptance among users: While [13] does not investigate user acceptance, users tend to reject the nudged advise of [10, 18] as they do not feel addressed individually. In contrast, *CbP* proposes nudging users in a self-adaptive, user-centric way.

Finally, a long line of research investigates on how to learn and configure users' access and sharing policies: [1] (semi-) automatically learns a user's group memberships and [4] automatically assigns privileges to a user's friends based on a limited amount of user input and settings of other users. Other approaches focus on predicting location sharing preferences [15,16]. [14] proposes to let users collaboratively manage access control to social media data. Our work is orthogonal as it engages the user one step earlier: We aim at nudging users towards treating their digital privacy more consciously, which could, e.g., lead the user to customize privacy preferences using one of the above approaches. However, our proposed *CbP* paradigms draws and extends on the idea of collaboratively managing privacy that is present in some of the discussed approaches.

3 Comparison-based Privacy

To enable self-adaptive, user-centric privacy nudges, we make the following three observations. First, comparison is a natural human behavior. People compare themselves to their peer groups everyday based on a wide set of criteria ranging from salary to health. Second, comparison does not require ground truth or training data. Instead, self-reflection and decision making is rather guided by relative values. The aggregated behavior of the peer group dynamically provides inidividual 'ground truth' for people to evaluate their own decisions. Third, people usually compare not to random strangers. They compare to people from their social environment who they can individually relate to, e.g., people with the same profession, age, or other demographics. In doing so, they harmonize individual and social factors that influence their decision-making process.

Based on these observations, we argue that comparing privacy relevant aspects of a user's social media activity allows her to intuitively understand and assess her privacy risk. Specifically, we propose to compare a user's sharing behavior along a number of *comparison metrics* to user-specific *comparison groups*. We refer to this novel approach as *Comparison-based Privacy (CbP)*. Notably, our approach renounces any fixed norm of privacy and fully embraces privacy as both an individual and a social concept. We now discuss our comparison metrics and groups and their combination, while deferring technical details to §4.

Comparison Metrics: Comparison metrics capture privacy-critical aspects of a user's sharing behavior. They are motivated from an analysis of the consequences of over-sharing on social media. Related work already proposed a wide range of such metrics: It has been recognized that employers and credit scorers look at *linguistic features* of applicants [8], e.g., correctness of grammar and spelling or abusive language. Other threats, e.g., stalking and cybercasing, exploit certain *content types* such as geo-location or pictures [5]. Embarrassment or loss of career opportunities often emerge from talk about *sensitive topics* such

as drug abuse or disease [6]. Finally, hints for mental diseases such as depression can be detected in users content [2]. It is important to note that our *CbP* approach neither obsoletes these related works nor is limited by it. Instead, *CbP* provides a unifying and extensible framework to integrate existing approaches as comparison metrics or devise new ones. We can, e.g., integrate as comparison metrics Kawase's job hater filter [10], Wang's nudge based on expressed sentiment [18], or Mao's disease and drunkenness classifiers [13]. The application of the *CbP* paradigm thereby transforms their fixed norm of privacy into a relative, comparison-based notion, thus increasing the acceptance among users.

Comparison Groups: Comparison groups allow a user to adapt *CbP*-based nudges to her specific norm of privacy. Hence, a user should select groups that she has an intuitive relation to. Social media sites already provide inherent structures and information, e.g., social graphs, profiles, lists of friends/followers, that provide such comparison groups and require no configuration at all. Besides these preexisting comparison groups, we can automatically build comparison groups based on user profile information, e.g., age, profession, interests and hobbies or even religion and political orientation, to provide an even more individualized nudging experience. Since not all users share this information publicly, comparisons for these groups would potentially be restricted to users of our system.

