SGX-Privinfer: A Secure Collaborative System for Quantifying and Mitigating Attribute Inference Risks in Social Networks

Hervais Simo and Michael Kreutzer
Fraunhofer SIT, Darmstadt, Germany

Keywords: Privacy, Social Networks, Attribute Inference, TEE, SGX.

Abstract: The growing popularity of Online Social Networks (OSNs) over the past decade has led to a significant portion of the global population sharing diverse personal information online, including relationship status, political affiliations, and religious views. However, research has shown that adversaries, such as third-party application providers and law enforcement agencies, can aggregate and correlate seemingly innocuous, publicly available data across various platforms. This process can uncover sensitive insights about individuals, often far beyond what users intend or realize they are disclosing. To mitigate this challenge, it is essential to provide OSN users with enhanced transparency and control over their digital footprints and the associated risks of attribute inference, as emphasized by regulations like the EU General Data Protection Regulation (GDPR). Innovative solutions in this domain often rely on Privacy Inference Detection Technologies (PIDTs), which empower users to understand and manage such risks. However, existing PIDTs raise significant privacy concerns, as they typically require highly sensitive data to be transferred to cloud services for analysis, exposing it to potential misuse or unauthorized access. To address these limitations, we introduce SGX-PrivInfer, a novel architecture that enables OSN users to collaboratively and securely detect and quantify attribute inference risks based on public profile data aggregated from multiple OSN domains. SGX-PrivInfer leverages Trusted Execution Environments (TEEs) to safeguard the confidentiality of both user data and the underlying attribute inference models, even in the presence of curious adversaries, such as cloud service providers. In its current design, we utilize Intel SGX as the implementation of TEEs to achieve these security guarantees. Our performance evaluation, conducted on real-world OSN datasets, demonstrates that SGX-PrivInfer is both practical and capable of supporting real-time processing. To the best of our knowledge, SGX-PrivInfer is the first architecture and implementation of a PIDT that offers strong security guarantees, data protection, and accountability, all backed by Intel SGX's hardware-enforced isolation and integrity mechanisms.

1 INTRODUCTION

Privacy Inference Detection Technologies (PIDT) on OSNs allow users' public data to be collected from different social media platforms and sent to a remote cloud service, where it is combined into aggregated ego graphs and inference risks per profile attribute are calculated. Unfortunately, by relying on a client-server model, existing PIDT proposals, e.g., (Simo et al., 2021,Talukder et al., 2010,Guha et al., 2008,Jia and Gong, 2018) and tools such as Apply Magic Sauce (https://applymagicsauce.com/demo), IBM Watson Personality insights (https://watson-developer-cloud.github.io/swift-sdk/services/PersonalityInsightsV3/index.html), require users to fully trust the remote server while receiving limited security guarantees. As a result, there are growing concerns about data protection and privacy. In fact, the data uploaded by users to the remote server, which is controlled by a third party or cloud provider, as well as sensitive inferences drawn from the aggregated user data, can be stolen or misused. For instance, an adversary, such as the partially trusted service provider, could use this data either for purposes other than those for which it was originally collected, or more broadly, without any legal basis. However, none of the previous work on PIDT addresses the challenge of ensuring the integrity and confidentiality of sensitive user data sent to and generated by the remote analysis server, nor the confidentiality of the attribute inference model.

Our Contribution. This paper proposes *SGX-PrivInfer*, a Trusted Execution Environment (TEE)-

Simo, H. and Kreutzer, M.
SGX-PrivInfer: A Secure Collaborative System for Quantifying and Mitigating Attribute Inference Risks in Social Networks.
DOI: 10.5220/0013390000003899

In *Proceedings of the 11th International Conference on Information Systems Security and Privacy (ICISSP 2025) - Volume 2*, pages 111-122
ISBN: 978-989-758-735-1; ISSN: 2184-4356

enhanced PIDT architecture that aims to prevent unauthorised/curious entities (e.g. entity operating the remote server) from accessing sensitive user data while still being able to provide inference risk assessment as a service. In the current design, we lever-age Intel Software Guard Extensions (SGX) (https://software.intel.com/en-us/sgx) to instantiate the TEE. We perform an empirical performance evaluation of SGX-PrivInfer on a real-world dataset of 27.520 ego graphs, demonstrating that the over-head induced by the proposed privacy extensions and mechanisms is low enough to enable real-time quantification of attribute inference.

Outline. The remainder of this paper is organized as follows: The next section provides the background and related work, including a review of private attribute inference attacks and an overview of transparency-enhancing technologies (TETs) and PIDTs. Section 3 introduces the design of SGX-PrivInfer, including the system model, threat model, design goals, and an architectural overview. Section 4 presents a proof-of-concept implementation along with its preliminary evaluation. Finally, Section 5 concludes the paper by summarizing our key findings and outlining future research directions.

