Improved Automated Framework to Improve Users' Awareness of Online Social Networks

Sameer Abufardeh *abufards@erau.edu*
Computer, Electrical, and Software Engineering Dept
Embry–Riddle Aeronautical University
Prescott, AZ, 86301, USA

Rahaf Barakat *rbarakat@ggc.edu*
School of Science and Technology
Georgia Gwinnett College,
Lawrenceville, GA 30043, USA

Abstract

The widespread usage of Online Social Networks (OSN) has introduced new privacy threats. These threats emerge when users intentionally but unknowingly share their information with a broader audience than intended(Ismail et al., 2021; Zhu et al., n.d.; Pew Research, 2019; Pew Forbes, 2022; Pew Ming, 2021; Bright et al., 2022; Cain et al., 2021). This paper introduces an improved framework for Users' Awareness of Online Social Networks. In our initial work, we identified critical privacy issues related to posting on social networks. We proposed a unique two-phase approach to address this problem (Barakat et al., 2016). The first phase involved recognizing key phrases in potential posts, particularly those that revealed location information. In this paper, we have expanded the detection rules to identify the location and other types of sensitive information, such as work and interests. We have developed a set of detection rules for this purpose and tested them with over 1500 actual Facebook posts. The initial detection system achieved a success rate of 85%, but the new system has improved the detection rate to approximately 88%.

Additionally, we conducted experiments with 500 actual Facebook posts from Arabic language users. However, the detection rates were lower due to the presence of English words mixed with the Arabic text.The second phase of our approach involves automatically grouping friends into sets called Circles of Trust. This allows messages containing sensitive information to be restricted to the appropriate Circles of Trust. We discuss an approach for initially assigning friends to circlesand the mechanisms for moving friends among circles as their relationship with the poster changes. This aspect has slight improvement but is still needs improvement.

Keywords: Online Social Networks OSN, privacy, risky, sensitive, dangerous, trust.

1. INTRODUCTION

Online social platforms, including Facebook, Myspace, Google, and Twitter, have become a fundamental component of our daily existence. People share a lot of personal information on these sites, such as profile information, status updates, and photos, which allows them to stay connected with each other. However, this information is not just seen by faithful friends but also by other parties with adverse interests, including identity thieves, stalkers, robbers, and sexual predators (Ismail et al., 2021; Zhu et al., n.d.; Pew Research, 2019; Pew Forbes, 2022).

While most users understand the importance of protecting their static profile information, many are unaware of the potential danger of sharing too much information through text posts with untrusted friends (Lenhart et al., 2007; Zhang et al., 2012). Information that people post is often intended for a specific group of people (Banks & Wu, 2009), but the sophistication of existing

privacy controls and their failure to capture the characteristics of the relationships between pairs makes it confusing for people when they try to make privacy decisions (Banks & Wu, 2009), (Facebook "Friend" Suspected in Burglary, n.d.). As a result, social network users are exposed to various types of threats (Ismail et al., 2021; Zhu et al., n.d.; Pew Research, 2019; Pew Forbes, 2022; Pew Ming, 2021; Bright et al., 2022; Cain et al., 2021; Gross & Acquisti, 2005; Lehrman, 2010; Lenhart et al., 2007). This type of information disclosure has also been a concern for government agencies, companies, and universities (Lenhart et al., 2007; Cloak, 2015).

In this paper, we introduce an improved and more comprehensive privacy control system that can be used to identify the location and other types of sensitive information, such as work and interest in users' text posts. The system is also designed to automatically separate friends into circles of trust to restrict information disclosure to the intended group of people. However, this aspect is still a work in progress. As an example of dangerous information, weuse location information in our discussion. We used Facebook to test our system. However, the system should work with other popular online social networks.

The remainder of the paper is organized in the following manner: Section 2 reviews research pertinent to the topic; Section 3 delineates the proposed system architecture and its evaluation; Section 4 showcases experiments and analyzes the results pertaining to the circles of trust; and lastly, Section 5 outlines prospective developments and draws conclusions.

2. RELATED WORK

Protecting users' private information on social networks is a critical issue that concerns both users and providers. Researchers have approached this issue in many ways, considering different aspects. Many researchers have studied and suggested techniques to protect users' privacy on social networking sites (SNSs) by focusing on static disclosure control.

For example, Reclaim Privacy,proposed by Zhang et al., (2012) is an open-source JavaScript privacy tool for Facebook that scans all user privacy settings and suggests a level for each setting to ensure the best privacy setting for the user's information. Another rule-based tool was proposed byBonneau et al., (2009)to help children manage their Facebook privacy settings and to serve as a monitoring mechanism for their parents. Furthermore, Bonneau proposed a "Privacy suite" where a user can adapt the privacy settings of another user, such as a trusted friend(Becker & Chen, 2009). PrivAware (Hirose et al., 2012)is another tool that measures privacy risks induced by friends. PrivAware suggests deleting risky friends or moving them to a private group to prevent profile information inference.

Although the proposed solutions help improve the privacy of SNS users, little research has been done on detecting information disclosure in user text posts or on countermeasures against information leaks (Watanabe & Yoshiura, 2010). Methods for detecting sensitive words in user text posts on SNSs were proposed by(Watanabe & Yoshiura, 2010; Nguyen-Son et al., 2012; Kataoka et al., 2010; Hart et al., 2009), where sensitive information is defined as "any information about a user that could be used to identify the user." The proposed solutions utilized search engines and user profile information to assess reachability between sensitive words in the user text post and the collected profile information. As a countermeasure, these tools employ different methods for text anonymization, ranging from omitting sensitive words from the post to defining generalization and synonymizing schemas to replace sensitive words.

Furthermore, Hart et al., (2007) and Pennebaker et al., (1999), developed a tag-based privacy control for online blogging websites was introduced to assist users with a non-technical background in describing their desired privacy policies. Instead of using per-object access control, users can define their privacy policies using tags. While this approach is a step forward, the dynamic nature of posts on online social networks makes it virtually impossible for users to know what topics they will discuss next and in what context. Therefore, predicting all tags for potentially revealing posts becomes a challenge.

Several trust models have been proposed in the literature to define trust communities in online social networks (OSNs),such as Liu et al., (2008), Caverlee et al., (2008),Hamdi et al., (2012),Moalla et al., (2010), and Kohavi & Provost, (1998). Existing approaches can be categorized as graph-based, interaction-based, and hybrid.

