VirtualIdentity: Privacy Preserving User Profiling

Sisi Wang[1], Wing-Sea Poon[1], Golnoosh Farnadi[2,3], Caleb Horst[1], Kebra Thompson[1],
Michael Nickels[1], Anderson Nascimento[1], Martine De Cock[1,2]
[1]Institute of Technology, University of Washington Tacoma
Email: {sisiwang, wpoon93, calebjh, kebrat, mnickels, andclay, mdecock}@uw.edu
[2]Dept. of Applied Mathematics, Computer Science and Statistics, Ghent University
Email: {golnoosh.farnadi, mdecock}@ugent.be
[3]Dept. of Computer Science, Katholieke Universiteit Leuven
Email: golnoosh.farnadi@cs.kuleuven.be

***Abstract*—User profiling from user generated content (UGC) is a common practice that supports the business models of many social media companies. Existing systems require that the UGC is fully exposed to the module that constructs the user profiles. In this paper we show that it is possible to build user profiles without ever accessing the user's original data, and without exposing the trained machine learning models for user profiling – which are the intellectual property of the company – to the users of the social media site. We present VirtualIdentity, an application that uses secure multi-party cryptographic protocols to detect the age, gender and personality traits of users by classifying their user-generated text and personal pictures with trained support vector machine models in a privacy preserving manner.**

I. Introduction

As more users are creating their own content on the web, there is a growing interest to mine this data for use in personalized information access services, recommender systems, tailored advertisements, and other applications that can benefit from personalization [12]. In addition to myriad applications in e-commerce, there is a growing interest in user profiling for digital text forensics [18]. Furthermore, the popularity of applications such as How-Old.net and HowHot.io shows that users are directly interested in their own personal features analysis as well [16], [15]. What is common across all of these existing personalized services is that the personal data of users, such as their pictures and text, is fully exposed to the user profiling service.

An obvious way to circumvent this would be to perform the user profiling entirely on the user's side. However, this would imply sharing proprietary, trained machine learning models for user profiling with each user of the social media site. Applying traditional cryptography to encrypt the personal data of the user (henceforth called the client) before sending it to the user profiling service (the service, or server) is not a solution either, as data encrypted with usual techniques becomes useless, and user characteristics can no longer be derived from it. Hiding the client's data from the service, while still allowing the client to use the service, requires novel cryptographic techniques that not only protect private information but also allow mathematical operations to be performed on encrypted data. To this end, the VirtualIdentity application that we present in this paper (see Figure 1) relies on secure multi-party computation, a process in which client and server jointly compute classification labels by exchanging encrypted messages, while keeping their own inputs private. As a result, VirtualIdentity allows a user to run our trained support vector machines (SVMs) for detection of age, gender, and personality traits, without leaking any personal text or profile picture to our server. In addition, the user does not learn anything about the coefficients of our SVM models.

Other services exist that will predict a user's age, gender, or personality based on UGC. For example, users can input their tweets or text and receive back scores of their personality, needs, and values[11]. Another site allows users to input a photo and receive an estimation of the gender and age of each face in the photo[16] while a third estimates the users attractiveness and age from a photo[15]. However, none of these services attempt to keep the user's data private. To the best of our knowledge, VirtualIdentity is the first platform to construct user profiles while preserving both the privacy of the user's data and the prediction models.

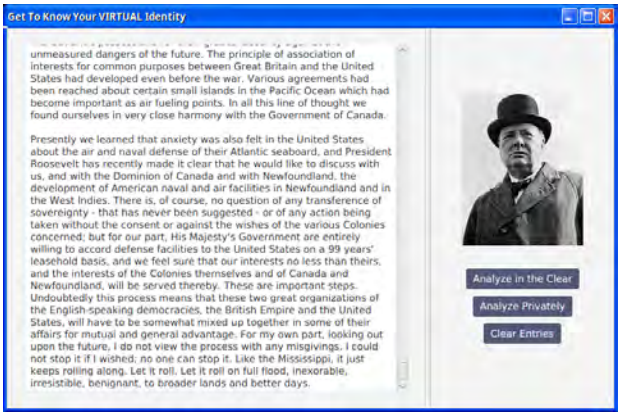
II. Predictive Models

Much work has been done recently using machine learning classification to predict age, gender, and personality based on images and text "in the clear", i.e. without any attempts for privacy preservation. In this paper we use SVMs, which are known as state-of-the-art classification techniques for detecting age, gender and personality traits from text and images [14], [9], [1], [10], [23].