2 RELATED WORK

Single-Domain Private Attribute Inference Attacks. OSN users often lack a full understanding of how their information is being collected and used, and what the trade-off is between having their data collected for the purpose of accessing services at virtually no cost, and the consequences that doing so could entail, especially in terms of implications for their privacy (Andreou et al., 2018, Maheshwari, 2019). This lack of both user awareness and support for transparency can actually cause or aggravate the privacy problems associated with the implicit self-disclosure of private information. That is, when sharing personal information on a social network platform, most users assume that their data will never leave the boundaries of that specific social network and only be accessible to a restricted group of users, i.e. their ”friends”. However, this assumption is far from reality. As a matter of fact, an adversary can easily jeopardize the privacy of the OSN user by searching the social network site and gathering seemingly non-sensitive and publicly available i.a. profile attributes, behavioral data (e.g., likes (Kosinski et al., 2013)) and other metadata such as location check-ins and topological properties of the victim’s social graph (Jurgens, 2013, Labitzke et al., 2013). Based on this information, the adversary can then perform local inference attacks to predict supposedly private and hidden details of the profile owner. Instances of such details includes sensitive identity attributes (gender, age, sex, ethnicity, occupation, income level, relationship / marital status, education level, family size, religion views) (Mislove et al., 2010, Gong and Liu, 2016, Gong et al., 2014), personality traits (concienciousness, agreeableness, neuroticism, openness, and extraversion) (Kosinski et al., 2013, Volkova and Bachrach, 2015), home location (Pontes et al., 2012, Li et al., 2012), political preferences and views (Idan and Feigenbaum, 2019, Volkova et al., 2014), current emotional state (Collins et al., 2015), depression and stress level (De Choudhury et al., 2013), sexual orientation (Wang and Kosinski, 2018, Zhong et al., 2015) and household income (Luo et al., 2017,Fixman et al., 2016). Other work on similar lines, yet with a focus on ego graph include Zheleva and Getoor (Zheleva and Getoor, 2009), He et al. (He et al., 2006), Ryu et al. (Ryu et al., 2013). In addition to texts posted on social media, typical sources of information for local attribute inference attacks also include user-generated images. Among approaches considering information from these various sources, a growing line of research on privacy inference from user-generated photographs and videos posted on OSNs is especially worth mentioning. Here, machine learning techniques from the field of computer vision are leveraged to deduce identity attributes (gender (Rangel et al., 2018, Ciot et al., 2013), race, and age (Chamberlain et al., 2017, Fang et al., 2015)) and personality traits from non-tagged images (Ferwerda and Tkalcic, 2018, Oh et al., 2016), sexual orientation from facial images (Wang and Kosinski, 2018), and detect hidden information like faces (Sun et al., 2018, Joo et al., 2015, Joon Oh et al., 2015), occupation (Chu and Chiu, 2014, Shao et al., 2013), social relationships (Sun et al., 2017, Wang et al., 2010), and hand-written digits (LeCun et al., 1989), sometimes even from images protected by various forms of obfuscation (Oh et al., 2017, McPherson et al., 2016).

Cross-Domain Private Attribute Inference Attacks. As highlighted by various researchers, e.g., (Xiang et al., 2017, Qu et al., 2022), a more powerful adversary, such as a malicious data aggregator, can collect, aggregate, and correlate disparate pieces of information about a targeted user and her contacts, across multiple, previously isolated

domains. By combining data from various OSNs and other platforms, the adversary can construct a highly detailed profile of the victim and/or link profiles across services. This practice significantly amplifies security and privacy risks (Irani et al., 2009), as it enables sophisticated profiling and identification techniques that go beyond what any single platform might reveal.

Transparency-Enhancing Technologies (TETs) and Privacy Inference Detection Technologies (PIDTs) for OSN. The General Data Protection Regulation of the European Union emphasizes transparency as a fundamental enabler of informational self-determination, as articulated in Articles 12 and 13. TETs (Janic et al., 2013, Murmann and Fischer- Hübner, 2017) play a crucial role in helping users understand and visualize how their data is collected, processed, and used. This is particularly significant in the context of challenges such as user profiling and private inference threats, where the lack of transparency can exacerbate privacy risks. Murmann et al. (Murmann and Fischer-Hübner, 2017) provide a comprehensive survey of ex-post TETs, including Privacy Inference Detection Technologies (PIDTs). Similar to TETs, PIDTs are designed to provide users with deeper insights into the data collected about them, while also helping them understand the potential privacy risks associated with that data. Over time, numerous PIDTs and related tools have been developed to address privacy risks in OSN, e.g., (Simo et al., 2021, Talukder et al., 2010, Jia and Gong, 2018, Aghasian et al., 2017, Cai et al., 2016). (Talukder et al., 2010) introduced Privometer, a tool for measuring the extent of information leakage from an OSN user profile. It leverages Chakrabarti et al.'s network-only Bayes Classifier inference model (Chakrabarti et al., 1998), assuming an adversarial model with access to the target user's ego graph. Privometer provides a variety of visual tools to empower users, including ego graph visualization, analysis of friends' contributions to privacy leakage, and suggestions for self-sanitization. In (Cai et al., 2016), Cai et al. propose data sanitization strategies (attribute-removal and link-removal) for preventing sensitive information inference attacks. Jia et al. (Jia and Gong, 2018) proposed AttriGuard, which leverages a two-phase noise-adding process to defend against attribute inference attacks. They show the feasibility of leveraging adversarial examples (specifically, evasion attacks) (Goodfellow et al., 2014) to the unique challenges of defending against private attribute inference attacks. Many other models have been proposed for scoring the privacy of a user by rating the risk of leakage of user profile attributes, e.g., (De and Imine, 2017, Zhang et al., 2017, Aghasian et al., 2017). In (Aghasian et al., 2017), Aghasian et al. proposed algorithm for calculating the privacy score across multiple social networks and across different types of content. In (Simo et al., 2021), Simo et al. introduced on PrivInferVis, a tool to help users assess and visualize individual risks of attribute inference across multiple online social networks. While these existing TETs and PIDTs provide valuable foundations, they fail to address critical challenges, such as ensuring the confidentiality of self-disclosed and inferred user data and the trustworthiness of the underlying inference model. In this paper, we build upon and extend this body of work by introducing SGX-PrivInfer, which leverages Trusted Execution Environments (TEEs) (Li et al., 2024) to enforce strict hardware-based isolation for critical data such as user inputs, inferred attributes, and the inference model itself.

3 SGX-PrivInfer DESIGN

This section presents the current design of the SGX-PrivInfer framework, detailing the underlying system model, threat model and associated assumptions, as well as an overview of the system's design objectives.

3.1 System Model

The proposed framework operates in a setting involving multiple end users and two remote entities. Specifically, there is a set of n Users who wish to collaboratively pool their data to conduct secure and privacy-preserving inference analyses on shared OSN data. These users interact with their respective Social Networks from where data is collected. The setting at hand also include a Model Supplier (MS) responsible for providing the software and inference models required for analysis, and a Service Provider (SP) that manages the analytic server hosted on untrusted third-party infrastructure. In this paper, we build on previous work from (Simo et al., 2021) by adding a TEE to the analytic server. This addition ensures secure and privacy-preserving attribute inference quantification on shared data. As a result, the analytic server consists of two components: a trusted component and an untrusted component. The trusted component or TEE securely hosts the attribute inference model and performs the inference analysis. The untrusted component

encompasses all other server resources that lie outside the TEE and do not have access to sensitive computations or data. In the current design, we utilize Intel SGX to instantiate the TEE, providing strong hardware-backed isolation and security guarantees.