Nudging the User: The user chooses the desired comparison metrics and comparison groups individually, e.g., *"compare the amount of abusive language to people of the same profession"*. This allows the user to individualize the used norm of privacy. Our *CbP* approach then evaluates each chosen metric on the target user and builds an aggregate (e.g., average or median) over each chosen comparison group. The aggregates serve as empirical ground truth *relative* to how the social environment behaves. Social aspects of privacy are thus factored in to the nudging decision. A particular comparison between one user and the group aggregate can result in the three different cases as depicted in Fig. 1. In the first case, the user and the group behave in similar ways, i.e., the target user's result is close to the group's aggregate. This information confirms the user in her behavior with respect to this group. If a particular comparison exceeds a threshold in either direction (Cases 2 and 3 in Fig. 1), a *CbP*-based nudge would alert the user to this fact. The nudge would, e.g., alert her that the amount of abusive language in her posts exceeds the average in her peer groups. Thresholds can be set individually by users or according to general profiles representing typical privacy attitudes of an unconcerned, critical, or very anxious user. It is a desired feature of our system that a user's behavior is evaluated only in relation to her peer groups, even if results may vary or contradict each other across different groups. Such personalized appeals have proven to be more effective than judging behavior by a fixed norm of 'good' and 'bad' [7].

4 Proposed System Design

We now describe a system architecture that leverages *CbP* to nudge social media users towards more privacy-conscious sharing decisions. As illustrated in

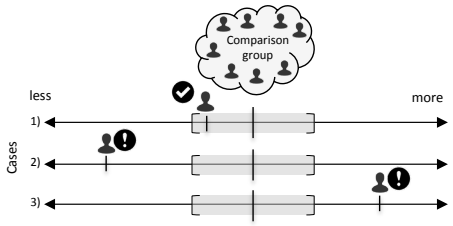

Fig. 1. Comparing a user against the aggregate group behavior.

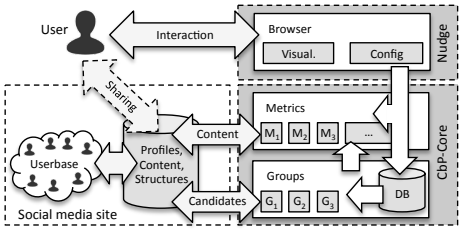

Fig. 2. Overview of the system components and their interaction.

Fig. 2, the system has two main components: (i) the *CbP*-core and (ii) the actual privacy nudge. The privacy nudge handles all interactions with the user. These interactions primarily include the initial discovery or configuration of the comparison groups and the comparison metrics as well as the actual nudging based on the comparison results obtained from the *CbP*-core. The *CbP*-core is a stand-alone application. It manages the comparison groups and implements comparison metrics. The user must grant it sufficient rights to query the social media site for the user's content to construct groups and evaluate the metrics. We now describe the details of the two *CbP* components and their interactions. Then, we discuss privacy implications of this design and possible alternatives.

Nudge: The nudge runs in the user's browser and represents the user interface. It asks the user to sign in and grant permission to access her social media accounts. To keep configuration efforts to a minimum, the user is presented with a pre-configured selection of comparison groups and metrics, but may refine this choice by filling in additional information. The nudge module triggers the *CbP*-core and then receives the results which it uses to nudge the user. Effective ways of actually presenting such nudging advise to a user is a question orthogonal to our approach and subject to ongoing research, e.g., [18] proposes to alter the control flow by delaying posts and [10] prods the user to delete certain content.

CbP-Core: The *CbP*-core contains the functionality to realize *CbP*, i.e., a groups module that builds and manages the comparison groups and a metrics module that implements the different comparisons (cf. §3). The *groups module* draws on standard structures of the social media site to build basic groups, e.g., the social graph or friend lists. If granted sufficient permissions, the *CbP*-core also accesses the user's protected profile information to build more specific groups. Further personal information that the user may supply during configuration, e.g., profession or age, is used to build more sophisticated groups. The *metrics module* takes a comparison group and evaluates the desired comparison metrics for each group member. Basic implementations of the metrics described in §3 can be realized using simple content filters based on word lists, e.g., for abusive language or sensitive topics, or by quantifying the amount of shared content, e.g., number of shared geo-locations. Quantifying how often similar content has been shared in the comparison groups additionally provides an indication of the sensitivity of the shared content. More comparison metrics can be built using

publicly available APIs, e.g., for sentiment analysis, and through the integration of related work. Finally, the metrics module provides the aggregated results of the comparison group and the result for the particular user to the nudge module.