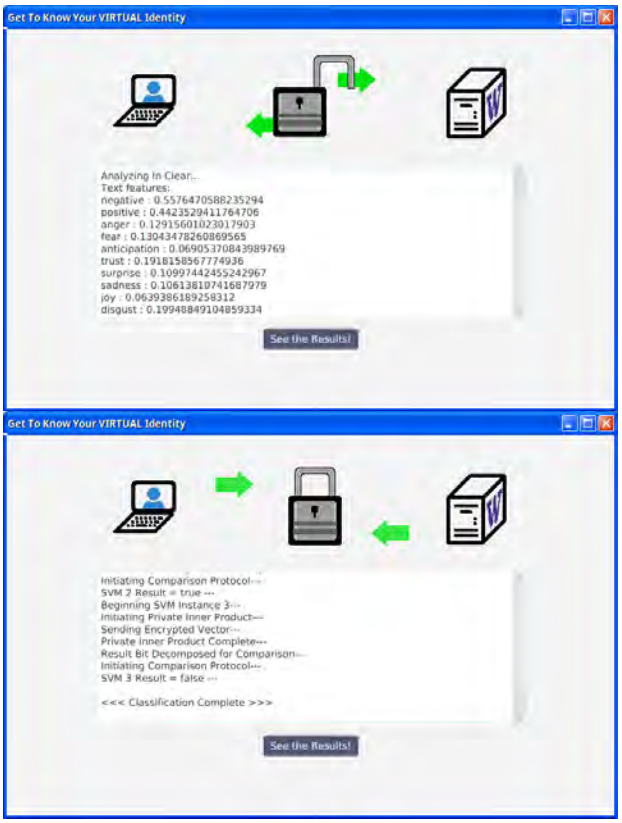
For age and gender classification we used the IMDB image dataset[21]. This set contains 460,723 images from which we extracted 136 facial landmark features using Dlib [22]. These features, which include attributes such as the exact locations of the eyes, nose, and mouth, were then used to train the models. After feature extraction, we have 318,562 valid instances remaining in the set. The set is divided into 4 similar-sized age groups: (7-26), (27-34), (35-43), (44-101). For age classification, each instance will be classified into one age bucket. For the actual training, we used 6000 of the IMDB dataset images such that the age and gender distributions of the selected images are representative of the full set.

For personality we report scores using the traits of the widely accepted model, the Big Five, consisting of the fol-

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Fransisco, CA, USA
978-1-5090-2846-7/16/$31.00 ©2016 IEEE



(a) The user inputs text and a profile picture.



(b) For demo purposes, the analysis is done both in the clear and in a privacy-preserving manner.

(c) The service returns age, gender, and personality analysis.

Fig. 1. Screenshots of VirtualIdentity application

lowing five results: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism [5]. Since more than one trait can be present in the same user, we trained a binary SVM classifier for each of the five traits that separates the users displaying the characteristic from those who do not. We trained our SVM models on a data set with 2467 essays (one empty instance was removed from the original 2468) from psychology students who were told to write whatever came to their mind for 20 minutes [14]. Each essay was analysed and given Big Five personality ground truth labels by Pennebaker et al. [19]. We extracted three kinds of features from the essays as input for the classifiers: 14 MRC features, 10 NRC features, and 19 LIWC features. MRC is a psycholinguistic database which contains psychological and distributional information about words such as the number of letters in the word, the concreteness, and the age of acquisition [3]. We used the same 14 MRC features as Farnadi et al., computing each MRC feature value of an essay by averaging the feature value of all the words in that essay [10]. NRC is a lexicon that contains more than 14,000 distinct English words annotated with 8 emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust), and 2 sentiments (negative, positive)[17]. For each document we counted the number of words in each of the 8 emotion and 2 sentiment categories, resulting in 10 features per document. The Linguistic Inquiry and Word Count tool (LIWC) is a well-known text analysis software which is widely used in psychology studies [20]. Part of the LIWC features rely on a proprietary dictionary. Our SVM models are trained on 19 LIWC features that relate to standard counts and that do not require the specific LIWC dictionary: word count, words per sentence, number of unique words, words longer than six letters, abbreviations, emoticons, question marks, periods, commas, colons, semi-colons, exclamation marks, dashes, quotation marks, apostrophes, parentheses, other punctuation marks, all punctuation marks, and interrogative sentences.

