INTERNATIONAL JOURNAL OF INFORMATION SECURITY SCIENCE
Y. Kilic. ,Vol.14, No.2, pp.1-17.
https://doi.org/10.55859/ijiss.1599063

A Supervised Evaluation Framework for Privacy Risk Scoring Models

Yasir Kilic

Adana Alparslan Turkes Science and Technology University,
Faculty of Engineering, Department of Computer Engineering, ADANA, TURKEY.
e-mail: ykilic@atu.edu.tr

Research Paper Received: 10.12.2024 Revised: 24.02.2025 Accepted: 24.04.2025

Abstract—The rise of online social networks (OSNs) has heightened concerns regarding user privacy, as sensitive attributes disclosed on profiles are increasingly susceptible to misuse, including identity theft and targeted manipulation. Each user's privacy risk varies based on the nature of the shared data and its intended audience. To quantify these risks, researchers have introduced privacy risk scores, inspired by credit scoring systems, to measure vulnerability to privacy violations. However, despite the proliferation of scoring models, their evaluation frameworks often rely on unsupervised methods, such as goodness-of-fit tests, which limit their practical reliability. To address this limitation, this study proposes a supervised evaluation framework named SPR-EVAL, which systematically assesses the performance of privacy scoring models using various real-world attack scenarios, offering a more robust and actionable approach to privacy risk assessment. SPR-EVAL integrates simulations of diverse real-world privacy attacks as a core evaluation mechanism. The framework is adaptable to any OSN dataset and supports the incorporation of various privacy risk scoring models and attack strategies.Extensive experiments on a real-world Facebook OSN dataset demonstrate the effectiveness of SPR-EVAL in evaluating and comparing popular privacy scoring models under supervised conditions. By providing a rigorous supervised evaluation metric, SPR-EVAL addresses the limitations of traditional unsupervised methods, representing a significant advancement in privacy risk assessment for OSNs.

Keywords—privacy risk scoring, online social networks (OSNs), privacy, attribute disclosure attacks, point multiserial correlation (PMS)

1. Introduction

In today's interconnected world, Online Social Networks (OSNs) such as Facebook, LinkedIn, and Instagram have become central to modern social interactions and activities [1]. These platforms offer users a digital representation of their lives, providing diverse opportunities to connect with others, share content, and build relationships.

Many users share a broad spectrum of personal information on online social networks (OSNs), including their name, location, contact details, interests, and daily activities. While this sharing facilitates social connections and digital engagement, it also introduces significant privacy risks. These risks include identity theft, where malicious actors exploit personal information for fraudulent activities [2], [3], and targeted advertising, in which user data is leveraged for potentially intrusive or manipulative marketing strategies [4], [5]. Additionally, such information may be exploited for phishing attacks,

social engineering, or profiling for surveillance purposes [6], [7]. As OSNs continue to evolve and user engagement increases, the potential for unintended misuse of personal data grows. These concerns underscore the need for robust privacy risk assessment frameworks to systematically evaluate and mitigate such threats, thereby enhancing user data protection.

Moreover, the widespread prevalence of data leakage incidents, such as the Cambridge Analytica scandal involving Facebook [8] and, more recently, the alleged sale of a leaked dataset containing records of 489 million Instagram users [9], highlights growing concerns about the privacy risks associated with sharing content on OSNs. These incidents expose significant vulnerabilities in the digital ecosystem, impacting millions of individuals and raising critical questions about the adequacy of existing privacy safeguards.

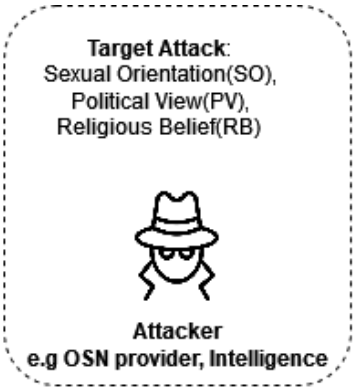
Although the right to privacy is recognized globally through frameworks such as the Universal Declaration of Human Rights and further reinforced by regulations like the European Union's General Data Protection Regulation (GDPR), these protections often prove inadequate when users consent to share personal data on OSNs. This situation underscores the need to educate users about the potential privacy risks and long-term implications of sharing data in digital environments.

It is important to note that each user's privacy risk is inherently unique, determined by the type of data shared and the audience with whom it is shared. Recent studies in this field [10], [11], [12], [13], [14], [15], [16], [17] aim to quantify this risk on an individual level, introducing the concept called *privacy risk score(PRS)*. This concept is inspired by credit risk scores used in financial systems. A higher privacy risk score indicates a greater susceptibility to potential privacy violations.

Despite the availability of numerous PRS methods, the evaluation of these models often lacks robust evaluation,with most relying on unsupervised and subjective techniques such as goodness-of-fit tests and score correlations [10], [11], [18], [19], [20]. These methods are commonly employed because they do not require labeled data or direct user feedback, making them practical for large-scale studies where ground-truth data may be inaccessible. However, while these techniques provide some insight into model performance, they fall short of capturing users' actual privacy concerns. Rather than evaluating how well the model aligns with users' perceptions of risk, these methods primarily focus on statistical performance, such as the model's fit to the data or correlations between outputs. Consequently, they fail to address the subjective nature of privacy risk, underscoring the need for supervised evaluation methods in PRS research.

Consider a potential adversary exploiting Online Social Networks (OSNs), including entities such as intelligence services, advertising firms, or OSN providers with access to user data. As exemplified in Figure 1, this attacker conducts three distinct attribute disclosure attacks (ADAs) to infer —sexual orientation (SO), political view (PV), and religious belief (RB)—against five targeted users, denoted as $U_1, U_2, \ldots, U_5$. The table in Figure 1 presents the attack outcomes for each user alongside their corresponding privacy risk scores, PRS_1 and PRS_2. Upon evaluating the privacy risk scores, PRS_1 demonstrates stronger clustering performance, as evidenced by a superior goodness-of-fit test result, compared to PRS_2. Existing PRS evaluation methods would therefore classify PRS_1 as the superior scoring model. However, incorporating attack success rates into the evaluation alters the conclusion: PRS_2 provides a more

accurate representation of actual privacy risks faced by users. This discrepancy reveals a key limitation of existing evaluation methods, which emphasize unsupervised metrics (e.g., clustering quality) rather than the predictive reliability of privacy risk scores in real-world attack scenarios.



