2017 IEEE 33rd International Conference on Data Engineering

Privacy Cyborg: Towards Protecting the Privacy of Social Media Users

Theodore Georgiou
Computer Science Department
University of California, Santa Barbara
teogeorgiou@cs.ucsb.edu

Amr El Abbadi
Computer Science Department
University of California, Santa Barbara
amr@cs.ucsb.edu

Xifeng Yan
Computer Science Department
University of California, Santa Barbara
xyan@cs.ucsb.edu

***Abstract*—Towards the vision of building artificial intelligence systems that can assist with our everyday life, we introduce a proof of concept for a social media privacy "cyborg" which can locally and privately monitor a person's published content and offer advice or warnings when their privacy is at stake. The idea of a cyborg can be more general, as a separate local entity with its own computational resources, that can automatically perform several online tasks on our behalf. For this demonstration, we assume an attacker that can successfully infer user attributes, solely based on what the user has published (topic-based inference). We focus on Social Media privacy and specifically on the issue of exposing sensitive user-attributes, like location, or race, through published content. We built a privacy cyborg that can monitor a user's posted topics and automatically warn them in real time when a sensitive attribute is at risk of being exposed.**

Video: https://youtu.be/PfzC39i9nbg

I. Introduction

The public nature of Online Social Networks, like Twitter and Facebook, has introduced a different privacy danger from the more traditional linkage attack (identifying the real identity of an online user). Attribute inference, the process of inferring an OSN user's attributes like age, gender, location, race, political preference, etc., can be extremely useful for the purposes of personalization in content recommendation, advertising, and/or social media analytics. For example, large Social Media websites like Facebook and Twitter already have proprietary methods for inferring social attributes of their users that are not explicitly provided by them. Recently, it was discovered that Facebook is able to learn a user's political preference between values like "Liberal", "Moderate", or "Conservative". However, if a third-party attacker is capable of inferring attributes that are sensitive or private then it is important to build techniques that can protect OSN users.

Additionally, the concept of *Trending Topics*, a popularly utilized mechanism for the detection of breaking news, hyper-local events, and memes, can make things even more complicated. Trending topics are often correlated with specific attributes, for example Twitter reports trending topics by location or interests. In the presence of even more sophisticated trending algorithms that capture several attributes apart from location, reports of trending topics further enable attribute inference attacks. For *example*, if it is reported that people that mentioned topic $\#BlackLivesMatter$ are 79% teenagers, 86% African Americans, and 67% live in Chicago, then an attacker can infer the age, race, and location of any user that mentions this hashtag with some statistical confidence. We introduced a method for the extraction of trending topics and correlated attributes in [1] and showed that it can be done on Twitter data realistically in real time and significantly improve the quality of reported topics. On the other hand, in the presence of such a reporting system, users must be mindful of which topics they discuss in order to protect themselves from such inference attacks. This can be particularly tedious and time consuming given the nature of social media which promotes public and frequent posting, something that usually seems harmless when considered at the level of a single post. Towards this end, we built a privacy cyborg, that can undertake the task of monitoring its owner's posts in social media and automatically warn them if necessary.

The idea of a cyborg fighting our social battles on our behalf was recently introduced by Anand Rajaraman [2]. The vision involves a local computational software resource that runs continuously (whether its owner is online or offline) and performs various tasks like protecting from online attacks, filtering out people that try to connect with malicious intend, following up on discussions, etc. Different tasks require different levels of sophistication and technology, but the cyborg in this demonstration mainly constitutes a proof of concept that targets online privacy. Similar to Privometer [3], the privacy cyborg can monitor a Twitter user's profile and what is being posted to identify potential risks of leaking sensitive information such as the user's location, race, or age. In contrast to Privometer which considers structural actions, i.e., connecting with a friend, joining a group, or liking a page, our focus is on the posted content and its role in revealing sensitive attributes.

II. Privacy Model

The goal of the cyborg is to preserve the privacy of its owner, hence, we need to understand how sensitive attribute inference works, and implement it locally to simulate a hypothetical attack. Through this process the cyborg can identify what an attacker could infer, make a judgment call whether an attack can indeed be successful or not, and in case of the former warn the user in an appropriate way.

The Bayesian model is often used for inference attacks in the literature. We follow the same model here, and assume that the attacker can acquire knowledge that involves the correlation of topics and attributes. Based on this knowledge an attacker can identify **the most probable** value of a sensitive attribute by comparing each value's probability (e.g., 79% chance a user is a woman versus 21% chance the user is a man). More

2375-026X/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDE.2017.193

specifically, based on the individual prior probabilities of the topics T mentioned by a user, an attacker can calculate the probability $P(A|T)$ for each attribute A.

We assume that an attacker can acquire the following knowledge: The general prior probability distribution of a sensitive attribute A, $P(A)$. The observed conditional probability distribution of a topic t given the attribute A, $P(t|A)$, which can be derived from a rich trending topics report (similar to [1]). With this information, the attacker can approximate the probability distribution of a user's sensitive attribute A, given the user's set of topics T: $P(A|T)$. We assume that the attacker can successfully infer the value of an attribute A if any value of $P(A|T)$ is greater than a desired threshold, e.g., 0.75.