3.2 Threat Model & Assumptions

In this study, two realistic adversaries are considered: 1) honest-but-curious adversaries, which may include SGX-PrivInfer clients, other users, or the model provider; and 2) malicious adversaries, which may include external hackers. While honest-but-curious adversaries permit the attribute inference process to proceed as intended, their primary objective is to compromise the privacy of users or participants, for instance, by identifying or tracking them. Malicious adversaries have complete control of the server used for inference analysis, including root privileges. In addition to seeking to derive sensitive details about the attribute inference model, such as the model weights, and training parameters, they may also seek to gain further insights into the model's inner workings. Such details can then be exploited by the adversary to undermine the trustworthiness of the attribute inference model. Furthermore, the MS may not behave faith-fully, for example by deploying a corrupted inference model, which would result in skewed inference results. In light of the above, our primary concern is the preservation of the privacy of all involved users, and the integrity of the inference model.

Assumptions. We assume that OSN users will provide correct and valid inputs to the server, as both parties share a mutual interest in producing accurate inference results. Similarly, we assume that the server will exchange only valid data with its TEE, ensuring proper and secure data flow in both directions. However, SGX-PrivInfer does not defend against the injection of false information by relevant stakeholders (e.g., users, the curious Service Provider, or the Model Supplier), nor does it offer protection against Denial-of-Service attacks caused by the server or TEE forwarding malicious or excessive data. Furthermore, we consider availability to be an orthogonal issue that lies beyond the scope of our current investigation. SGX-PrivInfer does not attempt to prevent the SP from halting the processing of client requests. However, our solution can be extended in future work to include concepts such as server/service replication (Kapritsos et al., 2012), which would address availability concerns. Additionally, we assume that the SGX-PrivInfer server is equipped with a sufficiently large TEE to handle the sensitive inference operations. This requirement can be fulfilled by leveraging one of the many TEE-enabled trusted containers specifically designed for such purposes (Paju et al., 2023). Furthermore, we assume that malicious adversaries do not have physical access to the TEE's internal environment. This prevents invasive attacks, such as key extraction or tampering with the code running within the secure processor, as outlined in prior work (Nilsson et al., 2020, Fei et al., 2021). Protecting against physical attacks, side-channel attacks, or potential exploitation of vulnerabilities in TEEs and their associated SDKs lies beyond the scope of this work. Instead, we rely on TEE developers to implement state-of-the-art protections against such threats, as outlined in (Nilsson et al., 2020, Fei et al., 2021). Consequently, we assume that the TEE provides robust integrity and confidentiality guarantees for its internal state, code, and data.

3.3 Design Goals & Requirements

The core concept of SGX-PrivInfer is to allow OSN users to securely upload their data to an untrusted remote server, where ego graphs are reconstructed and subsequently fed into an attribute inference model hosted within a TEE. Based on the system and threat models described earlier, SGX-PrivInfer is designed to satisfy the following three sets of requirements: 1) *Functional Requirements*: Ensuring the correct and seamless execution of attribute inference tasks, including data handling, graph reconstruction, and model processing, while maintaining usability for OSN users; 2) *Privacy and Security Requirements*: It is imperative that no entity is able to determine, extract or corrupt the input data transferred to the remote server, the output of the attribute inference model or other sensitive details about the inference model itself; and 3) *Performance Overhead and Extensibility Requirements*: Minimizing computational and communication overhead to enable practical real-time processing, while ensuring that the framework is scalable and extensible to accommodate future enhancements or alternative TEE implementations.

3.4 Architecture Overview

To achieve the aforementioned goals, we designed SGX-PrivInfer as a client-server framework, with a central analytics server operating on Intel SGX-enabled, untrusted third-party infrastructure. The

architecture is modular and scalable, allowing adaptability to various use cases and environments. At its core, our framework combines TEEs and advanced analytics to facilitate secure and privacy-preserving attribute inference analysis.

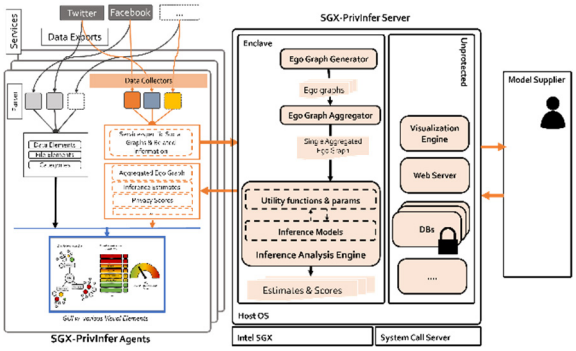

Figure 1: Architectural Overview of SGX-PrivInfer.

3.4.1 SGX-PrivInfer Client Application

The SGX-PrivInfer Client Application serves as the primary interface for user interaction, functioning as a dedicated client application or user agent. Each user agent comprises two main subcomponents: a Data Collector and a set of Data Visualization elements. Together, these components offer a comprehensive suite of features, including user registration, account and group management, and intuitive visualization of both user data and inference analysis results.