Target Users	Target Attacks: SO	PV	RB	Privacy Scores: PRS_1	PRS_2
u1	Fail	Fail	Unknown	0.9	0.2
u2	Fail	Success	Fail	0.8	0.25
u3	Unknown	Success	Fail	0.7	0.3
u4	Success	Success	Unknown	0.6	0.3
u5	Success	Success	Success	0.5	0.35

Figure 1. A Motivational Scenario

To address these limitations, this study proposes a novel **S**upervised **P**rivacy **R**isk **EVAL**uation framework, named **SPR-EVAL**, which evaluates privacy risk scores based on real-world attack outcomes. In contrast to prior studies that predominantly rely on unsupervised methods, such as clustering or score correlations [10], [16], [19], [21], SPR-EVAL adopts a supervised approach, integrating disclosed attributes to provide a more comprehensive evaluation. By grounding its assessment in actual attack success rates, the framework calculates practical measures of privacy vulnerability, offering a more realistic reflection of privacy risks in real-world scenarios. The inclusion of the attacker's perspective and the application of supervised evaluation techniques make SPR-EVAL a robust and systematic tool for privacy risk assessment.

Moreover, unlike existing studies that rely on synthetic or survey-based OSN data, this study conducts experiments using real-world data collected from the Facebook OSN. This approach enables an analysis grounded in actual user behavior rather than self-reported sharing patterns, which may be biased due to users' interpretations of their own privacy settings and the tendency to provide socially desirable responses, as noted in [22].

Extensive experiments demonstrated the effectiveness of SPR-EVAL by comparing privacy risk scoring (PRS) models under a simulated attack scenario targeting two sensitive user attributes: political views and religious beliefs. Specifically, the results indicate that granularity-aware and Item Response Theory (IRT)-based scoring models provide a more accurate assessment of users' privacy risk perceptions compared to traditional methods. The findings confirm that SPR-EVAL effectively quantifies privacy risks by integrating real-world attack success rates, offering a robust

and reliable evaluation framework for PRS models. By leveraging a supervised evaluation approach, the proposed method overcomes the limitations of unsupervised techniques, ensuring that privacy risk scores more accurately reflect the actual threats faced by users.

To sum up, the main contributions of the study are as follows:

1. **First-of-its-Kind Supervised Evaluation Framework:** To the best of author's knowledge, SPR-EVAL is the first framework that systematically evaluates privacy risk scoring models through a supervised methodology. Unlike prior approaches that rely on unsupervised techniques such as goodness-of-fit tests, SPR-EVAL integrates real-world attack success rates, providing a more practical and actionable evaluation.
2. **Versatile and Adaptable Design:** SPR-EVAL is designed to be compatible with any state-of-the-art privacy risk scoring model and privacy-oriented attack model. It is flexible and can be applied to diverse OSN datasets, making it a widely applicable tool for privacy risk assessment.
3. **Improved Reliability of Privacy Risk Scores:** While the primary focus of SPR-EVAL is on evaluation, experimental results demonstrate that it enhances the reliability of privacy risk scores. By incorporating attack success rates into the evaluation process, the framework ensures that privacy scores better reflect real-world privacy risks.
4. **Evaluation with Real-World OSN Data:** The framework is validated using real-world Facebook OSN data, allowing for an assessment based on actual user behavior rather than self-reported or assumed sharing patterns. This ensures a more accurate representation of privacy risks in practical scenarios.

The remainder of the paper is organized as follows: Section 2 reviews related work, while Section 3 presents preliminary concepts in the field. The proposed framework is outlined in Section 4, followed by a discussion of empirical results from experiments conducted on real-world OSN data in Section 5. Finally, the paper concludes in Section 6.

2. Related Work

Most OSN users remain largely unaware of the privacy risks associated with sharing personal information. While numerous research efforts aim to address this issue, the primary focus has been on raising awareness about privacy vulnerabilities through various methodologies. One widely adopted approach involves quantifying users' privacy risk scores based on the personal information they share, providing a measurable perspective on the potential risks they face.

Numerous privacy risk scoring methods have been proposed in the literature. These methods for online social networks (OSNs) have evolved over time, incorporating diverse computational models to assess user vulnerability. One of the earliest approaches, proposed by Liu and Terzi[10], introduced an Item Response Theory (IRT)-based framework that quantified privacy risk based on attribute sensitivity and visibility. This foundational model inspired subsequent works[11], [16], [23], which either extended or refined its scoring mechanism. For instance, the Privacy-Functionality Score (PFS) model [13] assessed privacy by evaluating the balance between information a user shares and what they can see from others. Other studies focused on alternative formulations, such as the Susceptible-Infectious-

Recovered (SIR) model [18], which drew inspiration from epidemiology to model the spread of information and privacy risks within OSNs. These methods primarily centered on profile-based risk assessment, where privacy exposure was determined by the visibility of structured user attributes such as age, gender, and location.

Beyond profile-based methods, several studies have shifted towards network-aware privacy risk estimation[19], [20]. These models consider the influence of a user's social connections on their privacy risk. For example, graph-theoretic approaches[19] incorporate centrality-based measures to assess the risk propagation through OSN structures, adapting the PageRank algorithm to measure exposure based on friend networks. Similarly, SONET [14] introduced the Privacy Index (PIDX), which defines attribute-to-actor, actor-to-actor, and attribute-to-attribute relationships to quantify privacy vulnerability. Additionally, privacy awareness tools like Friend Inspector[24] employ game-based mechanisms to educate users about potential privacy risks, further bridging the gap between theoretical scoring models and practical privacy risk mitigation.

Despite the numerous privacy scoring models proposed in the literature, a supervised approach to evaluate their effectiveness remains absent. Most existing evaluations rely on unsupervised methods, which, despite their widespread use, fail to resolve contradictions between the privacy risk scores generated by different models. At its core, a robust scoring model should accurately reflect the types of attacks users are likely to encounter, yet current methodologies often fall short in aligning risk assessments with real-world privacy threats.

A limited number of studies have attempted to evaluate privacy scores [10], [11], yet they lack a supervised assessment that directly validates risk estimations against real-world threats. For instance, in [11], an unsupervised hypothesis testing approach is proposed to evaluate privacy scores. However, SPR-EVAL highlights a fundamental bias in such methods, as they primarily assess clustering quality rather than establishing a direct correlation with real-world privacy risks. Unlike these conventional approaches, SPR-EVAL introduces a supervised evaluation framework that leverages actual attack outcomes to assess the reliability of privacy scoring models. By correlating privacy risk scores with an attacker's success rate, SPR-EVAL not only refines risk quantification but also establishes attack success as a critical benchmark for evaluating privacy models, ensuring a more accurate and actionable privacy risk assessment.