The proposed approaches in Zhang et al., (2006), Kim & Y.A., (2008),Golbeck et al., (2003), Golbeck, (2005), and Carminati et al., (2009)fall under the category of graph-based trust evaluation. These approaches leverage the network's structure and require user feedback to estimate trust. For instance, the trust model in Golbeck et al., (2003) and Golbeck, (2005)suggests a simple ontology for representing information about people and their connections on OSNs. They propose a trust rating scale of {1-9}, allowing users to annotate their relationships with information on how much they trust their friends. The trust rating value can be utilized to deduce the trust level between two nodes that are not directly connected.. Additionally, an algorithm called TidalTrust was proposed to infer trust relationships between nodes on the social network. However, results from Golbeck et al., (2003) and Golbeck, (2005)indicate that the effectiveness of these approaches depends on the connectivity of the trusted network and may perform poorly when the connectivity is sparse, which is often the case in online social networks. Furthermore, similar techniques have certain limitations. They only capture a few aspects of the trust notion, such as how individuals are related and how trust flows through the network. Consequently, they fail to capture other important computational trust properties, such as temporal and context dependency.

On the other hand, interaction-based trust models aim to compute trust values based on a user's interactions with others within the network. For example, in the context of online product reviews, a semi-supervised approach was proposed byLiu et al., (2008)to build a web of trust based on user interactions. The approach observed that a user trusts another user either because of their excellent reputation or because of positive interactions between them. To facilitate this, two taxonomies of trust factors were developed: one for user factors and another for interaction factors. The approach was evaluated using Epinions, a broad product review and rating community supporting various interactions. However, no evidence or supported study indicates that this approach and the taxonomies apply to OSNs similar to Facebook.

Another interaction-based trust model for OSNs is Strust, proposed by Nepaland Sherchan, (2011). This model aims to encourage positive interactions to increase social capital and improve social trust. The trust level in Strust consists of popularity trust and engagement trust, allowing the model to recommend two different types of people: leaders and mentors. Leaders are trustworthy members recognized through metrics such as the number of positive feedback/opinions on their posts or their followers. Mentors are members who actively engage in the community and are recognized based on metrics such as the number of members they follow or the number of posts they comment on. Zhan & Fang, (2011), proposed another approach that calculates trust scores between two directly connected members based on profile similarity, information reliability, and social opinions. The system then provides a trust score representing the actual trust between one member to another.

3. THE IMPROVED APPROACH

In this work, we have developed a system that can process user messages written in natural language without relying on search engines, accessing user profiles, or using a complex tagging system. The proposed system begins by identifying dangerous information that should be restricted. It then automatically divides potential message recipients into groups and suggests which group is considered safe to view the post.

A dangerous post is defined as revealing information about someone, either the poster or another person, which could lead to harassment or put that person or their property at risk (Ismail et al., 2021). Categories of dangerous posts on social networks include identity, location, work, and activities-related posts. Criminals often exploit information about a user's activities to find their time and location (Facebook "Friend" Suspected in Burglary, n.d.; Police: Thieves Robbed Homes, n.d.; Poulsen, n.d.). Our initial work in Barakat et al., (2016)focused on identifying dangerous location-

related posts, but in this paper, we have expanded our approach to cover other categories as well. We have extended the detection rules to identify locations anddifferent types of sensitive information, such as work and interests.

The proposed system comprises two major components: the Awareness System and the Circles of Trust System. The Awareness System aims to bridge the gap between the user's understanding and the countermeasures against revealing dangerous information on social networks. It detects dangerous information in the user's text posts using a set of detection rules and displays a warning message concerningsuch information disclosure.

On the other hand, the Circles of Trust System is designed to mitigate the risk of revealing dangerous information to all of the user's friends without any restriction. It categorizes the user's text posts based on topics and restricts the flow of information only to trusted friends. A friend may be part of multiple circles based on their topical interaction behavior with the user. The following section provides a detailed explanation of how the system operates, which is divided into two phases:

I. Phase One: The Awareness System Component.
II. Phase Two: The Circles of Trust Component.

3.1 Phase One: The Awareness System Component

The first phase of the system involves recognizing and detecting key phrases in the post that could compromise the safety of the poster or others. This phase consists of three steps:

Step 1: Natural Language Parsing: The plain text input (user's text post) is grammatically parsed using a natural language parser. This process generates a part-of-speech tagged text, where each word in the post is assigned a specific part-of-speech tag.

Step 2: Dangerous Information Extraction: The system then extracts potentially dangerous information from the tagged text. It identifies any information that matches the predefined set of detection rules for dangerous posts. If any dangerous information is found, it is flagged accordingly.

Step 3: Categorization: The extracted dangerous post is assigned an appropriate tag or topic category. These tags help the system categorize the topic of the dangerous post, enabling the restriction of sharing such information to only relevant and trusted parties.

The algorithm for extracting dangerous information is run as the user uploads their text post. Firstly, the algorithm utilizes the Stanford POS tagger library to associate each word in the post with its corresponding part-of-speech tag. Next, it searches the database for matching patterns. If the detected keyword and its associated tags match any detection pattern in the database, the post is flagged as containing dangerous information. Figure 1 illustrates the steps of the dangerous information extraction process, where the tagged text serves as input. The algorithm extracts any temporal data and compares it with existing detection patterns in the database to determine if the post contains dangerous informationor not.

The detection patterns for location information in our system are defined as follows:

1. Subset of time and location prepositions followed by a specific tag: This pattern captures phrases where a preposition like "to" is followed by a proper noun, indicating a destination. For example, "going to New York."
2. Subset of time and location prepositions accompanied by a pair ofspecific tags: This pattern captures phrases where a preposition like "at" is followed by a determiner and then a proper noun, indicating a specific location. For example, "at the park."

3. Keyword usage: This pattern captures the usage of specific symbols or keywords that reveal the user's location. For example, the symbol "@" is used in multiple posts instead of the word "at" to indicate a location. For instance, "@coffee shop."
4. Specific verbs followed by a specific tag: This pattern captures phrases where a verb like "heading" is followed by a noun, indicating a direction or destination. For example, "heading downtown."
5. Specific verbs followed by two specific tags: This pattern captures phrases where a verb like "going" is followed by a preposition like "to" and then a determiner, indicating an intended destination. For example, "going to the beach."

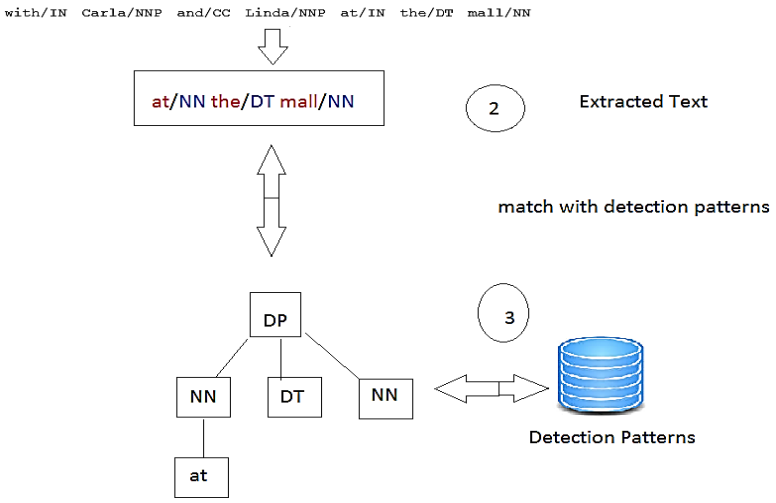


FIGURE 1: An Example of Dangerous Information Extracting.