Data Collector. This component allows users to efficiently retrieve their profile data from various OSNs. It supports two primary options for data collection: (1) data retrieval via the OSN's official API using an access token and the required user permissions, and (2) a web scraper that directly extracts information from the user's profile page(s). Additionally, the first option includes the ability to utilize 'Download Your Data' tools offered by most OSNs, such as Facebook's data download functionality (https://www.facebook.com/help/212802592074644) and LinkedIn's equivalent tool (https://www.linkedin.com/help/linkedin/answer/a1339364/download-your-account-data?lang=en). This dual approach ensures flexibility and robustness, accommodating diverse data collection scenarios and user preferences. Building on (Simo et al., 2021), SGX-PrivInfer features a graphical user interface (GUI) enriched with advanced visualization elements to provide OSN users with greater transparency and control over their exposure to attribute inference attacks on social media. Specifically, SGX-PrivInfer's Visualization Elements allow users to explore their privacy-related online data, comprehend the inferences derived from it, and take appropriate actions to mitigate associated privacy risks. The client application offers a variety of visualization features to present results from the data analysis module. These features include an *overall privacy score* that presents and rates the user's privacy level, *a location map* for visualizing location-based attributes of group members, an *ego graph visualization* that highlights connections and relationships, and a *detailed overview of derived attributes* along with the associated level of certainty for each inference. Additionally, the GUI serves as an interactive gateway for users to engage with other components of the SGX-PrivInfer framework. It provides access to essential functionalities such as user authentication and account management, remote server attestation, and inference control. These features ensure seamless operation, user-friendly interaction, and enhanced transparency throughout the system.

Privacy Inference Indicator. SGX-PrivInfer offers users a comprehensive view of inference results derived from public information aggregated from their social graph. The Inference Indicator uses an intuitive gauge-based visualization, with scores ranging from 0 to 100. A score of 0 indicates no attributes have been inferred, while a score of 100 reflects complete privacy exposure. To aid user understanding, the gauge is color-coded: green indicates low exposure, while red signifies high exposure. For deeper insights, the Inference Indicator also displays detailed information about the user's input data, including attributes from each ego graph and the top inference results. These inference scores are securely computed within the TEE on the server side and leveraging WAPITI (Simo and Kreutzer, 2024).

3.4.2 Analytics Server

The SGX-PrivInfer analytics server consists of two main components: a trusted enclave and an unprotected server application. The trusted enclave hosts critical components that ensure security and confidentiality, including the User Authenticator Component for validating client credentials, the Remote Attestation Component for verifying the integrity of the enclave, the Sealing-Unsealing Component for secure data storage and retrieval, and the Inference Analysis Engine for performing attribute inference computations. Outside the enclave, the unprotected server application manages components such as Encrypted Storage for the

inference model, provided and encrypted by the Model Supplier, and databases for storing user records, group membership information, and other non-sensitive housekeeping data. Persistent storage, such as hard disks, is also part of the untrusted region and is primarily used for storing non-sensitive information. Interactions between the trusted and untrusted regions occur through OCALLs and ECALLs, ensuring controlled and secure transitions between the two components. Client-server interactions in SGX-PrivInfer are conducted over a network using the HTTP(s) protocol, with the framework relying on TEE-supported TLS (Knauth et al., 2018) to ensure all communications are authenticated and encrypted, thereby maintaining confidentiality and integrity. Additionally, the SGX-PrivInfer server facilitates the modeling and aggregation of ego graphs (Simo et al., 2021), performs attribute inference analysis on the aggregated graphs, and generates visual representations of the analysis results. These visual outputs are securely transmitted to the SGX-PrivInfer Client Application, enabling users to effectively interact with and interpret the inference results. ***Inference Analysis Engine.*** SGX-PrivInfer provides attribute inference assessment as a service, using a probabilistic model to evaluate an adversary's ability to infer sensitive profile attributes from observations of aggregated ego graphs. At its core, it leverages WAPITI (Simo and Kreutzer, 2024), a model specifically designed to quantify the risk of attribute inference in aggregated OSN ego networks.

The SGX-PrivInfer Web Server. The SGX-PrivInfer server hosts a web server instance that provides multiple REST endpoints to serve client requests. These endpoints support a range of functionalities, including user and group registration, authentication and management, data analysis, secure communication, and data transfer.

Visualization Engine (VE). The VE is a versatile component built on top of free and open-source data visualization libraries. It allows SGX-PrivInfer to create rich and meaningful visual representations of processed data and deliver them to the client for display within SGX-PrivInfer's GUI. As previously described, the GUI serves as a modular and user-friendly front-end interface, offering extensive options for menus, settings, and visualization. It provides users with detailed insights into their digital footprint across various OSNs, their aggregated ego graph, and the sensitive attributes that can be inferred from it, thereby enhancing transparency and empowering user control.

3.5 *SGX-PrivInfer*'s Workflow

We now describe the operational workflow of SGX-PrivInfer, which consists of a four-phase interaction flow: 1) *Groundwork Phase* , 2) *Bootstrapping Phase*, 3) *User Onboarding Phase* and 4) *Secure Inference Analysis Phase*. The workflow is as follows:

Phase 1. Groundwork. The operational workflow begins with the Model Supplier (MS) specifying and provisioning essential, non-privacy-sensitive yet critical software or code to be executed inside the enclave. This software may include, for example, a parameterized Python environment required for running the inference model. To ensure the authenticity and integrity of the provided code, it is digitally signed using the Model Supplier's private key. This allows both the users and the enclave to verify the origin and integrity of the code before execution. However, users may require additional assurances regarding the trust-worthiness of such critical code, particularly concerning potential risks like data leakage or the presence of backdoors. While these guarantees could be obtained through approaches such as crowdsourced security vetting processes, addressing such processes lies beyond the scope of this paper.

Phase 2. Bootstrapping. As part of the Boot-strapping Phase, the enclave is created, loaded, and initialized using the SGX-enabled server CPU and the code provisioned earlier by the Model Supplier (MS). Specifically, the enclave is instantiated from the supplier's enclave code, which is copied into memory. During this process, a cryptographic hash of the initial memory content is generated and securely stored. Leveraging the hardware security features of the SGX-enabled CPU and the platform certificate issued by the device vendor, the creation process ensures both the confidentiality and integrity of the enclave code and its memory content. Once the enclave creation is complete, it is initialized, resulting in the generation of an asymmetric key pair for the enclave. The secret key is used, among other functions, to sign a proof of attestation. This proof serves as verifiable evidence to assure users and the Model Supplier that the correct, untampered enclave code is running on the SGX-PrivInfer server. The proof of attestation includes the cryptographic hash of the enclave's initial memory content and the enclave's public key. If needed, this proof is