Compared to author's recent work [17], the current study introduces several key distinctions. While the former focuses on the development of a novel privacy scoring model incorporating data granularity within OSNs, it evaluates model performance through unsupervised metrics such as clustering consistency and score correlations. In contrast, SPR-EVAL proposes a supervised framework that leverages real-world attack outcomes to directly validate privacy scores. This shift from unsupervised to supervised evaluation marks a significant advancement, as it enables a more actionable and attack-aware assessment of user privacy risks. Although both studies aim to improve privacy risk assesment, SPR-EVAL explicitly addresses the gap in evaluation methodology by using attacker success rates as ground truth, a component absent from the recent work.

3. Preliminaries

Naturally, an OSN dataset is represented as a directed multigraph $G = (V, \mathcal{E})$, where V denotes

the set of vertices representing OSN users, and $\mathcal{E}$ denotes the set of edges representing various types of relationships between these users. In social networks, it is common for users to share multiple types of relationships, such as friendship and followership, as illustrated in Figure 2. This multigraph structure effectively captures the complexity and diversity of interactions within an OSN.

3.1. Profile Attributes

Let A denotes profile attributes that introduce themselves to other users in OSNs as part of their profile such as gender,education history.

According to privacy literature, A may fall into one of three categories :identifying attributes(A^I),quasi-identifying attributes(A^Q) and sensitive attributes(A^S). A^I are attributes, which identify an user uniquely such as phone number,A^Q are set of attributes that identify an user uniquely and A^S is the attribute set that may violate the privacy of the user when disclosed such as sexual orientation,political view. For each user's sensitive attribute value $v.a^s$ for $a^s \in A^S$ can be disclosed by attackers (e.g Advertisement company, Intelligence Agency).

3.2. Attacker

A malicious entity or user that seeks to disclose sensitive user attributes (A^S), such as demographic information, preferences, or private behaviors. These attackers can include OSN providers exploiting user data for profit, advertisement agencies conducting invasive profiling, or third-party actors engaging in unauthorized data harvesting. By leveraging advanced data analytics, machine learning models, or social engineering techniques, these attackers can exploit the vast amount of data shared on OSNs, potentially breaching user privacy and anonymity.

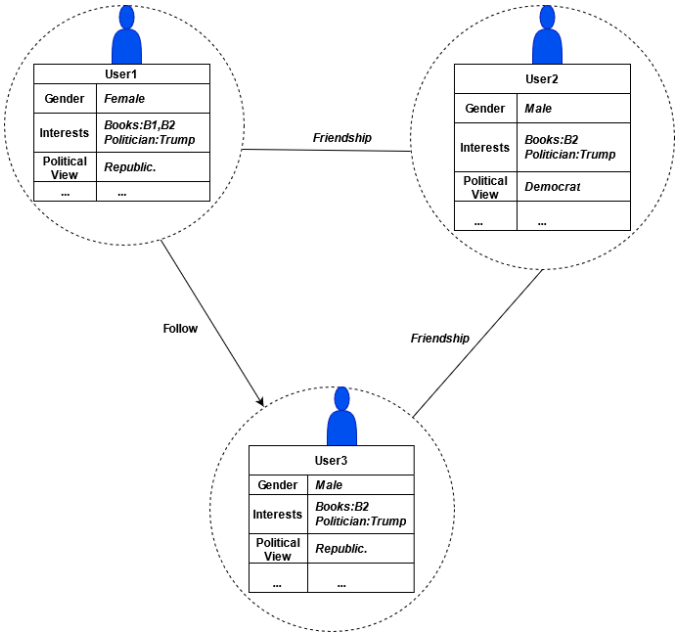


Figure 2. A Toy OSN Data Representation

3.3. Attribute Disclosure Attack (ADA)

Attribute Disclosure Attack (ADA) is one of the most prevalent privacy threats on Online Social Networks (OSNs). This attack occurs when a malicious entity accurately infers the true value of a user's sensitive attribute.

Assume that each sensitive attribute value, denoted as $v.a^s$, for a user v has a probability distribution across all possible values in the sensitive attribute's domain, represented as $dom(a^s)$. ADA aims to predict a value $v.\hat{a}^s$ from this distribution, as defined by Equations 1 and 2. Specifically, $v.\hat{a}^s$ in Equation 1 represents the predicted value inferred by ADA.

In general terms, ADA is formalized in Equation 2. Here, ADA_m denotes a single instance of a disclosure attack model. To achieve optimal results, the most effective model from a predefined attack

set or the best-known model should be employed.

$$\hat{v.a^s} = prediction_{a^s}^{v,ADA_m} \in dom(a^s) \quad (1)$$

$$ADA_m(v) \rightarrow \hat{v.a^s} \quad (2)$$

3.4. Privacy Risk Score (PRS)

Privacy risk scoring, first introduced with formal notations by [11], is the process of assigning scores to users based on their information-sharing behavior. Intuitively, users who disclose more sensitive information with broader visibility are attributed a higher privacy risk score, while those who limit sensitive disclosures receive lower scores.

The model assumes that every user $v \in V$ ($|V| = n$) has a privacy policy governing the sharing of N profile attributes ($|A| = N$), such as *gender* and *location*. These privacy policies are represented as an $n \times N$ policy or response matrix R. The privacy policies in R vary based on the underlying model. For instance, [11] introduced two types of privacy policy matrices: 1. A *dichotomous policy matrix*, which captures binary sharing behavior (shared or not). 2. A *polytomous policy matrix*, which incorporates user privacy settings at varying levels of granularity.

Each cell in R, denoted as $R(a, v)$, represents the user privacy policy v' on the attribute a. A higher value of $R(a, v)$ indicates a more permissive or careless privacy policy for the attribute a, while a lower value signifies a more restrictive or conservative policy.

Within this framework, two primary factors are widely used in PRS models:

i. Sensitivity (β_a): A measure of the importance or criticality of the attribute a from a privacy perspective.

ii. Visibility ($V(a, v)$): A measure of how exposed or accessible the attribute a is to the user v. Visibility can be modeled differently depending on the context, including factors such as network position, privacy settings, or attribute granularity.