6. Other commonly used location phrases on social networks: This pattern captures phrases commonly used on social networks that indicate a location, such as "on my way home."

These detection patterns allow the system to identify posts that contain location information and flag them as potentially dangerous, as disclosing specific location details can put the user's safety at risk.

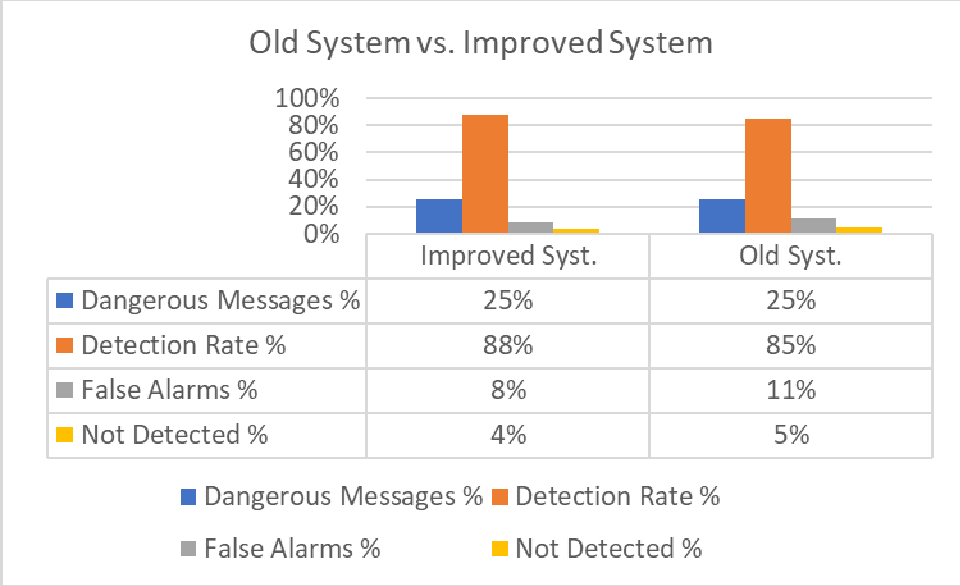
3.2 Experiments and Discussion of Results for the Awareness System

We developed a set of rules for detecting location, work, and interest information based on the abovementioned patterns. Our system uses these rules to identify key phrases that indicate potentially dangerous information in text posts. To evaluate the effectiveness of our detection rules, we collected an additional 1,500 Facebook posts from real users who agreed to participate in our research. We fed these posts into the system to measure the accuracy of our detection approach.

In the first part of our experiments, we aimed to demonstrate our system's true/false detection rate. The Awareness System tool uses our detection rules to scan text messages, which can consist of one or more words as well as symbols and emoticons, for potentially dangerous information. If the system detects such information, it prompts the user to confirm whether the post is indeed dangerous. If confirmed, the system increments the counter of detected dangerous posts; otherwise, it increments the counter of false alarms. If the system does not identify a post with harmful content, then the counter of undetected dangerous posts is incremented. By using this method, we were able to determine the number of revealing posts, detected posts, undetected posts, and false alarms and evaluate the accuracy of our detection approach. Table 1 summarizes our findings for both text status updates and check-in messages.

As observed, 21% of the amassed posts contained details regarding the user's present and/or upcoming whereabouts, events, or social agendas. This includes announcing travel itineraries, sharing summer holidays, scheduling visits to family or friends, and other social arrangements like participating in a soccer match, going to a cinema, planning hiking excursions, shopping trips, and so forth. Furthermore, a portion comprising 3% of the gathered data fell under the check-in category.

TABLE 1: Awareness System Old vs. Improved.



	Improved Syst.	Old Syst.
Dangerous Messages %	25%	25%
Detection Rate %	88%	85%
False Alarms %	8%	11%
Not Detected %	4%	5%

The system has achieved a promising detection rate of 88% for dangerous posts. However, it does make occasional mistakes. For instance, there have been instances where the system generated false alarms, flagging a post as dangerous when it was not actually harmful. An example of this is the post: "In Finland, we say: Happy friend's day." The system detected the word "in" and the proper noun "Finland," which satisfied one of the detection rules, making this post misclassified as dangerous. To address this issue, the system allows users to exclude specific phrases from being flagged as dangerous. Additionally, we have observed that long conversational-style sentences, quotes, or spelling mistakes often characterize text posts generating false alarms.

The system had a detection failure rate of 12%. However, since the false alarm was 8%, this makes the not detected only 4% for dangerous posts. An example of such a failure is the post: "Five more days and Vermont is calling." From this post, we can deduce that the author intends to visit Vermont in five days. Additionally, we noticed that the text posts triggering false alarms were typically lengthy, conversational-style sentence quotes or contained spelling errors. In contrast, the system could not identify 4% of the dangerous posts, with an example being a post similar to "Five more days and Vermont is calling." This type of post suggests that the author plans to travel to Vermont in five days. However, the current set of detection rules could not identify this as a dangerous post, highlighting the challenge of detecting similar text messages.

We implemented a parsing mechanism in the system to broaden the detection capabilities for various types of dangerous information, including work, home, and personal interests. This parsing process involves analyzing the text posts and comparing their content to a predefined set of categorized words using Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 1999). LIWC is a text analysis tool that utilizes an internal default dictionary containing word groups associated with specific personal concerns across seven categories: work, home, family, religion,

health, and more. By leveraging LIWC, we aim to enhance the system's ability to identify and address a broader range of potential risks and dangers in the posted content.

3.3 Phase Two: The Circles of Trust Component.

Trust has received the attention of many researchers with different expertise, such as computer science, software engineering, business and economics, cognitive sciences, psychology, and sociology. For that, we can find a plethora of definitions of trust in (Chen, Shi, 2009; Sherchan et al., 2013; Zhan & Fang, 2011); Zhang et al., 2006; Kim & Y.A., 2008; Liu et al., 2008; Caverlee et al., 2010). In the context of this work, we are interested in defining the level of trust at which a user can safely share their personal or sensitive information with another withoutprivacy concerns.The function of this component is to enhance the current trend of sharing dynamic information on social media platforms, steering it towards a trusted-friends paradigm.