securely transmitted to users and the Model Supplier through TLS-like secure channels, each originating directly from within the enclave. This process ensures that all entities involved can trust the integrity of the enclave and its operations. The final step of the Bootstrapping Phase involves the Model Supplier (MS) delivering its attribute inference model (Simo and Kreutzer, 2024), along with functions for ego graph generation and ego graph aggregation, to the SGX-PrivInfer server. Since this input is sensitive and constitutes the MS's intellectual property, it must be encrypted to prevent unauthorized access or misuse, particularly by the untrusted Service Provider (SP). In the current design, the sensitive input from the MS is encrypted using a symmetric encryption key, which is securely derived from the enclave's public key and a nonce. The encrypted data is then transmitted directly to the server through a secure communication channel. To address the constraints imposed by the limited enclave memory size, our solution allows the enclave to offload and store the Model Supplier's (MS) encrypted data in the untrusted part of the SGX-PrivInfer server. This design ensures that while the encrypted data resides in untrusted storage, it remains fully protected, as only the enclave can decrypt and access it when necessary. When needed, such as during inference analysis, the encrypted model can be securely retrieved from the untrusted storage and decrypted inside the enclave using the previously mentioned symmetric encryption key. To securely establish this key, the Model Supplier encrypts the symmetric encryption key with the enclave's public key and transmits it to the enclave. Upon receipt, the enclave decrypts the message, ensuring that both the MS and the enclave share the same symmetric key. Because the symmetric key is derived using a unique nonce n, the Model Supplier retains full control over who can access and use its inference model. This mechanism guarantees that only authorized enclaves that meet the attestation requirements can utilize the model, thereby protecting the MS's intellectual property and ensuring secure, controlled access to the inference process.

Phase 3. User Onboarding. In this phase, each user registers with the platform using a dedicated client application, the SGX-PrivInfer Client Application, which enables the collection of ego graphs from the OSN platforms where the user is registered. The collected data is securely uploaded to the analytics server, though at this stage it is not yet loaded inside the enclave. A key component of the registration and sign-in process is Server Attestation and Key Provisioning, which relies on the remote attestation feature of Intel SGX (Barbosa et al., 2016). During this step, the client application verifies the trustworthiness of the analytics server's enclave by checking that it is authentic and has not been tampered with. Upon successful attestation, the client and the enclave establish a secure communication channel, and the client securely provisions a symmetric key directly to the enclave. This symmetric key is persistently stored within the enclave for exclusive use in secure operations. The symmetric key serves multiple purposes. For example, the client application encrypts all input data, such as the user's ego graphs, before forwarding it to the server. The data remains encrypted and can only be decrypted inside the enclave. Additionally, the symmetric key is used to encrypt sensitive data that needs to be securely transmitted and made accessible only to the respective user. After successfully joining the SGX-PrivInfer platform, users can use the client application to create groups and invite their contacts on OSNs to participate. Contacts who accept the invitation are redirected to a website where they can download the client application. They then use the app to collect their individual public graph data from various social networks and upload it to the SGX-PrivInfer server for further secure analysis. This collaborative workflow ensures a seamless aggregation of ego graph data from multiple users while maintaining strong privacy and security guarantees.

Phase 4. Secure Inference Analysis. A registered user who wishes to analyze her data and assess her exposure to attribute inference risk can query the SGX-PrivInfer server by submitting the encrypted values of hidden attributes as key parameters. These inputs are encrypted using the user's previously established symmetric key, ensuring that only the enclave can access and process this sensitive information. Inside the enclave, our solution reconstructs the querier's aggregated ego-graph by combining all group members' public data. This aggregated ego-graph is then combined with the querier's private data to compute the probabilities of correctly inferring the values of each specified hidden attribute. To construct the aggregated ego-graph, the enclave securely loads the Model Supplier's models and functions from the untrusted part of the server, decrypts them, and forwards the data to the Inference Engine. It is important to note that aggregated ego-graphs can be

generated either prior to an inference analysis request or on-demand, depending on the system's configuration and user needs. To further protect against access pattern leakage, we incorporate techniques to randomize any observable information during the process of loading models and ego-graphs into the enclave, as outlined in (Chandra et al., 2017). This mechanism mitigates the risk of an attacker inferring sensitive details based on access patterns. Once the analysis is complete, the results—including the set of inferred attributes, attribute-specific inference scores, and other voluntarily shared attributes—are encrypted with the user's symmetric key and securely sent back to the SGX-PrivInfer Client Application. To ensure seamless and secure communication, the SGX-PrivInfer client interacts with the analytics server and the respective OSN platforms through REST endpoints exposed over a secure communication channel, achieved using Transport Layer Security (TLS). This guarantees data confidentiality and integrity throughout the interaction workflow.

4 PoC & EVALUATION

To demonstrate the feasibility of our proposed approach, we present a Proof-of-Concept (PoC) implementation of SGX-PrivInfer along with its preliminary evaluation.

4.1 Implementation Details

The current prototype of SGX-PrivInfer is implemented natively using Intel SGX SDK 2.0, primarily in C++, with some C and JavaScript components. Our prototype leverages WAPITI (Simo and Kreutzer, 2024), a weighted Bayesian model tailored for attribute inference in social ego networks, and extends PrivInferVis (Simo et al., 2021) a framework to provide OSN users with enhanced-transparency over attribute inference. We implemented our proposed framework with three key components. An *application server as a Node.js-server*, that stores and processes user account and content data across distinct databases. It serves as backend, handling the communication between users and the secure enclave. An *Angular.js-based client* used to access, manage, and display data and analysis results. This component integrates a *Chrome browser extension* to retrieve profile information directly from the user's OSN accounts.