The combined privacy risk score for any user v is calculated as the weighted sum of the sensitivity and visibility of all attributes, as shown in Equation 3. This results in a continuous privacy risk score assigned to each user. The range of this score, depending on the model, can range from zero to one, as indicated in Equation 4.

$$PRS_k(v) = \sum_{a \in A} \beta_a \times V(a, v) \quad (3)$$

$$PRS_k\ (v) \rightarrow\ [0, 1] \quad (4)$$

4. Proposed Framework

This section presents the proposed solution, *SPR-EVAL*, designed to evaluate privacy risk scoring (PRS) models within the proposed framework. As illustrated in Figure 3, the proposed solution consists of the following key steps:

- **Data Collection and Attack Simulation:** The solution utilizes an *OSN dataset* as the foundational data source, from which ground truth labels for sensitive attributes (e.g., political view and religious belief) are extracted. *Attribute Disclosure Attacks (ADAs)* are then implemented to infer these sensitive attributes, producing an *Attack Success Report* that records whether each attribute was successfully disclosed, misclassified, or remains unknown.
- **Encoding of Attack Success Levels:** Attack outcomes from the Attack Success Report are transformed into numerical success levels (see Figure 3). This encoding categorizes outcomes

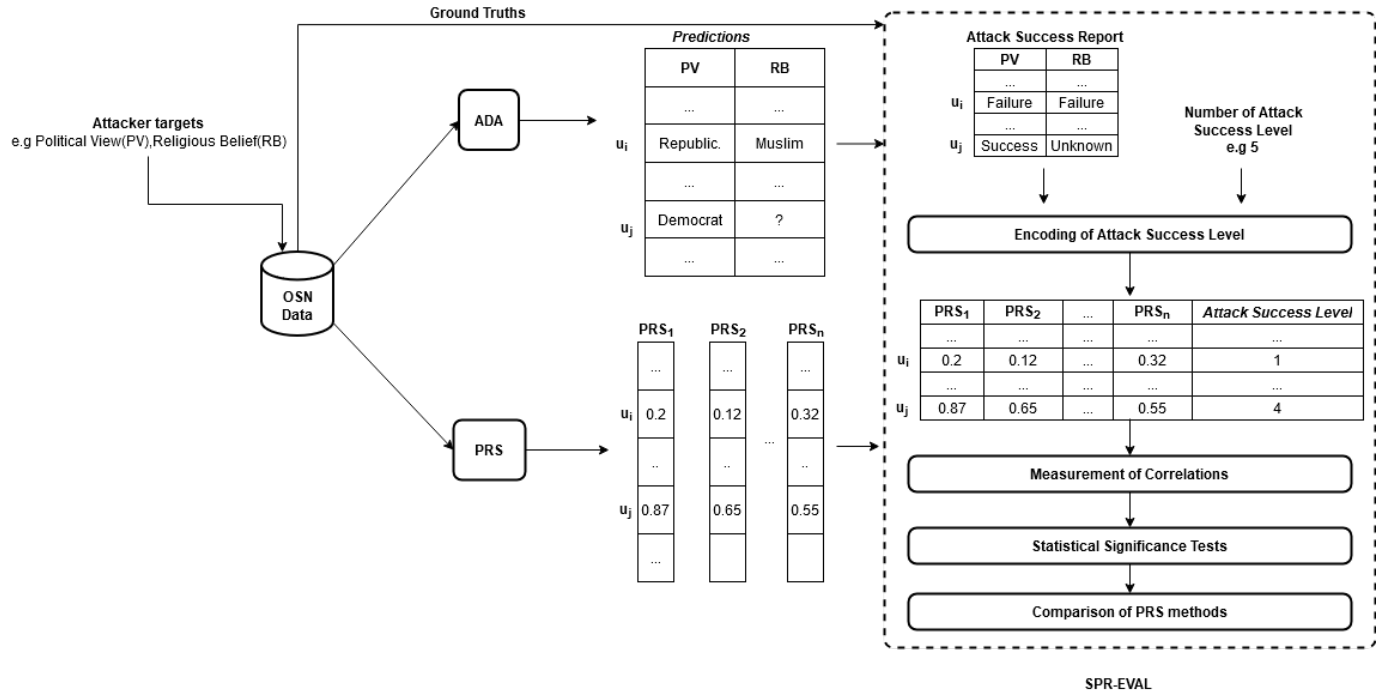


Figure 3. Overview of Proposed Framework

(e.g., success, failure, unknown) into structured levels, enabling a quantitative comparison of PRS models and facilitating a rigorous evaluation of their predictive accuracy.

- **Privacy Risk Scoring (PRS) Computation**: Multiple PRS models are applied to the dataset, assigning privacy risk scores to each user. These scores aim to measure users' vulnerability risk based on the information shared.
- **Correlation Measurement**: To assess the predictive validity of PRS models, the correlation between privacy risk scores and attack success levels is computed. This step evaluates how well each PRS model aligns with real-world attack outcomes.
- **Statistical Significance Testing**: The robustness of the correlations is verified using statistical tests. This ensures that observed relationships between PRS scores and attack outcomes are not due to random variation.
- **Comparison of PRS Methods**: The final step involves comparing different PRS models to determine their effectiveness in quantifying users' actual privacy risks. This analysis provides insights into which models are more reliable for privacy risk assessment in adversarial settings.

The following subsections provide a detailed technical explanation of these steps.

4.1. Encoding of Attacker Success Level(ASL)

$$Succ_v^{a^s} = \begin{cases} UNKNOWN & \text{if } v.a^s \text{ is null} \\ SUCCESS & \text{if } v.a^s = \hat{v.a^s} \\ FAIL & \text{if } v.a^s \neq \hat{v.a^s} \end{cases} \quad (5)$$

As formulated in Eq. 5, the attacker success level(ASL) is determined based on the success of the ADA performed for each user. For each

Table 1.
Three-valued Logic Encoding for Attacker Success State

Attacker Success State	Encoding
FAIL	0
UNKNOWN	1
SUCCESS	2

sensitive attribute a^s, the attacker success states are encoded using three-valued logic as shown in Table 1. Specifically, a partial ordering is defined for the three possible states resulting from the *ADA*. In this ordering, states representing higher levels of attacker success are assigned larger values. The state in which the ground truth value of a sensitive attribute on the OSN is unknown is categorized as *UNKNOWN*. If the sensitive attribute is correctly predicted by the attacker when the true value is known, the state is categorized as *SUCCESS*. Conversely, if the attribute is incorrectly predicted, it is categorized as *FAIL*.

To illustrate this system, consider two sensitive attributes: **Political View (PV)** and **Religious Belief (RB)**. Table 2 demonstrates how the success states can be combined to compute an ASL using the ternary system. Each combination of outcomes is assigned a single decimal value by interpreting the ternary code as a base-3 numeral. For s predefined sensitive attributes, the total number of possible ASLs is 3^s.

Notice that this exponential growth poses a challenge by reducing the statistical significance of results, particularly when sample sizes are limited. To address this issue, SPR-EVAL propose constraining the number of attack success levels to a predefined interval using 1-dimensional clustering [25]. The lower bound ensures sufficient granularity, while the upper bound prevents excessive fragmentation of the data.