We trained the SVMs using scikit-learn in Python, with a linear kernel and penalty parameter $C = 1$. We trained a binary SVM classifier for each of the personality traits (5 SVMs total), a binary SVM classifier for gender classification, and three binary SVMs for age classification. We use the results of all three age classifiers to determine the most likely age bracket, similar to what Han et al. do[23]. While they use additional models to then predict an actual age inside the bracket, we return the result determined from the three, original SVMs.

The trained SVMs are part of a private machine learning model bank that resides on the server, as shown on the right side in Figure 2. When a user requests analysis of a snippet of text and a picture, the features described above are extracted from the text and the image on the client side, as shown on the left side in Figure 2. Neither the user's text, nor the user's image, nor any of the extracted features are leaked to the server. Instead, both the client and the server engage in cryptographic protocols and exchange encrypted messages that ultimately allow the server to classify the feature vectors of the client, without ever seeing them in the clear, as we explain

in Section III.

III. Adding Privacy to our Classifiers

Only a limited amount of work has been done in cryptographically secure privacy-preserving machine learning classification and none of it is aimed specifically at user profiling.

Cryptographically secure privacy-preserving SVM classification protocols have been proposed in [6], [13]. The basic idea behind these protocols is to decompose the task of scoring an SVM into smaller tasks and to implement each one of them in a privacy-preserving way. To better understand these previous approaches we recall that the general process for *SVM classification in the clear* is as follows [4]: the client holds an n-dimensional input feature vector , and the server holds a trained model $(, b)$, where is an n-dimensional vector of weights and b is a real number learned from the training data. The result of the classification is obtained by computing $\mathsf{sign}\left(\langle , \rangle + b\right)$, where $\mathsf{sign}\left(y\right)$ is $+$ if $y > 0$ and $-$ otherwise. For instance, in the case of personality prediction, is a 43-dimensional vector with features extracted from the client's text and $(, b)$ are the weights and the bias that make up the trained SVM model for e.g. "neuroticism". A classification outcome $+$ means that the user is neurotic, and an outcome of $-$ means that he is not. Therefore, to score SVMs privately, one needs to build two sub-protocols: a protocol for computing inner products privately and a protocol for obtaining privacy-preserving comparisons.

In [13], private inner products and comparisons are obtained by using additive homomorphic encryption and oblivious transfer, while in [6] the proposed protocols are based on Paillier encryption - a specific kind of additive homomorphic encryption scheme. These operations are usually expensive from a computational complexity point of view, demanding costly modular exponentiations.

In [7], highly efficient protocols for privacy-preserving comparison and argmax were proposed. The comparison protocol is based on the commodity-based model [2] which assumes that some data is pre-distributed by a trusted initializer during an off-line setup phase. The trusted initializer does not engage in the remaining steps of the protocol and never learns the client's or server's inputs. If this trusted authority is not desired, a pre-processing phase performed by the client and the server over an off-line phase can be used as an alternative to compute the pre-distributed data (also known as commodities)[7]. The online phase remains the same with or without the trusted authority. The protocol proposed in [7] has a highly efficient online phase, requiring only modular additions and multiplications. In [7], the authors comment on a potential application of their protocol for scoring SVMs by combining it with a protocol for computing inner-products in the commodity-based model proposed in [8]. However, to the best of our knowledge, an implementation of the protocol for evaluating SVMs hinted at in [7] has never been presented in the literature.

Here we show not just the first implementation of a system for solving the problem of privacy-preserving user profiling. We base VirtualIdentity on an optimized implementation of the comparison protocol proposed in [7], and the inner product protocol proposed in [8], thus showing that these protocols are practical within the context of a real world application.

