Iron Mask: Trust-Preserving Anonymity on the Face of Stigmatization in Social Networking Sites

Hamman Samuel[1][0000-0002-3053-6047] and Osmar Zaïane[1][0000-0002-0060-5988]

[1] University of Alberta, Edmonton, Canada
hwsamuel@ualberta.ca, zaiane@ualberta.ca

Abstract. Social networking sites are pervasively being used for seeking advice, asking questions, giving answers, and sharing experiences on various topics including health. When users share content about sensitive health topics, such as sexual dysfunction, infertility, or STDs, they may wish to do so anonymously to avoid stigmatization and the associated negative effects on mental health. However, a user masking their name with a pseudonym may still be inadvertently exposing their identity because of various quasi-identifiers present in their profile. One such quasi-identifier that has not been investigated in literature is the content itself, which could be used for authorship identification. Moreover, an anonymous user's credibility cannot be established because their profile is no longer linked with their reputation. This study proposes the Iron Mask algorithm for providing enhanced anonymity while preserving trust. Iron Mask improves anonymity by using a probabilistic machine learning approach based on writeprint identification and inclusion of content as a quasi-identifier. Iron Mask also introduces the concept of a trust-preserving pseudonym which masks user identity without loss of credibility. We evaluate the proposed algorithm using datasets from Quora, a question-answering social networking site, and demonstrate the efficacy of our algorithm with satisfactory recall and survey feedback results.

Keywords: Anonymity, Pseudonymity, Trust.

1 Introduction

Social Networking Sites (SNS) provide various mechanisms to facilitate sharing of information, advise, questions and answers related to various topics. Different types of actions are available on various instances of SNS such as Quora, Stack Exchange, Facebook, and Twitter. Users can "friend" or "follow" other users, thereby creating connections. Different types of content can be created and shared on SNS, including text-based articles, blogs, microblogs, or multimedia content such as pictures and videos, or links to other users' postings and external websites. Users can also subscribe to topics of interest, thereby creating online communities of like-minded individuals. Normally, the user who is sharing a posting is identified as the author of the post by displaying their registered user name or full name. However, situations can arise in which the user does not want to be identified.

If a user shares a link with connections about sexual dysfunction or infertility, the user may wish to do so anonymously to avoid any potential stigmatization which may result from the assumption that the user sharing the content suffers from the condition [1]. Pseudonyms have proven effective within online forum communities for supporting stigmatized issues and people tend to discuss and learn more openly about stigmatized issues when the perceived risk of being publicly associated with the issue is taken away [2]. On the other hand, the negative effects of users experiencing social stigma can be severe, with outcomes ranging from poorer mental health to increased risk behaviors [3]. It is known that users have been increasingly using the internet for sharing personal experiences and seeking advice about various personal issues, which increases the likelihood of stigmatization from online activities [4].

Despite the potential severity of online social stigma, options and controls to anonymously post content are not well-supported in most SNS. Users on SNS may hide their real identity by creating a new account with a fake name or pseudonym, thereby duplicating the SNS user base. This is not ideal and unnecessarily complicates the process of information sharing. From the list of popular social media websites such as Facebook, Twitter, LinkedIn, YouTube, Google+, Stack Exchange and Quora, only the latter allows asking questions anonymously without needing to create a new account.

There are also potential drawbacks with the approach to replace the user's real identity with the generic pseudonym "anonymous". Firstly, despite their name being replaced by "anonymous", users may be inadvertently releveling their identity because of the similarities between the content they have posted in the past. Phrases, wordings, topics and other nuances about the writing style in the user's past postings may constitute a quasi-identifier that can be associated to a specific user. Secondly, the generic anonymous pseudonym also eliminates the user's associated credibility, thereby motivating the need for trust-preservation during anonymization. Information from a known source is easier to identify as being either more or less trustworthy than if it is coming from an unknown source [5]. For example, someone unknown suggesting in a post to take a certain medication will be less credible than a person who is known to be a medical expert. At the same time, advise from a person confirmed to have little knowledge of medicine would give a clearer indication of distrust, in contrast with when an unknown person gives similar advice.