Table 2.
An Example for Decimal Attack Success Level(ASL) assignment

Attack on PV	Attack on RB	ASL
Failure	Failure	0
Failure	Unknown	1
Failure	Success	2
Unknown	Failure	3
Unknown	Unknown	4
Unknown	Success	5
Success	Failure	6
Success	Unknown	7
Success	Success	8

4.2. Measuring Statistical Correlations

To assess the relationship between privacy risk scores(PRSs) and attack success levels(ASLs), SPR-EVAL employs a statistical approach utilizing the Point Multiserial Correlation (PMS) [26] coefficient. As outlined in Section 4.1, users are divided into k distinct groups based on their ASLs, which indicate the degree to which sensitive attributes are disclosed in attack scenarios. The PMS coefficient serves as a statistical metric to measure the correlation between a continuous PRS values (described in Section 3.4) and an ordered categorical ASLs (described in Section 4.1). This coefficient provides a quantitative measure of how effectively privacy risk scores predict the extent of attacker success.

4.3. Statistical Significance Tests

Figure 4 illustrates the statistical framework for grouping users based on attack success level (ASL), designed to ensure the validity of conclusions drawn from the data. The process begins with Bartlett's test [27], which evaluates the assumption of equal group variances. Although Bartlett's test is a standard approach for assessing homogeneity of variances, its sensitivity to deviations from

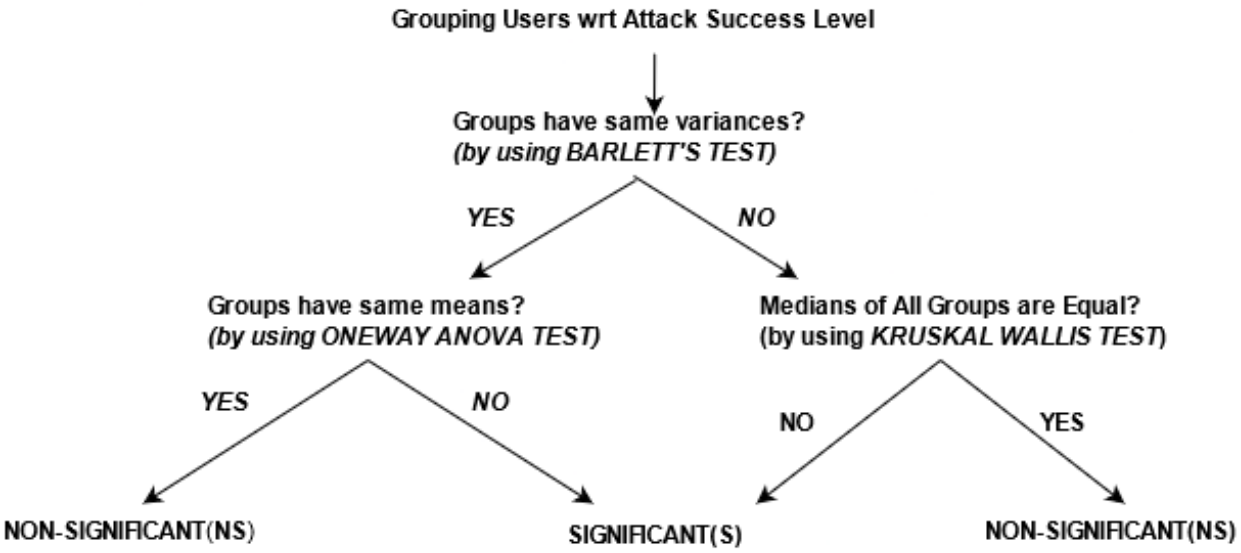


Figure 4. Flowchart Diagram for the Statistical Significance Tests

normality poses challenges when applied to real-world data, which often exhibits non-normal distributions. Despite this limitation, Bartlett's test is employed here due to its ability to identify variance discrepancies that could affect the validity of parametric tests like One-Way ANOVA [28].

If Bartlett's test confirms equal variances, the One-Way ANOVA test is used to determine whether group means differ significantly. A significant result indicates meaningful differences between groups, suggesting that ASL impacts user groupings. If the variances are unequal, a non-parametric alternative is required. In this framework, the Kruskal-Wallis test [29] is selected because it does not rely on the assumption of homogeneity of variances and is effective for comparing group medians, particularly in datasets with non-normal distributions. While other non-parametric tests such as the Welch ANOVA exist, Kruskal-Wallis is preferred here due to its robustness and simplicity in handling ordinal data or ranks.

Significant (S) results from ANOVA or Kruskal-Wallis highlight meaningful differences between groups, while non-significant (NS) results suggest similarities. By aligning test selection with the data's underlying assumptions, the framework delivers reliable and accurate evaluations.

5. Experiments

This section demonstrates the application of the proposed framework over a real-world OSN data. First, we outline the experimental settings for the framework in Section 5.1. Following this, a case study is presented to show the implementation of the framework Section 5.2.

5.1. Experimental Settings

As depicted in Figure 3 and detailed in Section 4, the necessary configurations for the proposed framework are described in the following the sub-headings.

5.1.1 Requirement 1 (OSN Dataset)

The dataset used in this study was collected by the author using a custom-developed web crawler, following the same methodology as in [30]. During the data collection process, profile attributes and liked pages of 20,000 users from Turkey were gathered. Due to privacy restrictions, liked page information was accessible for only 13,023 users, as given in Table 3. Among these users, 134 explicitly disclosed their political views, while 276 openly shared their religious beliefs. The class distributions of these attributes are presented in Table4.

Table 3.
Crawler statistics

Attribute	Number of Users
Total Users Collected	20,000
Users with Accessible Liked Pages	13,023
Users Disclosing Political Views	134
Users Disclosing Religious Beliefs	276

Table 4.
Sensitive Attribute Value Distribution

Political View	Number of Users
Political View 1	79
Political View 2	55
Total	134
Religious Belief	**Number of Users**
Religious Belief 1	262
Religious Belief 2	14
Total	276

To support further research in this domain and facilitate reproducibility, an anonymized version of the dataset is publicly avaliable in the github page[1]. The dataset has been carefully processed to remove personally identifiable information while preserving its analytical utility. The anonymization process involves replacing user-specific identifiers with randomly generated values to prevent direct re-identification. Additionally, while class labels extracted from publicly shared profile items, their interpretability has been deliberately obfuscated. For instance, categorical attributes such as political view, which originally contained explicit labels (e.g., ”Democrat”), have been anonymized to generic representations (e.g., ”political view=Political View 3”).

