Personality-based Knowledge Extraction for Privacy-preserving Data Analysis

Xuan-Son Vu [†], Lili Jiang[†], Anders Brändström[§], Erik Elmroth[†]

[†]Department of Computing Science, Umeå University, Sweden
sonvx@cs.umu.se,lili.jiang@cs.umu.se,elmroth@cs.umu.se
[§]Demographic and Aging Research Centre, Umeå University, Sweden
anders.brandstrom@umu.se

ABSTRACT

In this paper, we present a differential privacy preserving approach, which extracts personality-based knowledge to serve privacy guarantee data analysis on personal sensitive data. Based on the approach, we further implement an end-to-end privacy guarantee system, KaPPA, to provide researchers iterative data analysis on sensitive data. The key challenge for differential privacy is determining a reasonable amount of privacy budget to balance privacy preserving and data utility. Most of the previous work applies unified privacy budget to all individual data, which leads to insufficient privacy protection for some individuals while over-protecting others. In KaPPA, the proposed personality-based privacy preserving approach automatically calculates privacy budget for each individual. Our experimental evaluations show a significant trade-off of sufficient privacy protection and data utility.

KEYWORDS

Differential Privacy, Privacy-preserving Data Analysis

ACM Reference Format:
Xuan-Son Vu [†], Lili Jiang[†], Anders Brändström[§], Erik Elmroth[†]. 2017. Personality-based Knowledge Extraction for Privacy-preserving Data Analysis. In *K-CAP 2017: K-CAP 2017: Knowledge Capture Conference, December 4–6, 2017, Austin, TX, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3148011.3154479

1 INTRODUCTION

Privacy issues are getting more attention due to the growing volume of personal data such as social network data, demographic, and health data. Cross-disciplinary studies have been conducted with the need of integrating these personal data from multiple sources. This data integration dramatically increases the risk of privacy leakage. For example, Narayanan et al. de-anonymized the published Netflix Prize data by matching to IMDB [11]. Moreover, the registration procedure to access personal data even for research purpose is time-consuming in terms of privacy guarantee and ethical reviews [1]. Therefore, the main goal of this paper is to propose a differential privacy preserving approach by using a self-adaptive privacy concern detection algorithm and present a system for interactive privacy-preserving data analysis.

The rest of this paper is organized as follows. Section 2 presents related work and introduces to differential privacy. The KaPPA framework is shown in Section 3. Section 4 presents the proposed methodology. Experiments and results are discussed in Section 5. Section 6 shows conclusions followed by future work.

2 RELATED WORK

There has been privacy protection work on anonymization [2] and sanitization [15]. **Differential Privacy** later emerged as the key privacy guarantee by providing rigorous, statistical guarantees against any inference from an adversary [3]. Based on differential privacy, some privacy-oriented frameworks arose including PINQ [9] and GUPT [10]. However, they only provide a library for technical people to create their own privacy guaranteed data analysis tool. Moreover, they apply a unified amount of noise for privacy protection. To our knowledge, there does not exist any end-to-end privacy guaranteed system for sensitive data analysis. To resolve these limitations, we propose a personality-based differential privacy approach and implement an end-to-end system called KaPPA to guarantee privacy on data analysis. Notably, we propose a personality-based differential privacy approach to calculate privacy concern for reasonable personalized privacy budget.
Personality Profiling is the task of predicting user personality traits based on user-generated contents (i.e., Facebook). It can facilitate various personalized intelligent applications including recommender system [6], mental health diagnosis [14], recruitment and career counseling [5]. In Section 4, we will discuss on personality and privacy-concern in detail.