Figure 2 depicts the distinctions between the two methods of broadcasting. In Figure 2 (a), the user shares status updates with all friends or connections on the social network. Conversely, in the approach proposed in Figure 2 (b), the user restricts the dissemination of status updatewhich could disclose the timing or location of their activities and social engagementsto only a select group of trusted friends.

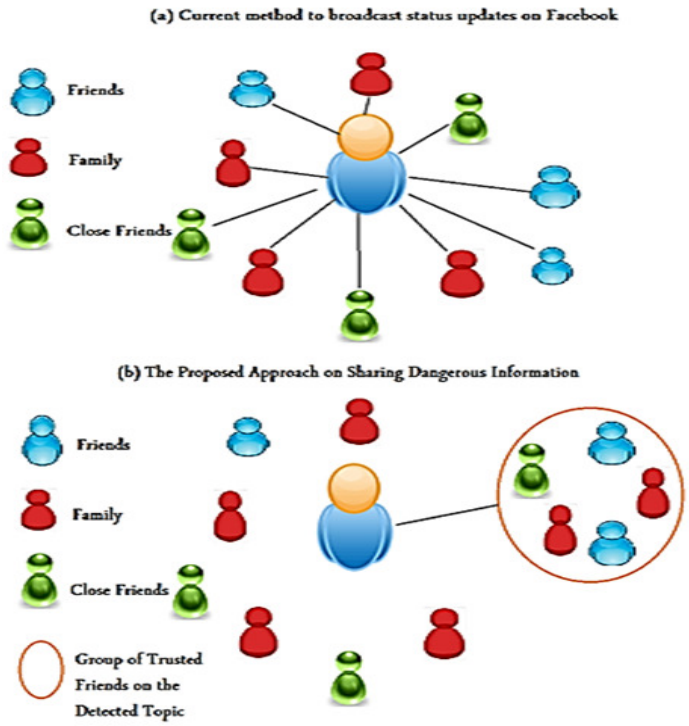


FIGURE 2: Illustrates the difference between the two broadcasting methods a & b.

Users on social networking platforms like Facebook have access to a range of static data sources to gauge the trustworthiness of their friends, such as profile details, mutual friends, and so on. Nevertheless, implementing trust-based privacy measures to safeguard dynamic data on social networks is considerably more complex, given the broad spectrum of topics that this data can encompass. Generally, some data is benign and not considered to be hazardous; hence can be accessible to the entire social network. On the other hand, certain data might expose the user to dangers, such as attacks or theft, due to revealing their location, political views, and similar details. In instances where users share updates containing potentially harmful information, the suggested system advises restricting the visibility of these posts to a group of trusted friends or those to whom the updates pertain directly.

Our strategy for limiting the dissemination of hazardous information exclusively to trusted individuals involves assembling a collection of interaction files. These files document user posts, their friends, and the interactions occurring between them, utilizing this data to pinpoint a roster of friends with whom the user feels at ease sharing and discussing particular subjects. Figure 3 delineates the comprehensive framework of our envisioned method for identifying trustworthy individuals.

The mechanism to recommend a suitable group of friends for sharing information is dynamic in nature. This signifies that whenever a user opts to publish a text message or status update, the system is obligated to:

(1) Scrutinize the impending content to assess its potential danger level, followed by
(2) Proposing a variable group of trusted friends, contingent upon the specifics of the status update content.

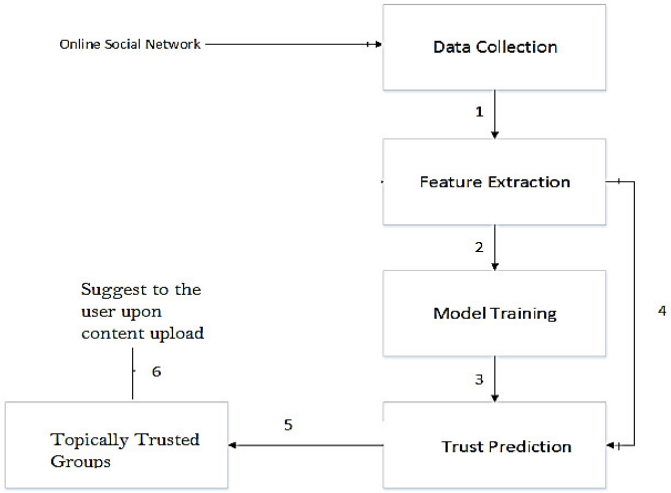


FIGURE 3: Identifying trusted parties' process.

The information found in status updates might pertain to a general subject and hence can be circulated to all active connections. Conversely, the content of the status update could touch upon certain specified sensitive topics prevalent on the social network, such as details about location, employment, family, or personal identity; in these cases, it should only be shared with active connections who are engaged in discussions on that specific topic.

We assume that friends who show continuous signs of interaction with the user are trusted parties.This is based on the observation that friends who lack any interaction markers are either:

(1) either not engaged or interested in the user's status updates or
(2) have a passive online social behavior and therefore are not active on the social network in general, or
(3) they have malicious motives and do not want to be noticed. With this in hand, we proceed to create the Circle of trusted parties.

3.3.1 Defining the Circle of Trust:

A circle of trust (*COT*) is defined as a collection of trusted friends or users $U = \{u_1, u_2, \ldots u_n\}$ with a specific sensitive topic t_n€ $T= \{t_1, t_2, \ldots, t_n\}$ by indicating continuous interaction with the individual about that particular topic t_n and, therefore, are permitted to view content identified as dangerous by the awareness system in and to other normal posts. We use $COT_{dtn}(U)$to denote the Circle of trust where a user $u_n$can be in more than one Circle of trust. Furthermore, it is possible for a user to not be allocated to any circle. In such scenarios, the user would be denied

access to view any posts classified as potentially dangerous. These circles function as an automated system to segment users' contacts on the social network into trusted groups regarding specific topics, thereby automatically regulating the visibility of posts deemed hazardous within relevant circles.People share major categories of dangerous information on social networks: identity, family, work, and location revealing information. Figure 4 provides a visual illustration of

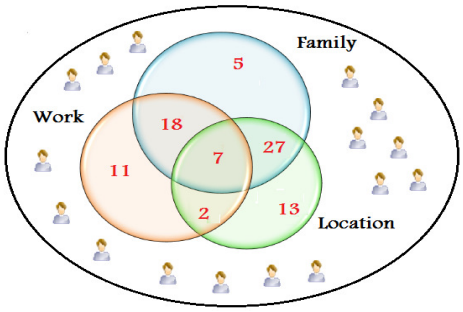


FIGURE 4: Illustration of Circles of Trust.