In particular, to address the risk of revealing sensitive information , the labels used in the dataset were derived primarily from self-declared, publicly visible user profile fields. For political view and religious belief fields,the annotation was performed on publicly shared free-text content, but only when explicit and unambiguous indicators were present. For instance, references to political figures or parties such as “Recep Tayyip Erdoğan,” “Kemal Kılıçdaroğlu,” or “AK Parti” were used to infer political alignment. Similarly, mentions of religious identifiers like “Islam,” “Christianity,” “Atheist,” or religious practices such as “Ramadan” or “church” were used to infer religious belief. No third-party inference or ambiguous interpretation was applied. The disclosure status of a sensitive attribute was determined based on whether the corresponding field was completed in the user’s public profile or contained these clear indicators. This approach ensures that no additional sensitive information was exposed or inferred beyond what users had already made publicly available on the OSN. The classification of fields as sensitive (e.g. political view, religious belief) aligns with established categorizations in the privacy literature and is consistent with privacy literature.

1. https://github.com/Snooper14/TR-FB-Dataset

5.1.2 Requirement 2 (Definition of the target of the attacker)

In this context, the attack scenario is defined based on the sensitive attributes available in the selected dataset, specifically the political view (PV) and religious belief (RB). These attributes were chosen to simulate attribute disclosure attacks (ADAs) because they represent highly sensitive information. Such attributes are frequently targeted in privacy breaches such as the Cambridge Analytica scandal [8], making them ideal for evaluating the effectiveness of the framework in assessing privacy risks.

5.1.3 Requirement 3 (Simulation of Attacks)

As technically described in Section 4.1, determining the attack success result for each user under a sensitive attribute inference attack requires identifying the most effective attack model.

To select the best-performing attack model, commonly used performance metrics named F1-score,accuracy,sensitivity and specificity are employed. In these metrics, The F1-score is the harmonic mean of precision and recall, providing a balance between the two:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

where

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

and

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

Here, TP denotes true positives, FP false positives, and FN false negatives.

Accuracy, a broader measure of overall correctness, is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

where TN represents true negatives. Accuracy provides an overall measure of how well the classifier performs across all classes but may be misleading in imbalanced datasets.

Sensitivity, also known as the true positive rate, measures the proportion of actual positive cases correctly identified:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{10}$$

This metric reflects the model's ability to capture positive instances.

Specificity, or the true negative rate, measures the proportion of actual negative cases correctly identified:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{11}$$

It indicates how effectively the method avoids false alarms by correctly rejecting negative instances.

To simulate attribute disclosure attacks (ADAs), various state-of-the-art attacker models are implemented. These models are designed to represent the most successful attacker for each scenario. Specifically, the following models are employed:

- unFriendly-WC [31] infers users' sensitive attributes by analyzing publicly exposed wall posts in OSNs. It first collects wall messages leaked due to asymmetric privacy settings between friends. Then, it aggregates all posts associated with a user, treating them as a document for text analysis. Next, it applies term frequency–inverse document frequency (TF-IDF) weighting to convert the text into feature vectors. Finally, a multinomial logistic regression classifier predicts sensitive user attributes, such as political affiliation, based on linguistic patterns and topic distributions in the exposed conversations.

- LDA-IFV [32] infers users' sensitive attributes based on their interests. It first augments user interest names with Wikipedia-derived semantic categories for enhanced feature representation. Then, it applies Latent Dirichlet Allocation (LDA) to extract topic distributions and compute the probability of each interest belonging to a topic. Next, it constructs Interest Feature Vectors (IFVs) by representing users based on their interest topic distributions. To identify similar users, it employs chi-squared distance to find nearest neighbors. Finally, it predicts sensitive attributes through majority voting among the nearest neighbors.
- SVD-LR [33] infers users' sensitive attributes based on their liked pages in OSNs. It first applies singular value decomposition (SVD) to reduce the dimensionality of the user-like matrix and then employs a logistic regression(LR) classifier to predict sensitive attributes.

Table 5.
Experimental results for targeted attack performance

(a)
Political View

Method	Acc.	F1-score	Spec.	Sens.
unFriendly-WC	0.99	0.99	0.99	0.99
LDA-IFV	0.99	0.99	0.99	0.99
SVD-LR	0.537	0.347	0.5	0.5

(b)
Religious Belief

Method	Acc.	F1-score	Spec.	Sens.
unFriendly-WC	0.989	0.941	0.994	0.994
LDA-IFV	0.985	0.918	0.875	0.875
SVD-LR	0.347	0.193	0.5	0.5

The attack models are implemented by adapting them to the structure of the dataset introduced in Section 5.1.1. To enhance the robustness of the results, cross-validation is employed during implementation. The obtained results are outlined in Table 5.1.3, where a comprehensive analysis indicates that the unFriendly-WC model achieves the highest performance in inferring both political view and religious belief attributes. Based on this observation, unFriendly-WC is selected as the reference attack model for subsequent experiments to simulate users' exposure to privacy-based attacks.

5.1.4 Requirement 4(Privacy Risk Scoring Models)

In this part, the scoring framework introduced in Section 3.4 is utilized to populate the response matrix $R(v, a)$ with user-shared attributes from the OSN. Specifically, four scoring models are implemented: PSN [10] and PSI [11], which represent foundational frameworks for privacy risk assessment, and $PSGN$ [17] and $PSGI$ [17] introduce the latest methodologies, incorporating sharing granularities and polytomous Item Response Theory (IRT) [34], respectively.

Specifically, the PSN model [10], as defined in Eq. 3, computes privacy risks based on the probability of attribute sharing, while PSI [11] builds on this approach by leveraging Item Response Theory(IRT) to account for probabilistic variations in user behavior. $PSGN$ [17] extends these models by incorporating sharing granularities, providing a more detailed representation of how users disclose attributes. $PSGI$ further enhances this approach by adapting polytomous IRT to capture multi-level attribute sharing dynamics [17].

The inclusion of these models serves a dual purpose. PSN and PSI offer a benchmark for evaluating traditional privacy scoring methods,

while $PSGN$ and $PSGI$ provide insight into how state-of-the-art techniques address the limitations of earlier frameworks. By employing these models, the analysis not only evaluates existing methodologies but also investigates the potential advancements in privacy risk scoring.

5.2. A Case Study on A Real-world Facebook Dataset

This section evaluates SPR-EVAL's effectiveness in identifying the optimal scoring method for modeling privacy risk perceptions. Leveraging a real-world OSN dataset, the analysis examines the relationship between privacy risk scores and attack outcomes under the experimental settings outlined in Section 5.1. This evaluation provides insights into how well the scoring methods capture the complexities of privacy risks associated with user behavior and susceptibility to attacks.