2.1 Differential Privacy Preliminaries

Differential privacy (DP) [3] has established itself as a strong standard. The key idea behind differential privacy is to obfuscate an individual's properties, but not the whole group's properties in a database. So the probability for any individual in the database to have a property should barely differ from the base rate. Then, when an attacker analyzes the database, he/she cannot reliably learn anything new about any individual in the database, no matter how much additional information he/she has. Here is the formal definition of $\epsilon - \delta$ differential privacy. We assume a database D consisting

K-CAP 2017, December 4–6, 2017, Austin, TX, USA

ACM ISBN 978-1-4503-5553-7/17/12…$15.00
https://doi.org/10.1145/3148011.3154479

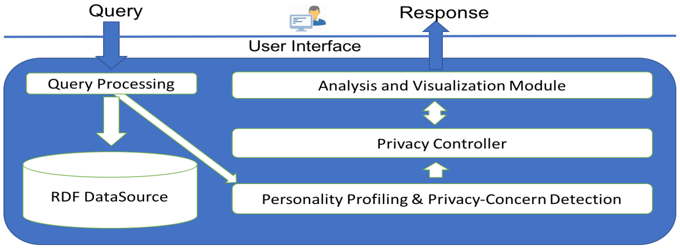

Query
Response
User Interface
Query Processing
Analysis and Visualization Module
Privacy Controller
RDF DataSource
Personality Profiling & Privacy-Concern Detection

Figure 1: KaPPA's System Architecture.

of n vectors of m-components over some set $\mathcal{F}$ represented as a $m \times n$ matrix over $\mathcal{F}$.

Definition 2.1. Define $\text{dist}(D, D') := \left|\{i \in \{1, 2, \dots, m\} \colon D_i \neq D'_i\}\right|$ $\forall D,\ D' \in (\mathcal{F}^m)^n$ as the number of entries in which the databases D and D' differ.

Definition 2.2. Let $\mathcal{A}$ be an algorithm processing D and Range($\mathcal{A}$) its image. Now $\mathcal{A}$ is called ϵ-δ-differentially private if $\forall \mathcal{S} \subset \text{Range}(\mathcal{A})$:

$$\forall D' \colon \text{dist}(D, D') \leq 1 \Rightarrow \Pr\left[\mathcal{A}(D) \in \mathcal{S}\right] \leq e^{\epsilon} \cdot \Pr\left[\mathcal{A}(D') \in \mathcal{S}\right] + \delta$$

Intuitively, differential privacy controls the degree to which D and D' can be distinguished. When $\delta = 0$ then ϵ-δ-differential privacy is also called ϵ-differential privacy. Smaller ϵ gives more privacy and worse utility. Then, given the result of a randomized algorithm $\mathcal{A}$, an attacker cannot learn any new property about data subjects with a significant probability.

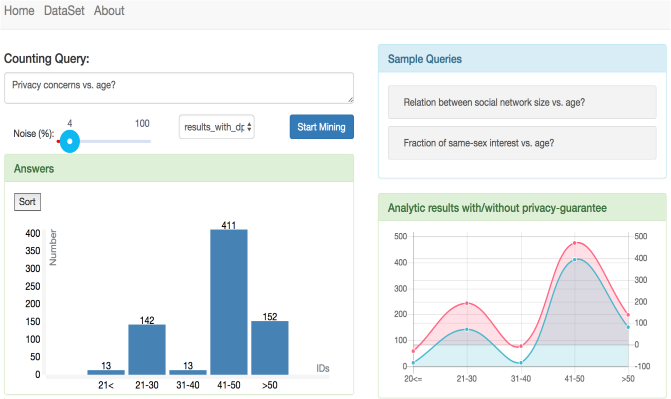
2.2 The Global Privacy Budget

PINQ [9] is an implementation of interactive differential privacy which ensures, at runtime, that queries adhere to a global privacy budget ϵ. Its central principle is that multiple queries (e.g. with differential privacy ϵ_1 and ϵ_2 respectively) have an additive effect $\epsilon_1 + \epsilon_2$ on the overall differential privacy. PINQ also tracks sensitivity of functions to track how much to deduct from the global privacy budget on each invocation of a primitive query. As mentioned in [4], the global privacy budget has limitations when it applies to an interactive system: (1) data analysts, who use the system, can run out of privacy budget even before finding useful results and (2) a global budget is not capable of handling a live database when new records are updated frequently.