Circles of Trusts. Therefore, as illustrated in Figure 4, each inner circle will represent one of the categories, such as location circle, family circle, work circle, etc. Each of these circles will encompass individuals who engage with the poster based on specific topical interests.

Essentially, their engagement is centered around a distinct theme, like the timing and venues of the user's activities and social engagements. Circles can have intersections, as some users might be members of multiple circles. Individuals within a particular circle are granted permission to view text posts pertaining to that specific theme, alongside those posts that are not classified as dangerous. Depending on their interaction pattern, some individuals may not meet the criteria to be included in any circle. As a result, the privacy settings will be configured to automatically restrict any posts identified as dangerous from being visible to them, while maintaining their access to other non-hazardous posts.

3.3.2 Identifying the Trust Metrics

This step aims to identify and extract a set of trust metrics automatically from users' interaction records. Interaction methods on social networks can be implicit and/or explicit. Given the restriction on collecting data related to implicit interaction methods (privacy issues), we are limited to explicit interaction methods.For example, likes, tags, and replies, to create our interaction-based trust metrics. Furthermore, our interaction-based trust metrics included the number of positive and negative feedbackreceived from friend ***y*** to user ***x***. We also introduce different metrics to each comment structure:the flat comments discussion structure and the threaded comments discussion structure.

Online social networks serve as distinct adaptations of the broader network concept and can be characterized as a consortium of two or more nodes linked through various types of relationships (Chen, Shi, 2009; Zhang et al., 2006). In this context, nodes within online social networks might signify individuals, corporations, brands, and the like. The nature of relationships that users maintain with others can vary greatly, encompassing friendships, familial ties, professional connections, or acquaintanceships. Networks are often visually represented through graphs, where each node (or user) can monitor and document the interaction patterns of its neighboring nodes (or friends) with whom it shares a direct connection or friendship concerning its content.Therefore, trust metrics centered on interactions have been distinctly formulated as metrics assessing comments, hierarchical threaded discussions, appreciations or likes, and tagging activities (Chen, Shi, 2009; Sherchan et al., 2013; Zhan & Fang, 2011).

First, we define the trust-based social network; then, we proceed to explain our four defined metrics in detail in the following sections:

3.3.3 Trust-based social network:

The trust-based social network constitutes a fragment of the original social network, wherein members can rely on one another and securely disseminate personal or sensitive data without worrying about privacy violations.

The trust-based social network is denoted as OSN=(V, E, TL, T), is a trust-based network model consisting of the social graph, trust between users, and the topic of trust, Where:

- V: is the set of vertex, where a vertex in the graph represents a social network user.
- E: is the set of edges, where a direct edge from a user x to a user y indicates the existence of a social link between a user x and a user y. For example, x is a friend of y.
- TL: The trust level set, which encapsulates the degree of trust that one user places in another user, is delineated as:

$$TL = \{(x, y, TS\,(x, y, t)) \mid x \in V, \quad y \in V, \quad x \neq y\}$$

And $t \in$*{' Location','identity','work','family'}*

A trust feature vector from user x to user y is defined as:

$$v(x,y) = (W_{fdr}\, CR_{\,t}(x,y), W_{hdr}\, HDR_{\,t}(x,y),$$
$$W_{ar}\, AR_{\,t}(x,y), W_{tr}\, TR_{\,t}(x,y))$$

Where:

- $0 \leq W_{fdr}, W_{hdr}, W_{ar}, W_{tr} \leq 1$ and
- $0 \leq W_{fdr} + W_{hdr} + W_{ar} + W_{tr} \leq 1$ and

- $CR_{\,t}(x,y)$ Denotes the Comments Trust Metric: The act of friend y leaving several comments on user x's text post concerning topic t suggests that friend y is keen on fostering a connection with user x, demonstrated by investing time to comment on x's content. Consequently, this budding connection can potentially cultivate a trust-based relationship between the two, specifically regarding that particular topic.
- $HDR_{\,t}(x,y)$ Denotes the Hierarchical Threaded Discussion Trust Metric: we assume that trust is a key factor in triggering discussions between pairs on the social network, and more discussion increases the trust.
- $AR_{\,t}(x,y)$ Denotes the Appreciation or like.
- Trust Metric: The number of a friend y "likes" received to user x text posts indicates that y is interested in x's content.
- $TR_{\,t}(x,y)$ Denoted the Tagging Trust Metric: we assumethattagging is an essential indicator of an existing trust relationship between the pair.

TABLE 2: Elements of Trust Vector.

Metric	Description
$CR_{\,t}(x,y)$	The ratio of $noc_{\,t}(x,y)$ and $tnotp_{\,t}(x)$
$HDR_{\,t}(x,y)$	The ratio of $notc_{\,t}(x,y)$ and $noc_{\,t}(x,y)$
$AR_{\,t}(x,y)$	The ratio of $nol_{\,t}(x,y)$ and $tnotp_{\,t}(x)$
$TR_{\,t}(x,y)$	The ratio of $not_{\,t}(x,y)$ and $tnotp_{\,t}(x)$

Where:

- $noc_t(x,y)$: is the number of comments user *y*made on the text posted by*x* about the topic *t*.
- $tnotp_t(x)$: represents the cumulative number of text posts shared by user x on the social network about a particular topic, such as the topic of location.
- $notc_t(x,y)$: is the number of threaded comments by user *y* on the text post *p* of *x*.
- $tnoc_t(x,y)$: is the number of comments user *y* made on user x text content.
- $nol_t(x,y)$: is the number of likesuser *y*made on x's textabout the topic *t*.
- $not_t(x,y)$: is the number of tags user x made a friend *y* about the topic *t*.

Operating on the premise that not all interaction metrics hold equal significance in identifying trusted friends, we allocate specific weights to each interaction technique. For instance, certain interaction methods, such as likes and comments, enjoy widespread popularity among social network users. Meanwhile, the Facebook user community has recently introduced other interactive functionalities like replies (or threaded comments). Our strategy in determining these weights grants greater emphasis to the metric that witnesses frequent utilization by friends within a series of status updates while assigning lesser weight to metrics that aren't commonly engaged within a batch of status updates. Moreover, this method can be tailored to suit individual user interaction patterns with their friends.

It's important to note that trust manifests as a subjective preference; users might regard some methods of interaction as more crucial than others when it comes to establishing trustworthy friendships. To accommodate users in projecting their perception of trust onto social networks, we propose offering users the latitude to modify the weights assigned to different metrics, thereby allowing them to accurately embody their trust perspectives within Online Social Networks (OSNs).. For example, the weight of the comment metric could be defined as:

$$W_{fdr} = \sum_{k=1}^{n} \frac{CR_t(x,k)}{n}$$

Where:

- *n:* the number of friends (neighbor nodes)
- $CR_t(x,k)$*:* represents the metric rate assigned by friend k to the content posted by user x.