The analysis begins by examining the relationship between discretized attack success levels for each user and privacy risk scores, with a focus on varying the number of attack success levels. Following this, Table 6 presents detailed statistical test results as described in Section 4.3, while Figure 5 illustrates the magnitude of these correlations. Predefining the number of attack success levels is crucial for ensuring the stability of the results. Notably, the data provides insufficient evidence to support the hypothesis for grouping users into two or three levels. However, a negative correlation is observed in Naive models (PSN, $PSGN$), which produces statistically significant results. This finding highlights a mismatch between the structure of these models and the actual risk perceptions of users, suggesting that simpler grouping strategies fail to capture meaningful distinctions among privacy risk levels.

When four or more attack success levels are defined, the statistical results achieve significance, indicating that risk perceptions are better modeled with a finer granularity. Both Figure 5 and the ρ entries in Table 6 provide complementary illustrations of this observation, showing the statistical correlation's magnitude as the number of predefined attack success levels increases. For lower numbers of levels, Naive models (PSN, $PSGN$) exhibit a negative correlation with attack success, underscoring their inability to accurately measure user risk perceptions. Conversely, other models fail to achieve statistical significance under these conditions, further supporting the inadequacy of coarse grouping.

A broader analysis of Figure 5 reveals that IRT-based models (PSI, $PSGI$) outperform Naive models (PSN, $PSGN$) in modeling attack success. This advantage is attributed to IRT-based models' ability to account for user sharing behaviors as demonstrated before in the studies [10], [11], [17]. Additionally, Granularity-based mechanisms ($PSGN$, $PSGI$) demonstrate further improvements over baseline mechanisms like those proposed by Liu et al. [11], (PSN, PSI). By incorporating profile-based information, these granularity-aware approaches refine the risk assessment process, suggesting that they are more suitable for accurately modeling user risk perceptions.

6. Conclusion and Future Work

In conclusion, this work introduces a framework for evaluating privacy risk scoring mechanisms under the supervision of real-world privacy attacks. Unlike existing methods that often rely on hypothetical analysis, the proposed framework employs real-world attack scenarios to assess the effectiveness of various privacy risk models. This approach provides a more realistic and actionable evaluation of how well these models capture

Table 6.
Statistical Correlation Results between Attack Success Levels and Privacy Risk Scoring Mechanisms at Significance Level 0.01

# Attack Success Levels	Metric	PSN	PSI	PSGN	PSGI
8	ρ	0.046	0.132	0.058	0.141
	BT(p-val)	<0.01	<0.01	<0.01	<0.01
	ANOVA(p-val)	-	-	-	-
	KW(p-val)	<0.01	<0.01	<0.01	<0.01
	Significance Test?	Accept	Accept	Accept	Accept
7	ρ	-0.033	0.103	-0.01	0.131
	BT(p-val)	<0.01	<0.01	<0.01	<0.01
	ANOVA(p-val)	-	-	-	-
	KW(p-val)	<0.01	<0.01	<0.01	<0.01
	Significance Test?	Accept	Accept	Accept	Accept
6	ρ	0.002	0.137	0.013	0.15
	BT(p-val)	<0.01	<0.01	<0.01	<0.01
	ANOVA(p-val)	-	-	-	-
	KW(p-val)	<0.01	<0.01	<0.01	<0.01
	Significance Test?	Accept	Accept	Accept	Accept
5	ρ	-0.137	0.059	-0.099	0.096
	BT(p-val)	<0.01	<0.01	<0.01	<0.01
	ANOVA(p-val)	-	-	-	-
	KW(p-val)	<0.01	<0.01	<0.01	<0.01
	Significance Test?	Accept	Accept	Accept	Accept
4	ρ	-0.039	0.09	-0.013	0.122
	BT(p-val)	<0.01	<0.01	<0.01	<0.01
	ANOVA(p-val)	-	-	-	-
	KW(p-val)	<0.01	<0.01	<0.01	<0.01
	Significance Test?	Accept	Accept	Accept	Accept
3	ρ	0.018	0.069	0.025	0.07
	BT(p-val)	<0.01	<0.01	<0.01	<0.01
	ANOVA(p-val)	-	-	-	-
	KW(p-val)	<0.01	0.255	<0.01	0.487
	Significance Test?	Accept	**Reject**	Accept	**Reject**
2	ρ	-0.291	-0.073	-0.235	-0.011
	BT(p-val)	0.278	0.902	0.127	0.934
	ANOVA(p-val)	<0.01	0.339	<0.01	0.881
	KW(p-val)	-	-	-	-
	Significance Test?	Accept	**Reject**	Accept	**Reject**

users' actual risk perceptions. By incorporating attack success levels across multiple sensitive attributes, the framework enables a more nuanced and comprehensive assessment compared to prior studies in the field.

The proposed framework has the potential to greatly benefit organizations handling large-scale OSN data, such as OSN providers. It offers a mechanism through which these organizations could assess the privacy risks faced by their users and issue timely warnings when necessary. This contribution could significantly enhance privacy management practices, ultimately improving the protection of users' sensitive information.

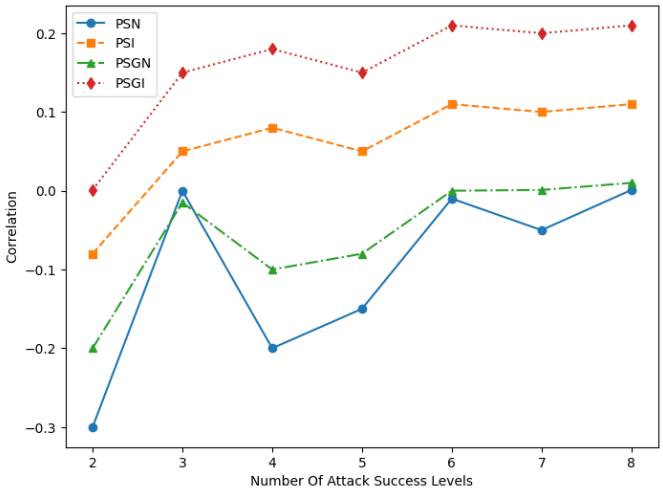


Figure 5. Statistical Correlation With Respect To Number Of Attack Success Levels

Looking ahead, there are several directions for future work. One notable limitation arises when the number of target attacks increases, which results in smaller group sizes for each attack success level, potentially diminishing the statistical significance of results. To address this, reducing the number of attack success levels (e.g., 5) is recommended to preserve statistical significance. Moreover, refining the attack models by integrating more diverse and complex threat scenarios could enhance the accuracy and relevance of privacy risk assessments. Finally, exploring the practical application of this framework in real-world OSN environments would provide valuable insights into its scalability and overall effectiveness.

References

[1] Anonymous, "Digital 2024: 5 billion social media users," , 2024. Accessed on September, 1, 2024

[2] Wang, Cheng and Yang, Bo and Cui, Jipeng and Wang, Chaodong, "Fusing behavioral projection models for identity theft detection in online social networks," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 4, pp. 637–648, 2019.