3 PRIVACY-GUARANTEE DATA ANALYSIS

3.1 KaPPA's Architecture

KaPPA as depicted in Figure 1 consists of four modules: *Query Processing, Personality Profiling & Privacy-Concern Detection, Privacy Controller, and Analysis and Visualization Module.* Given a user query and privacy customization, *Query Processing* parses user query and transforms to a SPARQL query[1]. *Personality Profiling & Privacy-Concern Detection* is designed to automatically learn and predict privacy budget for each individual record using our personality prediction algorithm. *Privacy Controller* applies differential privacy algorithm to calculate the privacy budget based on personality trait scores. Based on the transformed SPARQL query and calculated privacy budget, *Privacy Controller* only selects individual records having higher privacy budget than the privacy compensation of the given query. Afterwards, *Analysis and Visualization Module* receives data records and further generates numeral statistical results passing to show the answer in statistics (i.e., histogram).



Counting Query: Privacy concerns vs. age?

Noise (%): 4 100

results_with_dp

Relation between social network size vs. age?

Fraction of same-sex interest vs. age?

Figure 2: User Interface and Interaction on KaPPA

For *Query Processing*, we construct our own Natural2Sparql process engine to convert a natural query to the SPARQL query using regular expressions. For *Personality Profiling&Privacy-Concern Detection*, we build knowledge graph for personality profiling and privacy concern prediction based on personality. As our main contribution components, *Personality Profiling&Privacy-Concern Detection* and *Privacy Controller* will be especially explained in the next Section.

3.2 KaPPA's Demonstration

Figure 2 demonstrates a running example of KaPPA. A user (i.e., a data analytic researcher) issues a counting query "Relation between privacy concerns and age?" with a customized noise-level.

When the researcher prefers a better privacy-protection for their analytic results, he/she will inject more noise to guarantee higher privacy, although it will reduce data utility. Given the request, KaPPA presents answer as a histogram on the bottom left. On the bottom right, we show a comparison analytic results of the answer with privacy-guarantee and without privacy-guarantee. In the real running system, KaPPA does not present original results to users, though we present them here for demonstration purpose to show the scientific process. To the top right panel of the interface, we provide a list of sample queries. In the current version, we only support the histogram query based data analysis. However, the system is capable of supporting many different data analysis queries such as clustering, mean, variance, principal component analysis.

Intuitively, KaPPA enables users especially researchers to interactively work on sensitive data for statistical studies. Moreover, their registration time to access the authorized personal data is reduced compared with regular application procedure (see [1]).

4 PERSONALITY-BASED DIFFERENTIAL PRIVACY PRESERVING APPROACH

As mentioned above, one limitation of differential privacy is the unified privacy budget on the same dataset for all individuals in data. To address this limitation, we propose a personality-based

[1] https://www.w3.org/TR/rdf-sparql-query/

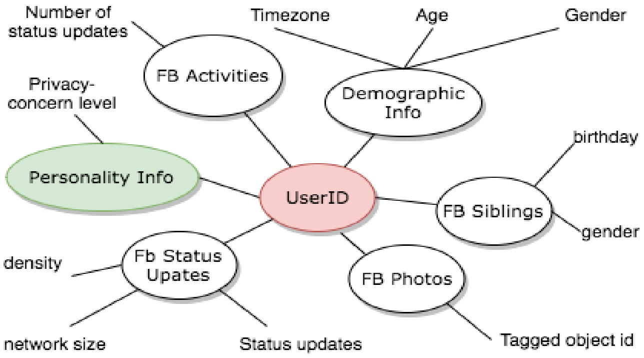


Figure 3: Knowledge graph for the personality-based privacy concern detection

Table 1: A simple statistic of the constructed knowledge graph. FB Siblings is the potential family users getting from FB Social Networks data.