4.1 Discussion

Any entity, i.e., the operator of the nudge system or other users in the comparison groups, may try to spy on or actively attack the nudged user. We thus discuss how to establish trust in our system and prevent information leakage and coercion.

Trust: In our proposed design, the *CbP*-core runs as a stand-alone third-party application, as this is the easiest deployment option. However, this requires the user to trust the *CbP*-core and grant it access to her social media content. We identify two alternatives to this approach: First, the site operator itself could run the *CbP*-core or provide a suitable query interface that allows evaluation of the comparison metrics without explicitly accessing the user's contents. The second alternative is to run the *CbP*-core on the user side, e.g., as a browser plugin, and collect only the aggregate results of the comparison groups centrally. Both alternatives would not require the user to trust an additional entity.

Information leakage: In all previously mentioned deployment scenarios, the user learns the results of the comparisons aggregated over the chosen comparison groups. However, this might be sensitive information, e.g., a malicious user may learn private information about outliers by choosing artificially small comparison groups. A trivial protection mechanism would be to only allow groups of a certain minimum size. To achieve rigorous privacy guarantees, we propose to apply Differential Privacy [3] to the aggregated outcome of the comparison.

Coercion: The aggregated behavior of a comparison group may unintentionally move into a harmful direction or a an attacker may try to manipulate it to steer a user's privacy decisions into a particular direction. We argue that a user can counter such attacks by choosing multiple, diverse, and sufficiently large comparison groups or even known reference groups, e.g., comprising the national data protectionists. As a second protection mechanism, extreme outliers, e.g., results contributed by an attacker who wants to manipulate the aggregate group behaviour, could be filtered out by the *CbP*-core component.

5 Preliminary Results

We demonstrate the feasibility of our approach by the example of Twitter. Information on Twitter is mostly public, which has lead to the many privacy violations [13], but also makes Twitter and its users a prime target for our proposed privacy nudge. Although other OSNs may enforce stricter access control on shared data and attract different categories of users, qualitatively similar privacy violations have been reported for them, e.g., for Facebook [6, 17]. Thus, we expect that our obtained results also generalize to other OSNs. We collected half a million tweets of 1 839 active Twitter users in four comparison groups by profession: teachers (659), nurses (542), journalists (559), and U.S. senators (79).

Groups were obtained through the Twitter Search API and scraped from public lists. We evaluate a choice of comparison metrics from §3 for each group.

The *location disclosure* metric measures the percentage of a user's tweets tagged with a geo-location. We find that all groups are very restrictive about location disclosure. Specifically, well above 90% of the users disclose their location in less than 7.8% of their tweets. Nearly all of them do not disclose their location at all. This result shows a wide consensus among Twitter users concerning location disclosure. This fact would immediately become apparent through the use of *CbP*. Unaware users (we observe outliers among the nurses and teachers) could therefore better assess their privacy risks with *CbP*. Results are less homogeneous for *abusive language*, defined as the percentage of tweets containing expressions regarded as offensive. Journalists and politicians use very little abusive language, while nurses and teachers show considerable use of it. Hence, it appears that some amount of abusive language is tolerable in particular groups. This confirms our relative norm of privacy and the need for user-specific comparison groups. *CbP* captures this fact and, e.g., would rather nudge the politician than the nurse. The *sensitive topics* metric measures the percentage of tweets containing references to work, diseases or drug abuse. We find that these comparisons are less useful as such topics are also referenced in many privacy irrelevant contexts. Using *sentiment analysis* on tweets as comparison metric, again shows the importance of nudging users individually with respect to their social environment. While nurses and teachers tweet with rather neutral sentiment, senators are clearly more upbeat. Surprisingly, journalists commonly display a negative mood, part of which relates to reports about crimes and disasters.