In addition to incorporating visualization elements from (Simo et al., 2021), the SGX-PrivInfer client application provides essential functionalities, including remote attestation and system initialization. These features are crucial for establishing the necessary keys during the Bootstrapping, User Onboarding, and Secure Analysis phases. Our implementation of the SGX-PrivInfer server, built following Intel SGX developer guidelines, is partitioned into two main components: a trusted enclave and an unprotected server application, enabling secure and efficient operations. Transitions between the trusted and untrusted components are handled through OCALLs and ECALLs. While the server executes as a single binary, the enclave shares the same virtual address space as the unprotected application but maintains strict isolation of its stack and heap. The enclave is designed as a single trusted entity, avoiding inter-enclave communication and thereby reducing the risk of secret leakage. The server consists of three layers: i) unprotected code for sensitive operations managed by SGX's untrusted runtime system (uRTS), ii) trusted code protected by SGX guarantees and managed by the trusted runtime system (tRTS), and iii) EDL files to define transition functions between the two regions. To ensure secure and efficient operations, additional configuration files such as the enclave configuration and the signing key are included, alongside makefiles for building the enclave binary. Key components on the server-side such as the User Authenticator, the Access Controller, the Sealing-Unsealing Component, and the Inference Engine extensively uses the SHA512 algorithm as hash function and the AES-GCM algorithm with a 256-bit key as the symmetric key encryption scheme.

4.2 Evaluation

Now, we report on SGX-PrivInfer's initial performance results.

Testbed Setup & Evaluation Methodology. We evaluated SGX-PrivInfer on a desktop machine equipped with an Intel i5 processor (2.50GHz core), 32 GB of memory and running Ubuntu 16.04 LTS 64-bit along with OpenEnclave (https://github.com/openenclave/openenclave/), an open-source SDK for building enclave-based applications on Intel SGX. To comprehensively evaluate the performance of SGX-PrivInfer, we compared it against PrivInferVis (Simo et al., 2021), which serves as a baseline framework. Unlike SGX-PrivInfer, PrivInferVis does not incorporate enhanced data protection mechanisms on the server side, nor does it

utilize TEEs. In PrivInferVis, ego graph aggregation and attribute inference are performed without the confidentiality and integrity guarantees provided by Intel SGX. Our evaluation primarily focused on quantifying the additional latency introduced by the TEE and associated security mechanisms in SGX-PrivInfer. Specifically, we aimed to measure the runtime overhead incurred by performing computations inside the SGX enclave, compared to PrivInferVis. To this end, we recorded and compared the execution time for private attribute inference on the same dataset in both systems. This approach allows us to precisely determine the performance cost of integrating Intel SGX enclaves. For our experiments, we evaluated attribute inference on gender, leveraging the Weighted Bayesian attribute inference model from (Simo and Kreutzer, 2024, Simo et al., 2021). To ensure a differentiated analysis, we measured execution times in two modes: Mode A and Mode B. In Mode A, users' data is encrypted, allowing us to assess the potential cost of handling additional encrypted data on the server side. In Mode B, users' data is provided in plain text, allowing us to evaluate potential performance overhead in a setting without an additional secure channel between the client and the server (or the enclave). In both modes, execution time was measured from the moment the server received the inference request to the completion of the analysis operation. This setup allowed us to capture the performance impact of Intel SGX-based operations and associated security mechanisms in SGX-PrivInfer.

Evaluation on Real Datasets. Our experiments were conducted using the Kaggle Social Circles dataset (Kaggle, n.d.), a widely utilized benchmark in the literature for social network analysis. The dataset comprises 27,520 ego graphs along with various profile attributes (features) derived from real Facebook users. Each user (ego) in the dataset is associated with a subset of up to 57 attributes, including location, birthday, gender, and other relevant personal details.

Initial Performance Results. For our experiments, we partitioned the dataset into four clusters of selected ego graphs based on their sizes: C1 (31-45), C2 (138-151), C3 (215-238), and C4 (341-357). For example, C1 (31-45) refers to a subset of the dataset containing ego graphs, each with a size ranging from 31 to 45 nodes. All experimental results presented below are averaged over 10 runs to ensure statistical reliability. The results, summarized in Figure 2, demonstrate that SGX-PrivInfer incurs a moderate runtime overhead when compared to PrivInferVis. Specifically, SGX-PrivInfer is, on average, approximately only 1.5 times slower than PrivInferVis in both evaluation modes. Indeed, while SGX-PrivInfer offers enhanced security, it consistently incurs moderate overhead in both modes, with the overhead being slightly higher in Mode A compared to Mode B. Such an overhead, as the cost of securing the attribute inference analysis process with Intel SGX enclaves and associated security mechanisms, is arguably justifiable.

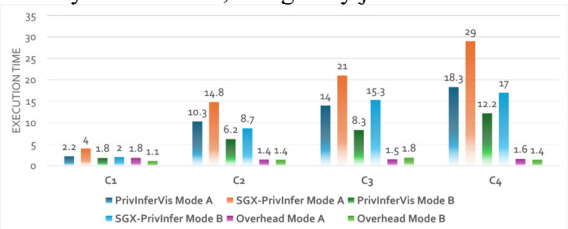

Figure 2: Runtime comparison (time in seconds).

5 CONCLUSION

We presented SGX-PrivInfer, a novel framework for collaborative and privacy-preserving attribute inference risk analysis in online social networks. By leveraging Trusted Execution Environments (TEEs), specifically Intel SGX, SGX-PrivInfer ensures the confidentiality of both user data and attribute inference models, effectively mitigating the risks posed by adversarial or curious actors in cloud environments. SGX-PrivInfer combines hardware-assisted security features with the isolation guarantees of Trusted Execution Environments (TEEs). Its scalable architecture supports real-time analytics while ensuring robust privacy and security protections. Indeed, our framework offers a range of features that collectively enable collaborative and privacy-preserving attribute inference risk analysis, regardless of the entity controlling the server on which the analysis is performed. Key features of SGX-PrivInfer include user and group management, effective ego graph aggregation and weighted Bayesian attribute inference as introduced in (Simo et al., 2021), and server attestation based on Intel SGX's remote attestation feature. Our preliminary evaluation demonstrates the feasibility of SGX-PrivInfer and its practical performance on real datasets, showing minimal overhead compared to privacy inference detection technologies without TEEs. While this evaluation validates the core functionality and architectural design of SGX-

PrivInfer, it also highlights several areas for future exploration.