Dataset	Number of records	Overlapping
Demographic Info	42,82,857	4,282,857 (100%)
FB Siblings	218074	217,936 (99%)
FB Status Updates	22,043,394	22,043,394 (100%)
FB Activities	1,674,260	1,622,368 (96.9%)
FB Photos	77,269,236	20,893,365 (27.04%)

differential privacy algorithm in *Privacy Controller*. This work is motivated by the findings from [13], which detected the significant correlation between personality and privacy concerns of individuals on Facebook. Moreover, this personality-based privacy can be characterized as personalized-differential privacy that is also satisfied ϵ-differential privacy by the proof of Ebadi et al. [4]. However, Ebadi et al. did not have an automatic way of detecting personalized privacy-concern level, which we are addressing. Below, we first introduce the most influential personality categorization model.

The Five Factor Model has become a standard model in psychology over the last 50 years [8].The five factors are defined as extraversion, neuroticism, agreeableness, conscientiousness, and openness to experience. Pennebaker et al., [12] identify many linguistic features associated with each of personality traits in *FFM*. (1) Extroversion (cEXT) tends to seek stimulation in the external world, the company of others, and to express positive emotions. (2) Neurotics (cNEU) people use more 1st person singular pronouns, more negative emotion words than positive emotion words. (3) Agreeable (cAGR) people express more positive and fewer negative emotions. Moreover, they use relatively fewer articles. (4) Conscientious (cCON) people avoid negations, negative emotion words and words reflecting discrepancies (e.g., should and would). (5) Openness to experience (cOPN) people prefer longer words and tentative expressions (e.g., perhaps and maybe), and reduce the usage of 1st person singular pronouns and present tense forms.

Algorithm	Majority	Naive Bayes	SVM
Accuracy	0.78	0.57	**0.80**

Table 2: Privacy concern detection performance in comparison with majority accuracy.

Personality-based Differential Privacy. Taking the five factor model as classification labels, we first extract personality features and build a knowledge graph with RDF representation, based on which we predict personality using SVM classifier. Figure 3 shows the skeleton of the knowledge graph built based on five datasets including Demographic data, FB Likes, FB Activities, FB Photos, and Facebook Status Updates with personality labels from myPersonality project [7]. It is worth to mention that the *Personality Info* and its property - i.e., privacy-concern level, are learned and predicted by the *Personality Profiling & Privacy-Concern Detection* module. Table 1 shows a simple statistics regarding the overlapping information between different datasets in the myPersonality project. Next, we build a machine learning model to automatically classify user privacy concerns to high, medium, and low level using their five personality trait scores and status updates. After applying the above model to determine privacy budgets of individuals, for each query, we subtract from the budget a corresponding amount of ϵ. When privacy-budget of an individual is used up, his/her data will no longer be included in the search results. Figure 4 presents a simple example to show the functionality of *Privacy Controller*.

5 EXPERIMENTS AND EVALUATION

We run two experimental studies to evaluate (1) the performance of privacy concern detection and (2) the trade-off of personality-based privacy preserving and data utility in data analysis.