We have already mentioned how we perform the private classification of personality traits. Now, we briefly describe how we proceed to obtain age and gender prediction. For age prediction, we first split the age groups into $n = 4$ classes, such that the frequency of each class is equal. Because there are $n = 4$ classes, there will be $n - 1 = 3$ splitting points. We therefore have an SVM for each of these splitting points; because each SVM is independent of the other, we run them all in parallel. Each SVM is a binary classifier which outputs whether the predicted age is greater than, or less than or equal to, the splitting point. We use the secure SVM as described above in combination with the argmax protocol from [7] to determine the final age group classification. The security of this protocol follows from the fact that the SVM described above is privacy preserving, as is the argmax protocol[7]. A separate SVM is evaluated in a privacy-preserving way to determine the gender of the user.

It should be noted, that the techniques used here for implementing privacy-preserving inner product, comparison, and argmax protocols only work for integer values. To account for this, real values must be converted into integers and lose some of the precision allowable by floating notation.

IV. System Overview

The overall architecture of our demo is shown in Figure 2. The framework consists of a client Java application, a server, and the cryptographic protocols embedded in client and server. Next, we describe these modules.

A. Client Application

The user interface of our client application shown in Figure 1 is developed with JavaFX. The client application consists of a feature extractor and its respective portion of the cryptographic protocols. It allows users to upload user generated content (i.e. to input written text and to upload a personal picture). It extracts features from the UGC, executes cryptographic protocols with the server, and interprets and displays the final prediction results from the machine learning models. The interpretation of personality refers to Personality Insights[11].

B. Server

The server contains its respective portion of the cryptographic protocols and the private machine learning model bank. The model bank contains the SVMs which are used for predicting personality traits, age and gender.

C. Cryptographic Protocols

The cryptographic protocols (protocols for computing privacy-preserving inner products, comparisons and argmax), are executed in both the client and server side. The trusted initializer pre-distributes correlated data to the client and the

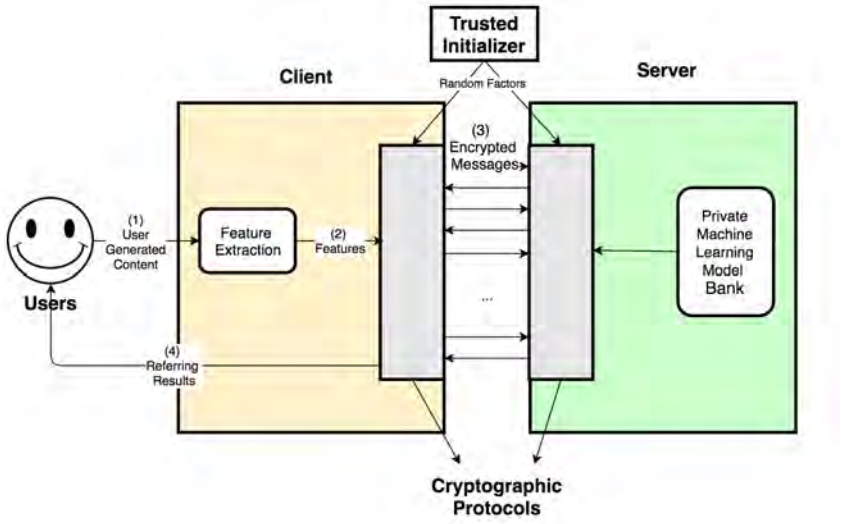


Fig. 2. System overview of the VirtualIdentity application

server as specified in the commodity-based model during an off-line phase [7], [2]. The communication between client, server is implemented using sockets. The whole VirtualIdentity application is programmed with Java under JDK 1.8.

V. Conclusion

Many data-driven personalized services require that private data of users such as user generated content, personal preferences, browsing behavior, or medical lab results is scored with proprietary, trained machine learning models. The current widespread practice expects users to give up their privacy by sending their data in the clear to the server where the machine learning models reside. In this paper we have demonstrated that the use of secure multi-party computation techniques allows the construction of user profiles from user generated content while preserving both the privacy of the users data and the prediction models. The overall architecture of the VirtualIdentity application is generic and can be extended to other applications; this would involve extraction of different features and training new models for the private machine learning model bank.