3.4 Topical Trust Scores Calculator:

The aggregation of metrics delineated in the preceding steps serves to compute individual trust scores. This amalgamation can facilitate the analysis of the correlation between these metrics and the accuracy rate in predicting trustworthy users. The trust score attributed to a user concerning a specific topic delineates the trust circle the user is associated with, consequently determining the user's entitlement to view subsequent posts within that category.

Every node (for instance, a user) within the social network has the ability to calculate the trust quotient with each of its adjacent nodes (like friends), pertaining to a certain topic, thereby establishing a bond between the pair as illustrated below:$TS(x,y,t) = T_{max}(W_{fdr} * CR_t(x,y) + W_{hdr} * HDR_t(x,y) + W_{ar} * AR_t(x,y) + W_{tr} * TR_t(x,y))$

Where:

- W_n : refers to the interaction metric weight, representing a portion of the highest attainable rating. Users can specify their preferred trust levels by allocating weights to each metric associated with interaction-based trust.
- T_{max}> 0: is a predefined constant that represents the maximum trust rating.

The calculated direct topical trust value falls within the continuous range of {0-100}. The trustworthiness of y in the eyes of x amplifies as the number of interactions - such as comments,

hierarchical threaded remarks, likes, and tags - relating to the content posted by user x on topic t escalates. Each node can determine the trust level with its neighboring nodes concerning topic t, and this trust level is asymmetric. Consequently, each node adheres to the aforementioned procedures to establish the trust level with its adjacent nodes.

Based on the trust computation,we can: (1) Separate users with malicious intentions, who do not engage in interactions, to prevent them from accessing posts that could potentially disclose harmful information; (2) Identify well-intentioned friends who regularly exhibit signs of interaction with the user; and 3) Enable the user to restrict content visibility to a trusted audience within a specific topic, safeguarding user privacy and safety while maintaining content relevance for the viewer.

3.5 Calculating the Threshold

In this phase, the system assesses the qualification of a friend to access another user's sensitive material (such as location information), predicated on their interactive behavior with the user concerning the relevant topic. A uniform threshold is implemented across all users to delineate the list of recommended trusted friends. This threshold is derived from the average trust scores computed in the preceding stage. Given the calculation of the above trust scores, the threshold employed to differentiate between trusted and non-trusted friends is defined as follows:

$$threshold\ (x) = \sum_{k=1}^{n} \frac{T(x,k,t)}{n}$$

Where:

- *n:* the number of friends (neighbor nodes)
- $T(x,k,t)$: individual trust scores

The criteria to determine whether friend y is permitted to view future posts from user x that disclose location information are defined as follows:

T (x,y,t)≥threshold(x)

Where:

threshold(x): refer to the owner's threshold or the required trust rating for user x location posts of out of a maximum trust rating of 100.

The circles of trust are fluid, necessitating recalculation each time a user logs into the social network. This module proposes a novel approach to depicting the dynamics of social networks, as the allocation of individuals into the Circles of Trust will vary in line with alterations in their topical interaction patterns over time.

4. EXPERIMENTS AND DISCUSSION OF RESULTS FOR CIRCLES OF TRUST

Although users on OSNs generate a massive amount of text status updates daily, the relationship between topic "Location" and the audiences' responsiveness has not been studied before. Therefore, our work on analyzing the collected interaction data set began with the goal of understanding this relationship. Facebook is considered the largest most studied social network. Therefore, we acquired interaction data from this platform and proceeded to do our analysis. The new data set has a total of 15,00 posts, which include 335 location revealing status updates and their corresponding likes, comments, tags, and replies. We also collected 500 Arabic posts.

4.1 High-Level Characteristics of Data

We begin with discussing the two high-level characteristics of the collected data set. First, the social links and active links, and a comparison between the two in terms of size. Then, the distribution of interaction methods.

First, we examine the difference in size between the friend list and the list of active friends who showed signs of interaction with the user in our data set. Using the collected data sets, we constructed each user's list of active friends. A friend is considered active if he/she interacted with the user at least once using comments, tags, or replies. We performed this analysis to prove our hypothesis that users on Facebook only interact with a small subset of their social links (friends), as opposed to the null hypothesis that assumes users interact with all their social links. Therefore, privacy management that limits information disclosure to a subset of active friends will not compromise the enjoyment of social networks. Our analysis shows that users interact only with a small percentage of their friends when they post text posts, 33% based on the data in our data set. Therefore, we can conclude that even though users in our data set have established an average of 309 social links (friend relationships), they only interacted with a subset of these friends. Our findings confirm the findings from previous research(Hirose et al., 2012) that have shown that the size of the activity network is significantly smaller than that of the social network.

Second, we examined the difference between friends' engagement in each interaction method. On Facebook users' wall posts, friends can interact through different interaction methods, such as comments or replies. We analyzed the data sets to understand friends' engagement in the interaction methods. The null hypothesis assumes that all interaction methods' effects are equal when determining the trustworthiness of friends, against the alternative hypothesis that the interaction methods' effects are not all equal when choosing trusted friends.

We concluded that like was the most adapted explicit interaction method between Facebook users in our data set. Like on OSNs, it becomes a form of confirmation from a friend that he/she has seen the user's status updates. On the other hand, a comment is still the most adapted form of explicit interaction between users on Facebook right after like. While the percentage of users' engagement in comments is less than likes, we believe that comments are more valuable than likes when determining trusted friends. The interesting fact about the comment is that it reflects the commenter's opinion. Moreover, comments could be studied and analyzed to suggest positive or negative user interactions.

The percentage of tags compared to likes and comments is significantly low because tagging on Facebook is more often associated with photo status updates. In our study, we only focused on text status updates. However, tagging a friend in a location revealing post, or check-in status update could be interpreted differently to reflect the existence of the tagged friend and the user at the same event or location or their plans to be in the same location in the future. As we mentioned earlier in this section, reply/threaded comment is a new feature that was added to Facebook shortly before the collection of our data set; therefore, replies data in our data set were inconclusive.

Our analysis shows that the location-revealing posts,for example,evoke a response from a small percentage of active links on the user's social network. Based on our collected data set, only 19% of active links interacted with the user on location-revealing posts. On the other hand, 81% of active links did not show signs of interactions with the user on location posts. These results could indicate that not all active friends are interested in day-to-day activities and social plans the user likes to discuss in his/her status update. Moreover, this finding could mean that restricting location information disclosure to the subset of friends who interact with the user on this topic might not compromise the enjoyment of social network.