III. The Cyborg: Supported Operations

The privacy cyborg is implemented as a daemon process that runs constantly on a local machine. While the cyborg can obviously interact with its owner when they are using their personal computer, it still needs to monitor the correlations between topics and attributes even when the owner is offline. Thus, the cyborg can perform the following two tasks: 1) Inform the user of how the public perceives them, and 2) warn the user if something they are about to post can put their sensitive attributes in danger.

More specifically, since the cyborg is practically simulating an inference attack in real time, it is able to derive a description of its owner, as perceived from their publicly posted content. This description is given as a report with a list of attributes and the corresponding probabilities for each value (e.g., male 34%, Los Angeles 56%, etc.). Since we focus on Twitter data for this demo, this task can be performed on any non private account without additional permissions. Note that these reports can be returned on demand or triggered when something changes. Due to the nature of the inference process, probabilities can change without the user posting anything—could be the result of a population shift for a topic of interest.

However, since the publication of new topics remains the most effective way for these probabilities ($P(A|T)$) to change the cyborg can proof-read a new post and inform them whether any sensitive attributes will be compromised, i.e., an attacker will be able to successfully infer them following the publication of the post.

IV. The Cyborg: Technical Components

The cyborg needs access to the same knowledge as the attacker it tries to simulate. Namely, access to the historical tweets of its owner, access to reports of topics and the correlated attribute values with specific percentages, the prior general probability distributions of the attributes, and the user-provided settings for the privacy threshold (the attribute inference probability is high enough to pose a privacy risk).

Figure 1 is a visualization of how the cyborg works. The cyborg owners and social media users (bottom left and right) want to post content online that contains a growing list of topics (action 2). This list goes through the cyborg for proof-checking. In the background, the cyborg consumes information from the social media service (action 1) and monitors the topics of the user (both old and new). With this acquired knowledge

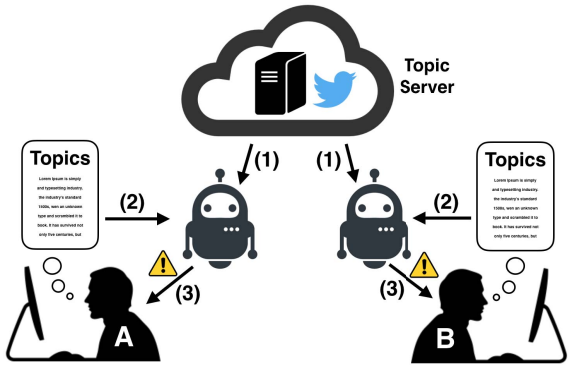

Fig. 1. Visualization of the cyborg, related actions, and components. Note that each owner/user has their own cyborg.

the cyborg can warn its owner when a sensitive attribute is at risk (action 3).

The cyborg itself comprises of a daemon running on a local computer with Internet access so it can acquire and maintain the necessary data for calculating the attribute probability distribution $P(A|T)$. Interactions with the owner are supported though a graphical UI where the proof checking and warning displaying can be performed. The prior probability distributions and correlated topics/attributes are provided by an external server (the cloud in Figure 1) and are assumed to be public knowledge, similar to how trending topics on Twitter are public. This server constantly collects tweets from the Twitter Streaming API, and computes in real time the correlation between topics with specific attribute values. To identify the attribute values of other users, a periodic job is executed on the server that infers the attributes of any other user in the social stream (detailed in [1]). This information is then used to calculate the percentages of association between attribute values and topics in the stream.

V. Conclusion

We introduce the notion of a privacy cyborg, an entity that runs constantly and monitors its owner's social media privacy in the context of sensitive attribute inference. This cyborg can simulate an attacker to identify when a sensitive attribute might be at risk to be revealed publicly and notifies in real time when this happens, while the user tries to post something new or is even offline. During the demo, attendees will be able to instantiate a cyborg process for their Twitter account id, and the cyborg will demonstrate its ability to infer attributes and warn owners when they tweet on a topic that may expose some of their sensitive attributes.

Acknowledgments: This work is supported by NSF grant CNS 1649469.

References

[1] T. Georgiou, A. E. Abbadi, and X. Yan, "Extracting topics with focused communities for social content recommendation," in *Proceedings of the 20th Conference on Computer-Supported Cooperative Work & Social Computing, CSCW, Portland, OR, USA, February 25 - March 1*, 2017.

[2] A. Rajaraman, "Data-driven disruption: The view from silicon valley," *PVLDB*, vol. 9, no. 13, p. 1620, 2016.

[3] N. Talukder, M. Ouzzani, A. K. Elmagarmid, H. Elmeleegy, and M. Yakout, "Privometer: Privacy protection in social networks," in *Workshops Proceedings of the 26th International Conference on Data Engineering, ICDE 2010*, pp. 266–269.