In future work, we plan to enhance the design of SGX-PrivInfer by exploring alternative hardware-based security technologies and isolation mechanisms beyond SGX. This includes evaluating other TEEs, such as secure Linux containers (e.g., Scone (Arnautov et al., 2016) and LKL-SGX (Priebe et al., 2019)), to assess their suitability and performance for privacy-preserving attribute inference tasks. Additionally, we aim to expand our evaluation of SGX-PrivInfer to account for potential end-to-end network overheads in distributed setups, as these could impact real-world deployments. Future efforts will also focus on designing and conducting a user study to investigate user intentions, perceived usefulness, and attitudes toward SGX-PrivInfer.

ACKNOWLEDGEMENTS

This work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

REFERENCES

Aghasian, E., Garg, S., Gao, L., Yu, S., and Montgomery, J. (2017). Scoring users' privacy disclosure across multiple online social networks. *IEEE Access*, 5:13118– 13130.

Andreou, A., Venkatadri, G., Goga, O., Gummadi, K. P., Loiseau, P., and Mislove, A. (2018). Investigating ad transparency mechanisms in social media: A case study of facebook's explanations. In *NDSS 2018-Network and distributed system security symposium*, pages 1–15.

Arnautov, S., Trach, B., Gregor, F., Knauth, T., Martin, A., Priebe, C., Lind, J., Muthukumaran, D., O'keeffe, D., Stillwell, M. L., et al. (2016). SCONE: Secure linux containers with intel SGX. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 689–703.

Barbosa, M., Portela, B., Scerri, G., and Warinschi, B. (2016). Foundations of hardware-based attested computation and application to sgx. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 245–260. IEEE.

Cai, Z., He, Z., Guan, X., and Li, Y. (2016). Collective data-sanitization for preventing sensitive information inference attacks in social networks. *IEEE Transactions on Dependable and Secure Computing*, 15(4):577–590.

Chakrabarti, S., Dom, B., and Indyk, P. (1998). Enhanced hypertext categorization using hyperlinks. *Acm Sigmod Record*, 27(2):307–318.

Chamberlain, B. P., Humby, C., and Deisenroth, M. P. (2017). Probabilistic inference of twitter users' age based on what they follow. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 191–203. Springer.

Chandra, S., Karande, V., Lin, Z., Khan, L., Kantarcioglu, M., and Thuraisingham, B. (2017). Securing data analytics on sgx with randomization. In *Computer Security–ESORICS 2017: 22nd European Symposium on Research in Computer Security, Oslo, Norway, September 11-15, 2017, Proceedings, Part I 22*, pages 352–369. Springer.

Chu, W.-T. and Chiu, C.-H. (2014). Predicting occupation from single facial images. In *2014 IEEE International Symposium on Multimedia*, pages 9–12. IEEE.

Ciot, M., Sonderegger, M., and Ruths, D. (2013). Gender inference of twitter users in non-english contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145.

Collins, S., Sun, Y., Kosinski, M., Stillwell, D., and Markuzon, N. (2015). Are you satisfied with life?: Predicting satisfaction with life from facebook. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 24–33. Springer.

De, S. J. and Imine, A. (2017). Privacy scoring of social network user profiles through risk analysis. In *International Conference on Risks and Security of Internet and Systems*, pages 227–243. Springer.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Fang, Q., Sang, J., Xu, C., and Hossain, M. S. (2015). Relational user attribute inference in social media. *IEEE Transactions on Multimedia*, 17(7):1031–1044.

Fei, S., Yan, Z., Ding, W., and Xie, H. (2021). Security vulnerabilities of sgx and countermeasures: A survey. *ACM Computing Surveys (CSUR)*, 54(6):1–36.

Ferwerda, B. and Tkalcic, M. (2018). You are what you post: What the content of instagram pictures tells about users' personality. In *The 23rd International on Intelligent User Interfaces, March 7-11, Tokyo, Japan.*

Fixman, M., Berenstein, A., Brea, J., Minnoni, M., Travizano, M., and Sarraute, C. (2016). A bayesian approach to income inference in a communication network. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 579–582. IEEE.

Gong, N. Z. and Liu, B. (2016). You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors. In *25th {USENIX} Security Symposium*, pages 979–995.

Gong, N. Z., Talwalkar, A., Mackey, L., Huang, L., Shin, E. C. R., Stefanov, E., Shi, E. R., and Song, D. (2014). Joint link prediction and attribute inference using a social-attribute network. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(2):27.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Guha, S., Tang, K., and Francis, P. (2008). Noyb: Privacy in online social networks. In *Proceedings of the first workshop on Online social networks*, pages 49–54.

He, J., Chu, W. W., and Liu, Z. (2006). Inferring privacy information from social networks. In *International Conference on Intelligence and Security Informatics*, pages 154–165. Springer.

Idan, L. and Feigenbaum, J. (2019). Show me your friends, and i will tell you whom you vote for: Predicting voting behavior in social networks. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 816–824.

Irani, D., Webb, S., Li, K., and Pu, C. (2009). Large online social footprints–an emerging threat. In *2009 International Conference on Computational Science and Engineering*, volume 3, pages 271–276. IEEE.

Janic, M., Wijbenga, J. P., and Veugen, T. (2013). Transparency enhancing tools (tets): an overview. In *2013 Third Workshop on Socio-Technical Aspects in Security and Trust*, pages 18–25. IEEE.

Jia, J. and Gong, N. Z. (2018). Attriguard: A practical defense against attribute inference attacks via adversarial machine learning. In *27th {USENIX} Security Symposium*, pages 513–529.

Joo, J., Steen, F. F., and Zhu, S.-C. (2015). Automated facial trait judgment and election outcome prediction: Social dimensions of face. In *Proceedings of the IEEE international conference on computer vision*, pages 3712–3720.

Joon Oh, S., Benenson, R., Fritz, M., and Schiele, B. (2015). Person recognition in personal photo collections. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3862–3870.