4.2 Evaluating the Circles of Trust Prediction Rate

Our experiment aimed to test our proposed approach's ability to use a user's existing interaction data set to predict their future friends' interaction. We only chose to perform this evaluation using the comment metric due to insufficient data on other interaction methods in our data set. We evaluated our approach in terms of its error rate (the percentage of friends classified incorrectly).

We used a training data set for each participant in building the initial model. The system accepted the user's interaction data set and output a list of friends we predicted would interact with the user.

Then we used a validation data set to test the accuracy of our model. The training and the validation data sets have pre-known information about posts and commenters on the posts.

For each participant, we compared the suggested list of trusted friends using the training data set against the pre-known list of friends who commented on the user's location, revealing posts in the evaluation data set. To evaluate the overall performance of our system on the evaluation data set, we used the overall error rate. The overall error rate and the approach accuracy were defined as follows:

- Overall error rate = (sum of misclassified records) / (total records).
- Accuracy = 1- error rate

We used the confusion matrix(Kohavi & Provost, 1998) to assess the accuracy and performance of the classification model by providing information on the type and frequency of errors made by the model. It provides a detailed breakdown of the model's predictions compared to the actual ground truth. The matrix displays the count of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions made by the model.The average system accuracy was 66% compared to 61%. In contrast, the average system error rate was 34% compared to 39%. We also had 4% of friends who did not appear in our training data set and made their first appearance in the evaluation data set.

We believe the 4% newly appeared friends result from our approach of dividing the data set into training and evaluations. For example, this 4% could include friends added recently by the user on the friend list. Therefore, this could be their first attempt to interact with the user in the evaluation data set. Unfortunately, we did not have access to information about the dates of friendship formation, which limited the ability to draw conclusions based on the timing of friend additions. This number could also indicate the dynamics of the change in the interaction behavior of some friends. For example, some friends might have recently started being active on Facebook or recently started interacting with the user. As these friends continue interacting with the user on the same topic, their interaction rate will increase and be included in the Circle of trust. The opposite is also true; the absence of friend interaction on the given topic would cause their exclusion from the Circle of trust over time.

In addition, we conducted experiments with 500 Arabic posts; however, the results were inconsistent and yielded less than 50% accuracy. This can be attributed to Arabic being a highly inflected language with complex morphology. In Arabic, words can undergo various inflections and change their form based on the grammatical context. These characteristics present challenges for part-of-speech tagging systems, whichmust account for the extensive morphological variations. Removing or replacing English words with Arabic is not straightforward and can introduce ambiguity, making it difficult to accurately assign the correct part-of-speech tag to a word without considering the context.

4.3 Limitations

Our data set was limited to the subset of Facebook users willing to participate in our study. However, many studies have demonstrated similarities across popular social networks such as Google+ or Twitter.

The interaction data sets lack a "Date Joined" field and a "Friendship Date," meaning that the system did not consider the time difference between newly added and previously added friends. As a result, the comparison between these two groups may be unfair.

5. CONCLUSION AND FUTURE WORK

OSNs continue to bethe dominant platform to broadcast information among all generations. While these sites are an excellent place for people to stay connected and socially active, the privacy issues involving the broadcasting of dynamic datacontinue to be challenging. Existing OSNs lack the privacy control needed to manage the broadcasting of this type ofdata.

In this work, we enhanced the performance of our first proposed privacy control framework for dynamic data on online social networks. The proposed approach goal is to automatically limit information disclosure on OSNs by providing a list of trusted friends with whom the user can safely share these dangerous posts. The new results are encouraging. For example, testing the dangerous information detection approach yielded an 88% detection rate, which is very promising. The testing results using the preliminary Circles of Trust approach yielded an accuracy rate of 66% with a reasonably acceptable error rate of 34%. The results of our Circles of Trust model were inconclusive when the data set did not contain active users and active friends. Overall, we believe these results can be further improved by adding and refining the detection rules and the categorization approach to the types of sensitive information people share on OSNs, such as work, education, family, and health. We plan to continue improving the system's accuracyto improve the system performance on data sets fromArabic and other languages. Arabic text analysis continues to be challenging because the language used on social media such as Facebook, Twitter, and YouTube is informal(colloquial) and formal (standard).

6. REFERENCE

Ali, B., Villegas, W., & Maheswaran, M. (2007). A trust based approach for protecting user data in social networks. In K. A. Lyons and C. Couturier, (Eds.). *Proceedings of the 2007 conference of the center for advanced studies on collaborative research*, CASCON' 07, (pp. 288-293).

C. Zhang, J. Sun, and X. Zhu. (2012). "Privacy and Security for Online Social Networks: Challenges and Opportunities," *Network. IEEE*. 24 (4), pp. 13-18.

Carminati, B., Ferrari, E., & Perego, A. (2009). Enforcing access control in web-based social networks, *ACM Transactions on Information & System Security*, 13(1): 1-38.

Caverlee, J., Liu, L., & Webb, S. (2008). Social trust: Tamper-resilient trust establishment in online communities. *In Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'08).* (pp. 104-114).

Caverlee, J., Liu, L., & Webb, S. (2010). The social trust framework for trusted social information management: Architecture and algorithms. Information Sciences 180(1): 95-112.

Chen, X. & Shi, S. (2009). A literature review of privacy research on social network sites, *In MINES* '09 1, (pp. 93-97).

Cloak. (2015). Reclaim your privacy. http://www.reclaimprivacy.org

Facebook "Friend" Suspected in Burglary. (n.d.)"http://www.cbsnews.com/news/facebook-friend-suspected-in-burglary/

Golbeck, J., Parsia, B., & Hendler, J. (2003). Trust networks on the semantic web. *In Proceedings of the 7th International Workshop on Cooperative Intelligent Agents* (pp. 238-249).

Golbeck, J. (2005). Personalizing applications through the integration of inferred trust values in semantic web-based social networks. *In Processing of the ISWC Semantic Network Analysis Workshop*.

H. Kataoka, A. Utsumi, Y. Hirose, and H. Yoshiura.(2010). "Disclosure control of natural language information to enable secure and enjoyable communication over the internet," *in Security Protocols, ser. Lecture Notes in Computer Science*, 2010, pp. 178–188.

Hamdi, S., Gancarski, A. L., Bouzeghoub, A., & BenYahia, S. (2012). Iris: A novel method of direct trust computation for generating trusted social networks. *In Proceedings of the 11th IEEE International Conference on Trust, Security and Privacy in Computing and Communications* (pp. 616-623).

Hoang-Quoc Nguyen-Son, Quoc-Binh Nguyen, Minh-Triet Tran, Dinh-Thuc Nguyen, Hiroshi Yoshiura, Isao Echizen. (2012). "Automatic Anonymization of Natural Languages Texts Posted on Social Networking Services and Automatic Detection of Disclosure",*2012 Seventh International Conference on Availability, Reliability and Security*.