[3] Wang, Cheng and Zhu, Hangyu and Yang, Bo, "Composite behavioral modeling for identity theft detection in online social networks," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 2, pp. 428–439, 2021.

[4] Tucker, Catherine E, "Social networks, personalized advertising, and privacy controls," *Journal of marketing research*, vol. 51, no. 5, pp. 546–562, 2014.

[5] Zhang, Kunpeng and Bhattacharyya, Siddhartha and Ram, Sudha, "Large-scale network analysis for online social brand advertising," *Mis Quarterly*, vol. 40, no. 4, pp. 849–868, 2016.

[6] Ali, Ahmad and Malik, Ahmad Kamran and Ahmed, Mansoor and Raza, Basit and Ilyas, Muhammad, "Privacy concerns in online social networks: A users' perspective," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 7, pp. 601–613, 2019.

[7] Jain, Ankit Kumar and Sahoo, Somya Ranjan and Kaubiyal, Jyoti, "Online social networks security and privacy: comprehensive review and analysis," *Complex & Intelligent Systems*, vol. 7, no. 5, pp. 2157–2177, 2021.

[8] Confessore, Nicholas, "Cambridge Analytica and Facebook: The scandal and the fallout so far," *The New York Times*, vol. 4, pp. 2018, 2018.

[9] Paulina Okunytė, "Almost 500 million Instagram users had their data scraped, hackers claim," *Access: https://cybernews.com/news/instagram-user-data-scraping/*, 2024.

[10] Liu, Kun and Terzi, Evimaria, "A Framework for Computing the Privacy Scores of Users in Online Social Networks," *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, pp. 288–297, 2009.

[11] Liu, Kun and Terzi, Evimaria, "A framework for computing the privacy scores of users in online social networks," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 1, pp. 6, 2010.

[12] Srivastava, Agrima and Geethakumari, G, "Measuring privacy leaks in online social networks," *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2095–2100, 2013.

[13] Domingo-Ferrer, Josep, "Rational privacy disclosure in social networks," *International Conference on Modeling Decisions for Artificial Intelligence*, pp. 255–265, 2010.

[14] Nepali, Raj Kumar and Wang, Yong, "Sonet: A social network model for privacy monitoring and ranking," *2013 IEEE 33rd International Conference on Distributed Computing Systems Workshops*, pp. 162–166, 2013.

[15] Sramka, Michal, "Evaluating Privacy Risks in Social Networks from the User's Perspective," *Advanced Research in Data Privacy*, pp. 251–267, 2015.

[16] Petkos, Georgios and Papadopoulos, Symeon and Kompatsiaris, Yiannis, "PScore: a framework for enhancing privacy awareness in online social networks," *2015 10th International Conference on Availability, Reliability and Security*, pp. 592–600, 2015.

[17] Kilic, Yasir and Inan, Ali, "Privacy Scoring over OSNs: Shared

Data Granularity as a Latent Dimension," *ACM Transactions on the Web*, vol. 17, no. 4, pp. 1–28, 2023.

[18] Bioglio, Livio and Pensa, Ruggero G, "Impact of neighbors on the privacy of individuals in online social networks," *Procedia Computer Science*, vol. 108, pp. 28–37, 2017.

[19] Pensa, Ruggero G and Di Blasi, Gianpiero and Bioglio, Livio, "Network-aware privacy risk estimation in online social networks," *Social Network Analysis and Mining*, vol. 9, no. 1, pp. 15, 2019.

[20] Alemany, J and del Val, E and Alberola, J and García-Fornes, Ana, "Estimation of privacy risk through centrality metrics," *Future Generation Computer Systems*, vol. 82, pp. 63–76, 2018.

[21] De, Sourya Joyee and Imine, Abdessamad, "Privacy Risk Analysis of Online Social Networks," *Synthesis Lectures on Information Security, Privacy, and Trust*, vol. 10, no. 1, pp. 1–109, 2020.

[22] Braunstein, Alex and Granka, Laura and Staddon, Jessica, "Indirect content privacy surveys: measuring privacy without asking about it," *Proceedings of the Seventh Symposium on Usable Privacy and Security*, pp. 15, 2011.

[23] Pensa, Ruggero G and Di Blasi, Gianpiero, "A centrality-based measure of user privacy in online social networks," *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 1438–1439, 2016.

[24] Bioglio, Livio and Capecchi, Sara and Peiretti, Federico and Sayed, Dennis and Torasso, Antonella and Pensa, Ruggero G, "A social network simulation game to raise awareness of privacy among school children," *IEEE Transactions on Learning Technologies*, vol. 12, no. 4, pp. 456–469, 2018.

[25] Wang, Haizhou and Song, Mingzhou, "Ckmeans. 1d. dp: optimal k-means clustering in one dimension by dynamic programming," *The R journal*, vol. 3, no. 2, pp. 29, 2011.

[26] Gupta, S Das, "Point biserial correlation coefficient and its generalization," *Psychometrika*, vol. 25, no. 4, pp. 393–408, 1960.

[27] Bartlett, Maurice Stevenson, "Properties of sufficiency and statistical tests," *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, vol. 160, no. 901, pp. 268–282, 1937.

[28] Tamhane, Ajit C, "Multiple comparisons in model I one-way ANOVA with unequal variances," *Communications in Statistics-Theory and Methods*, vol. 6, no. 1, pp. 15–32, 1977.

[29] Kruskal, William H and Wallis, W Allen, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[30] Coban, Onder and Inan, Ali and Ozel, Selma Ayse, "Towards the design and implementation of an OSN crawler: A case of Turkish Facebook users," *International Journal of Information Security Science*, vol. 9, no. 2, pp. 76–93, 2020.

[31] Thomas, Kurt and Grier, Chris and Nicol, David M, "unfriendly: Multi-party privacy risks in social networks," *Privacy Enhancing Technologies: 10th International Symposium, PETS 2010, Berlin, Germany, July 21-23, 2010. Proceedings 10*, pp. 236–252, 2010.

[32] Chaabane, Abdelberi and Acs, Gergely and Kaafar, Mohamed Ali and others, "You are what you like! information leakage through users' interests," *Proceedings of the 19th annual network & distributed system security symposium (NDSS)*, 2012.

[33] Kosinski, Michal and Stillwell, David and Graepel, Thore, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the national academy of sciences*, vol. 110, no. 15, pp. 5802–5805, 2013.

[34] de Ayala, Ralph J, "An introduction to polytomous item response theory models.," *Measurement and evaluation in Counseling and Development*, 1993.