References

[1] F. Celli, E. Bruni and B. Lepri, "Automatic Personality and Interaction Style Recognition from Facebook Profile Pictures," *Proc. of ACMMM2014*.

[2] Beaver, D., 1997, May. Commodity-based cryptography. In Proceedings of the twenty-ninth annual ACM symposium on Theory of computing (pp. 446-455). ACM.

[3] M. Coltheart, "The MRC Psycholinguistic Database", *Quarterly Journal of Experimental Psychology*, Vol. 33A, p. 497505, 1981.

[4] C. Cortes and V. Vapnik, "Support-vector networks", *Machine Learning*, Vol. 20(3), p. 273-297, 1995.

[5] P. T. Costa and R. R. McCrae, "The Revised NEO Personality Inventory (neo-pi-r)," in *The SAGE Handbook of Personality Theory and Assessment*, Thousand Oaks, CA, SAGE Publications Inc, 2008, p. 179-198.

[6] R. Bost, R. Popa, S. Tu, S. Goldwasser: Machine Learning Classification over Encrypted Data. 22nd Annual Network and Distributed System Security Symposium, NDSS 2015

[7] David, B., Dowsley, R., Katti, R. and Nascimento, A.C., 2015. Efficient Unconditionally Secure Comparison and Privacy Preserving Machine Learning Classification Protocols. In Provable Security (pp. 354-367). Springer International Publishing. Vancouver

[8] David, B., Dowsley, R., van de Graaf, J., Marques, D., Nascimento, A.C. and Pinto, A.C., 2016. Unconditionally Secure, Universally Composable Privacy Preserving Linear Algebra. Information Forensics and Security, IEEE Transactions on, 11(1), pp.59-73.

[9] E. Eidinger, R. Enbar and T. Hassner, "Age and Gender Estimation of Unfiltered Faces," *IEEE Trans. Inf. Forensic Secur.*, Vol. 9(12), p. 2170-2179, 2014.

[10] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.F. Moens, M. De Cock, "Computational Personality Recognition in Social Media", *User Model. User-Adapt. Interact.*, 2016.

[11] IBM Watson Developer Cloud, "Personality Insights," https://personality-insights-livedemo.mybluemix.net/ [Accessed 29 4 2016].

[12] M. Kosinski, Y. Bachrach, P. Kohli, D. Stillwell, T. Graepel, "Manifestations of User Personality in Website Choice and Behaviour on Online Social Networks," *Mach. Learn.*, Vol. 95(3), p. 357-380, 2013.

[13] S. Laur, H. Lipmaa and T. Mielikainen, "Cryptographically Private Support Vector Machines," *Proc. ACM SIGKDD2006*.

[14] F. Mairesse, M. A. Walker, M. R. Mehl and R. K. Moore, "Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text," *J. Artif. Intell. Res.*, Vol. 30, p. 457-500, 2007.

[15] Merantix & Blinq, "Let Artificial Intelligence Guess your Attractiveness and Age," http://howhot.io/ [Accessed 29 April 2016].

[16] Microsoft Cognitive Services, https://how-old.net/ [Accessed 29 April 2016].

[17] S. Mohammad, X. Zhu, J. Martin, "Semantic Role Labeling of Emotions in Tweets", *Proc. of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, 2014.

[18] PAN Evaluation Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse, http://pan.webis.de/ [Accessed 02 May 2016].

[19] J.W. Pennebaker, L.A. King, "Linguistic Styles: Language Use as an Individual Difference", *J. Pers. Soc. Psychol.*, Vol. 77(6), p. 1296-1312, 1999.

[20] Y.R. Tausczik, J.W. Pennebaker, "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods", *J. Lang. Soc. Psychol.*, Vol. 29, p. 2454, 2010.

[21] R. Rothe, R. Timofte, L. Van Gool, "DEX: Deep EXpectation of apparent age from a single image", *ICCV, ChaLearn Looking at People workshop*, December, 2015.

[22] Learn OpenCV: Facial Landmark Detection, http://www.learnopencv.com/facial-landmark-detection/ [Accessed 3 May 2016].

[23] H. Han, C. Otto, A.K. Jain, "Age Estimation from Face Images: Human vs. Machine Performance", in *Proc. 6th IAPR International Conference on Biometrics (ICB)*, 2013.