Jurgens, D. (2013). That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Kaggle. Learning social circles in networks. https://www.kaggle.com/c/learning-social-circles.

Kapritsos, M., Wang, Y., Quema, V., Clement, A., Alvisi, L., and Dahlin, M. (2012). All about eve:{Execute-Verify} replication for {Multi-Core} servers. In *10th USENIX Symposium on Operating Systems Design and Implementation (OSDI 12)*, pages 237–250.

Knauth, T., Steiner, M., Chakrabarti, S., Lei, L., Xing, C., and Vij, M. (2018). Integrating remote attestation with transport layer security. *arXiv preprint arXiv:1801.05863*.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805.

Labitzke, S., Werling, F., Mittag, J., and Hartenstein, H. (2013). Do online social network friends still threaten my privacy? In *Proceedings of the third ACM conference on Data and application security and privacy*, pages 13–24. ACM.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Li, M., Yang, Y., Chen, G., Yan, M., and Zhang, Y. (2024). Sok: Understanding design choices and pitfalls of trusted execution environments. In *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, pages 1600–1616.

Li, R., Wang, S., Deng, H., Wang, R., and Chang, K. C.-C. (2012). Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1023–1031. ACM.

Luo, S., Morone, F., Sarraute, C., Travizano, M., and Makse, H. A. (2017). Inferring personal economic status from social network location. *Nature communications*, 8:15227.

Maheshwari, S. (2019). Facebook advertising profiles are a mystery to most users, survey says.

McPherson, R., Shokri, R., and Shmatikov, V. (2016). Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*.

Mislove, A., Viswanath, B., Gummadi, K. P., and Druschel, P. (2010). You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM.

Murmann, P. and Fischer-Hübner, S. (2017). Tools for achieving usable ex post transparency: a survey. *IEEE Access*, 5:22965–22991.

Nilsson, A., Bideh, P. N., and Brorsson, J. (2020). A survey of published attacks on intel sgx. *arXiv preprint arXiv:2006.13598*.

Oh, S. J., Benenson, R., Fritz, M., and Schiele, B. (2016). Faceless person recognition: Privacy implications in social media. In *European Conference on Computer Vision*, pages 19–35. Springer.

Oh, S. J., Fritz, M., and Schiele, B. (2017). Adversarial image perturbation for privacy protection a game theory perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1491–1500. IEEE.

Paju, A., Javed, M. O., Nurmi, J., Savima¨ki, J., McGillion, B., and Brumley, B. B. (2023). Sok: A systematic review of tee usage for developing trusted applications. *arXiv preprint arXiv:2306.15025*.

Pontes, T., Magno, G., Vasconcelos, M., Gupta, A., Almeida, J., Kumaraguru, P., and Almeida, V. (2012). Beware of what you share: Inferring home location in social networks. In *2012 IEEE 12th International*

Conference on Data Mining Workshops, pages 571–578. IEEE.

Priebe, C., Muthukumaran, D., Lind, J., Zhu, H., Cui, S., Sartakov, V. A., and Pietzuch, P. (2019). Sgx-lkl: Securing the host os interface for trusted execution. *arXiv preprint arXiv:1908.11143*.

Qu, Y., Xing, L., Ma, H., Wu, H., Zhang, K., and Deng, K. (2022). Exploiting user friendship networks for user identification across social networks. *Symmetry*, 14(1):110.

Rangel, F., Rosso, P., Montes-y Go´mez, M., Potthast, M., and Stein, B. (2018). Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. *Working notes papers of the CLEF*, 192.

Ryu, E., Rong, Y., Li, J., and Machanavajjhala, A. (2013). Curso: protect yourself from curse of attribute inference: a social network privacy-analyzer. In *Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks*, pages 13–18.

Shao, M., Li, L., and Fu, Y. (2013). What do you do? occupation recognition in a photo via social context. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3631–3638.

Simo, H. and Kreutzer, M. (2024). WAPITI - a weighted bayesian method for private information inference on social ego networks. In *The 23rd IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*.

Simo, H., Shulman, H., Schufrin, M., Reynolds, S. L., and Kohlhammer, J. (2021). PrivInferVis: Towards enhancing transparency over attribute inference in online social networks. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1–2.

Sun, Q., Schiele, B., and Fritz, M. (2017). A domain based approach to social relation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3481–3490.

Sun, X., Wu, P., and Hoi, S. C. (2018). Face detection using deep learning: An improved faster rcnn approach. *Neurocomputing*, 299:42–50.

Talukder, N., Ouzzani, M., Elmagarmid, A. K., Elmeleegy, H., and Yakout, M. (2010). Privometer: Privacy protection in social networks. In *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, pages 266–269. IEEE.

Volkova, S. and Bachrach, Y. (2015). On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsychology, Behavior, and Social Networking*, 18(12):726–736.

Volkova, S., Coppersmith, G., and Van Durme, B. (2014). Inferring user political preferences from streaming communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 186–196.

Wang, G., Gallagher, A., Luo, J., and Forsyth, D. (2010). Seeing people in social context: Recognizing people and social relationships. In *European conference on computer vision*, pages 169–182. Springer.

Wang, Y. and Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246.

Xiang, L., Sang, J., and Xu, C. (2017). Demographic attribute inference from social multimedia behaviors: A cross-osn approach. In Amsaleg, L., Gumundsson, G. , Gurrin, C., Jo´nsson, B. , and Satoh, S., editors, *MultiMedia Modeling*, pages 515–526, Cham. Springer International Publishing.

Zhang, X., Zhang, L., and Gu, C. (2017). Security risk estimation of social network privacy issue. In *Proceedings of the 2017 the 7th International Conference on Communication and Network Security*, ICCNS 2017, page 81–85, New York, NY, USA. Association for Computing Machinery.

Zheleva, E. and Getoor, L. (2009). To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th international conference on World wide web*, pages 531–540.

Zhong, Y., Yuan, N. J., Zhong, W., Zhang, F., and Xie, X. (2015). You are where you go: Inferring demographic attributes from location check-ins. In *Proceedings of the eighth ACM international conference on web search and data mining*, pages 295–304. ACM.