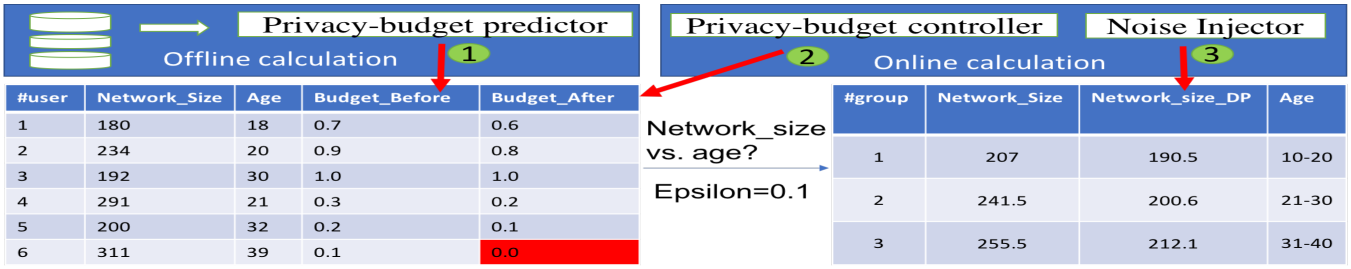
5.1 Dataset

We evaluate our approach on a subset of the *FB Status Updates* data in the myPersonality dataset [2]. It contains 9,917 Facebook statuses of 250 users in raw text, gold standard (self-assessed) personality labels, and several social network measures. It is a sample of personality scores and Facebook profile data collected by myPersonality project [7] using a Facebook application. The application obtained the consent from its users to record their data and use it for the research purposes. The status updates have been manually anonymized.

5.2 Gold Standard Labels and Evaluation

Ideally, we would evaluate downstream performance compared to a ground truth. Unfortunately, a ground truth is difficult to characterize for the privacy-concern task since people would have answered “as high as possible” if someone simply asked them “how much privacy-guarantee do you want to have?”. Therefore, we constructed our own labels, using all available information about users, and we use them as an approximation of the ground truth. We constructed these labels in order to evaluate downstream classification performance and they cover a set of users in three cases of privacy-concern levels, i.e., having high (HiPC), medium (MePC), and low privacy-concern (LoPC) level as proposed in [4]. Given the ground truth of personality labels, i.e., **yes** and **no** labels. Based on the findings of [13] we know that privacy concerns of different personality traits are ordered as following cNEU, cOPN, cCON, cAGR, cEXT from the highest privacy-concern to the lowest privacy-concern correspondingly. Therefore, we select users who belong to **yes** of (cNEU, cOPN) but **no** of (cAGR, cEXT) and put into the HiPC class. Conversely, users belong to **no** of (cNEU, cOPN) but **yes** of (cAGR and cEXT) will be put into the LoPC class. Remaining users are put

[2] http://myPersonality.com



#user	Network_Size	Age	Budget_Before	Budget_After
1	180	18	0.7	0.6
2	234	20	0.9	0.8
3	192	30	1.0	1.0
4	291	21	0.3	0.2
5	200	32	0.2	0.1
6	311	39	0.1	0.0

#group	Network_Size	Network_size_DP	Age
1	207	190.5	10-20
2	241.5	200.6	21-30
3	255.5	212.1	31-40

Figure 4: Example of Privacy-guarantee Data Analysis in KaPPA. Based on user personality, *privacy-budget predictor* ① decides privacy budget for each user data record. Since each query comes with a designed ϵ value, *privacy-budget controller* ② subtracts the same ϵ value from individuals' privacy budget. Finally, before releasing histogram statistic to the UI, *noise injector* ③ injects the noise to the histogram result.

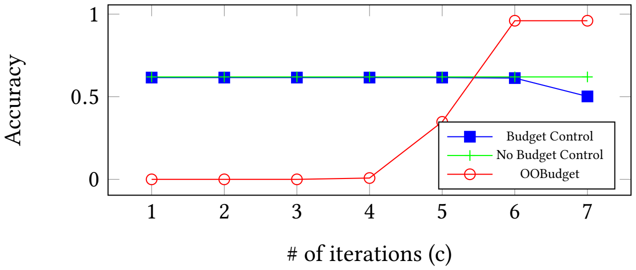


Figure 5: Evaluating the effect of privacy-budget control to the binary classification performance of the cEXT class. OOBudget is the ratio of out-of-budget user records.

into the MePC class. Eventually, our ground truth set consists of 29 users in HiPC, 212 users in MePC, and 9 users in LoPC.