We additionally scraped the top 300 job haters from the `FireMe!` site and used it as a contrast group. Those users are endangered of job loss and our system should detect and warn against this privacy risk. Indeed, job haters spike for all our metrics, i.e., disclosing more locations than others, having significantly higher rates of abusive language, and tweeting with clearly more negative sentiment. While our system could not directly point them to the risk of losing their job, it would still nudge them away from their harmful sharing behavior by pointing out their discrepancy with social norms established from the comparison groups.

6 Outlook and Conclusion

We are developing our proposed system for Twitter and Facebook to answer practical questions, e.g., how stable comparison results are. We also investigate further comparison metrics and groups as those briefly mentioned in §3. Finally, we intend to conduct a user study based on our developed system to answer non-technical questions: Our system may issue possibly contradicting advise, how do users respond to this? Usability is a major design goal; how much configuration is really necessary for inexperienced users?

To conclude, *CbP* presents a novel paradigm for nudging users in a best-effort manner towards more informed privacy decisions – an important challenge due to the increasing proliferation of social media among young and inexperienced

users. Our *CbP* approach promises to overcome the restrictions of related work by employing a relative norm of privacy that considers both individual and social factors and does not require training data or preconfigured rules. The preliminary results show that our *CbP* paradigm indeed has the potential to effectively nudge social media users towards more privacy conscious sharing decisions.

Acknowledgements. This work has been funded by the Excellence Initiative of the German federal and state governments.

References

1. Amershi, S., Fogarty, J., Weld, D.: Regroup: Interactive machine learning for on-demand group creation in social networks. In: CHI'12. ACM (2012)
2. De Choudhury, M., Counts, S., Horvitz, E.: Social media as a measurement tool of depression in populations. In: Web-Sci'13. ACM (2013)
3. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Halevi, S., Rabin, T. (eds.) Theory of Cryptography, LNCS, vol. 3876, pp. 265–284. Springer (2006)
4. Fang, L., LeFevre, K.: Privacy wizards for social networking sites. In: WWW'10. ACM (2010)
5. Friedland, G., Sommer, R.: Cybercasing the joint: on the privacy implications of geo-tagging. In: HotSec'10. USENIX (2010)
6. Garone, E.: Can social media get you fired? `http://www.bbc.com/capital/story/20130626-can-social-media-get-you-fired` [Accessed: 2015-07-10] (2013)
7. Goldstein, N.J., Cialdini, R.B.: A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. JCR (2008)
8. Huffington Post: 37 Percent Of Employers Use Facebook To Pre-Screen Applicants, New Study Says. `http://huff.to/1c5fvQg` [Accessed: 2016-07-10] (2012)
9. Jakob, M., Moler, Z., Pěchouček, M., Vaculín, R.: Content-based privacy management on the social web. In: WI-IAT'11. IEEE (2011)
10. Kawase, R., et al.: Who wants to get fired? In: WebSci'13. ACM (2013)
11. Liu, K., Terzi, E.: A framework for computing the privacy scores of users in online social networks. TKDD (2010)
12. Liu, Y., Gummadi, K.P., Krishnamurthy, B., Mislove, A.: Analyzing facebook privacy settings: user expectations vs. reality. In: IMC'11. ACM (2011)
13. Mao, H., Shuai, X., Kapadia, A.: Loose tweets: an analysis of privacy leaks on twitter. In: WPES'11. ACM (2011)
14. Squicciarini, A.C., Shehab, M., Paci, F.: Collective privacy management in social networks. In: WWW'09. ACM (2009)
15. Toch, E.: Crowdsourcing privacy preferences in context-aware applications. Personal and ubiquitous computing (2014)
16. Toch, E., Cranshaw, J., Drielsma, P.H., Tsai, et al.: Empirical models of privacy in location sharing. In: UbiComp'10. pp. 129–138. ACM (2010)
17. Wang, Y., et al.: "I regretted the minute I pressed share": a qualitative study of regrets on Facebook. In: SOUPS'11. ACM (2011)
18. Wang, Y., et al.: Privacy nudges for social media: an exploratory Facebook study. In: WWW '13. IW3C2 (2013)