J. Bonneau, J.Anderson, L. Church.(2009). "Privacy Suites: shared privacy for social networks," in *ACM International Conf. Proc. of the 5th Symposium on Usable Privacy and Security, 2009, pp.1-2.*

Justin Becker, Hao Chen. (2009). "Measuring Privacy Risk in Online Social Networks" *In Proceedings of W2SP 2009: Web 2.0 Security and Privacy.*

K. Poulsen. Myspace predator caught by code. (n.d.).http://www.wired.com/news/technology/0,71948-0.html.

Kim, Y. A. (2008). Building a web of trust without explicit trust ratings. *In Proceedings of the 24th IEEE International Conference on Data Engineering Workshop* (pp. 531-536).

Kohavi, R., and Provost, F. (1998). On Applied Research in Machine Learning. In Editorial for the *Special Issue on Applications of Machine Learning and the Knowledge Discovery Process.*

L. Banks and S. F. Wu. (2009). "All Friends are Not Created Equal: An Interaction Intensity Based Approach to Privacy in Online Social Networks," Proc. WSPOSN '09.

Lenhart, A., & Madden, M. (2007). Teens, privacy, & online social networks. Pew Internet and American Life Project Report. http://www.pewinternet.org/2007/04/18/teens-privacy-and-online-social-networks.

Liu, H., Lim, E., Lauw, H., Le, M., Sun, A., Srivastava, J., & Kim, Y. (2008). Predicting trusts among users of online communities: An Epinions case study. *In Proceedings of the 9th ACM Conference on Electronic Commerce* (pp. 310-319).

M. Hart, C. Castille, R. Johnson, and A. Stent. (2009). "Usable Privacy Controls for Blogs," *In Proceedings of the International Conference on Computational Science and Engineering*, pp. 401-408.

M. Hart, R. Johnson, and A. Stent. (2007). "More Content - Less Control: Access Control in the Web 2.0," *Proceedings of the Web 2.0 Security & Privacy Workshop.*

Midori Hirose, Akira Utsumi, Isao Echizen, Hiroshi Yoshiura.(2012). "A Private Information Detector for Controlling Circulation of Private Information through Social Networks," ares, pp.473-478, *2012 Seventh International Conference on Availability, Reliability, and Security.*

Moalla, S., Hamdi, S., & Defude, B. (2010). A new trust management model in p2p systems. *In Proceedings of the 6th IEEE International Conference on Signal-Image Technologies and Internet-Based System, SITIS'*, (pp. 241-246).

N. Zakaria, K. Y. Lau, N.M.A. Alias, W. Husain.(n.d.). "Protecting the privacy of children in social networking sites with rule-based privacy tool" High Capacity optimal Networks and Enabling Technologies (HONET) pp. 253-257, 2011.

Nepal, S. & Sherchan, W. (2011). STrust: A trust model for social networks. In Proceedings of the 10th IEEE International Conference on Trust, *Security and Privacy in Computing and Communications* (pp. 841-846).

Pennebaker, J. W. and Francis, M. E. Linguistic Inquiry and Word Count. Lawrence Erlbaum. (1999).http://www.liwc.net/.

Pew Bright, L. F., Logan, K., & Lim, H. S. (2022). Social Media Fatigue and Privacy: An Exploration of Antecedents to Consumers' Concerns regarding the Security of Their Personal Information on Social Media Platforms. *Journal of Interactive Advertising*, 1-16.

Pew Forbes (2022). The Top Security Threats of 2022. https://www.forbes.com/sites/splunk/2022/03/01/the-top-security-threats-of2022/?sh=4315d2a12e5d.

Pew Ming, S. S. Y. (2021). Research on Influencing Factors of Information Privacy Concerns of Social Media Users. *Information and Documentation Services*, 42(3), 94-104.

Pew Research. (2019). Social media fact sheet. https://www.pewresearch.org/internet/fact-sheet/social-media/.

Police: Thieves Robbed Homes Based On Facebook, Social Media Sites. (n.d.).http://www.wmur.com/Police-Thieves-Robbed-Homes-Based-On-Facebook-Social-Media-Sites/11861116.

Ralph Cain, J. A., Imre, I. (2021). Everybody wants some: Collection and control of personal information, privacy concerns, and social media use. *New Media & Society*, https://doi.org/10.1177/14614448211000327.

Ralph Gross, Alessandro Acquisti. (2005). "Information Revelation and Privacy in Online Social Networks (The Facebook case)" Pre-proceedings version. *ACM Workshop on Privacy in the Electronic Society (WPES).*

R Barakat, S Abufardeh, K Magel. (2016). "Automated framework to improve users' awareness on Online Social Network Automated framework to improve users' awareness on Online Social Networks". *2016 IEEE International Conference on Electro Information Technology (EIT).*

R Ismail, A., M. R. Hamzah, H. Hussin. (2021). "The roles of trust and perceived risks on online self-disclosure." *AIP Conference Proceedings*. Vol. 2347. No. 1. AIP Publishing LLC, 2021.

Sarah Palin email hack. (n.d.).http://en.wikipedia.org/iki/Sarah_Palin_email_hack.

Sterling, G. (2013). Pew: 94% of teenagers use Facebook, have 425 Facebook friends, but Twitter and Instagram adoption way up, Pew Research Center. http://marketingland.com/pew-the-average-teenager-has-425-4-facebook-friends-44847.

Wanita Sherchan , Surya Nepal , Cecile Paris.(2013). A survey of trust in social networks, *ACM Computing Surveys* (CSUR), v.45 n.4, p.1-33.

Watanabe, N. and Yoshiura, H. (2010). "Detecting Revelation of Private Information on Online Social Networks," *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Darmstadt, Germany, 2010, pp. 502-505, doi: 10.1109/IIHMSP.2010.128.

Yosef Lehrman. (2010). "The Weakest Link: The Risks Associated with Social Networking Websites" New York Police Department, *Journal of Strategic Security*, Vol. 3 No. 2 .
Zhan, J. & Fang,X. (2011). A novel trust computing system for social networks. *In Proceedings of the IEEE International Conference on Privacy, Security, Risk and Trust* (pp. 1284-1289).

Zhang, Y., Chen, H., & Wu, Z. (2006). A social network-based trust model for the semantic web. In *Proceedings of the 6th International Conference on Autonomic and Trusted Computing* (pp. 183-192).

Zhu, Tianqing, et al.(n.d.) "The dynamic privacy-preserving mechanisms for online dynamic social networks." *IEEE Transactions on Knowledge and Data Engineering* 34.6 (2020): 2962-2974.