5.3 Results and Discussion

Privacy-Concern Detection. Using the above ground truth data, we built two different privacy-classifiers with Naive Bayes and Support Vector Machine (SVM) algorithms. Table 2 shows the performance of privacy-concern detection in comparison with the majority accuracy. Clearly, due to the imbalance of class distribution, Naive Bayes does not perform well. SVM, however, can get a better result compared to the majority accuracy.

Privacy-budget Controller. We design a learning task with the privacy-budget controller to see how does it affect to the classification performance. A 10-fold SVM classification is designed to interactively request valid user records until it receives no records. Thus, this test is similar to a real scenario when an analyst requests to the system and retrieves information. Figure 5 shows that the accuracy goes down when the ratio of out-of-budget (OOBudget) increases. As in the global privacy budget method, when a data analyst uses up his/her privacy-budget, they will no longer get any search results. However, with personality-based privacy budget, the performance slowly goes down as the privacy budgets are embedded in user records to meet user-personalized privacy requirements.

6 CONCLUSION

This paper presents a differential privacy preserving approach and a privacy guaranteed system for data analysis. To address the limitation of unified privacy budget in differential privacy, we calculate privacy budget based on personality knowledge. The introduced system is a preliminary version, but supports natural query, adjustable query-based privacy guarantee configuration, and the returned results are privacy guaranteed by using the proposed privacy preserving approach. The system shows the potential to simplify the registration procedure to access personal sensitive data.

ACKNOWLEDGMENTS

This research is supported by Umeå University on federated database research. The authors also thank Michal Kosinski, David Stillwell and the myPersonality project for sharing data.

REFERENCES

[1] Australian National Data Service (ANDS). 2017. Application process to research on sensitive data with Ethics and Consent. (2017). ANDS's application process and ANDS's ethics and consent.

[2] R. J. Bayardo and Rakesh Agrawal. 2005. Data privacy through optimal k-anonymization. In *ICDE*. 217–228.

[3] Dwork Cynthia. 2006. Differential Privacy *(ICALP)*. 1–12.

[4] Hamid Ebadi, David Sands, and Gerardo Schneider. 2015. Differential Privacy: Now It's Getting Personal. In *Proceedings of the 42Nd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '15)*. 69–81.

[5] William L. Gardner, Brian J. Reithel, Claudia C. Cogliser, Fred O. Walumbwa, and Richard T. Foley. 2012. Matching Personality and Organizational Culture. *Management Communication Quarterly* 26, 4 (2012), 585–622.

[6] Rong Hu and Pearl Pu. 2011. Enhancing Collaborative Filtering Systems with Personality Information. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. 197–204.

[7] M. Kosinski, SC Matz, SD Gosling, V. Popov, and D. Stillwell. 2015. Facebook as a Social Science Research Tool. *American Psychologist* (2015).

[8] François Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *J. Artif. Int. Res.* (2007), 457–500.

[9] Frank D McSherry. 2009. Privacy Integrated Queries: An Extensible Platform for Privacy-preserving Data Analysis. In *SIGMOD*.

[10] Prashanth Mohan, Abhradeep Thakurta, Elaine Shi, Dawn Song, and David Culler. 2012. GUPT: Privacy Preserving Data Analysis Made Easy. In *SIGMOD*.

[11] A. Narayanan and V. Shmatikov. 2008. Robust de-anonymization of large sparse datasets (how to break anonymity of the netflix prize dataset). (2008).

[12] J. W. Pennebaker and L. A. King. 1999. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology* (1999), 1296–1312.

[13] C. Sumner, A. Byers, and M. Shearing. 2011. Determining personality traits and privacy concerns from Facebook activity. *Black Hat Briefings* (2011), 197–221.

[14] Laura Uba. 2003. *Asian Americans: Personality Patterns, Identity, and Mental Health.* Psychology, Guilford Press.

[15] Rui Wang, XiaoFeng Wang, Zhou Li, Haixu Tang, Michael K. Reiter, and Zheng Dong. 2009. Privacy-preserving Genomic Computation Through Program Specialization *(CCS)*. 338–347.