The notions of credibility and trust are homonyms related to the belief that a person's actions during an interaction will be beneficial rather than detrimental [6]. Credibility of a user in SNS is often expressed using a reputation system based on an aggregate of positive and negative feedback received from other users. This mechanism is used by Quora and Stack Exchange, where the aggregate points received can be used to determine a user's level of expertise. The assumption, barring Sybil attacks, is that the higher a user's aggregate points, the more knowledgeable they are, given that they have received more positive than negative feedback. Another form of trustworthiness is based on the personalized grouping of connections based on closeness of relationship. This strategy is available in Facebook, where connections are categorized as "family", "close friends", "friends", "acquaintances", "friends of friends". This hierarchy of closeness can be interpreted as being directly proportional to trustworthiness, the closer the user, the more trustworthy.

Our proposed algorithm, Iron Mask, uses the whiteprint or authorship identification approach to take into account the user's historical content, thereby enhancing anonymity by minimizing the risk of re-identification and decreasing the likelihood of online stigma. Iron Mask also provides trust-preservation to balance the social network's needs to generate credible content with the user's need for optional yet reliable anonymity. The naïve approach of explicitly revealing information related to user credibility would constitute a quasi-identifier, and could lead to identity being compromised through correlations [7], so a more sophisticated approach is required. To achieve this, Iron Mask introduces the concept of the Trust-Preserving Pseudonym (TPP), which provides a broader range of pseudonym labels, in addition to the generic "anonymous" pseudonym to mask or cover up the user's actual account name identity while appropriately summarizing credibility information.

The scope of our work is on self-contained SNS, and adversaries external to the social network are not considered. External adversaries would have additional information that is outside the network, while internal adversaries would be registered users within the SNS. Two aspects of the Iron Mask algorithm need to be evaluated. Firstly, the whiteprint identification approach is tested using datasets from the Quora question answering community. The evaluation demonstrates the accuracy of predicting the author of a post even when their user name is hidden. Secondly, the trust-preservation approach and TPP are evaluated using a survey-based approach to demonstrate usefulness and applicability.

The rest of the paper is organized as follows. Section 2 presents related work on anonymity in SNS, while Section 3 gives an outline of the Iron Mask algorithm. Section 4 provides details of our experimental design and evaluation results, including identification of content dealing with sensitive topics, while Section 5 concludes with comparative analysis and commentary on future directions.

2 Literature Review

Narayanan et al. [8] investigated different de-anonymization attacks on social networks such as Twitter. Their study looked at possible re-identification risks involved with user information available on more than one social network, i.e. Twitter and Flickr, and how intersection of common information could lead to re-identification. A similar study by Beach et al. [9] also looked at anonymity in social networks and the disadvantages of using traditional anonymization methods such as *k*-anonymity on SNS like Facebook. However, these studies focused on partial anonymization where some properties of the user are hidden, such as name, while others are visible, like gender or location. Our research looks at the scenario where the user's identity is completely hidden via full anonymization. Also, the adversaries considered in these studies had information that was external to the target SNS, while our study focuses on information being exclusively within the network.

The veiled viral marketing approach was suggested by Hansen and Johnson for sending anonymized messages to friends within Facebook [1]. In the study, research on an awareness campaign for Human Papilloma Virus (HPV) showed that people who knew HPV is a sexually transmitted were more likely to feel shame and stigma, and less likely to share or post information about it on their Facebook profiles. Moreover, people were willing to share links to websites about social causes like breast cancer awareness, but were unlikely to do likewise for links to syphilis or gonorrhea websites. The proposed veiled viral marketing approach allowed sending of anonymous "veiled" messages to friends, which essentially substituted the user's identity with the "friend" pseudonym. Users would know that the message came from one of their friends, but would not know which friend actually sent the message. However, this study did not take into account any risk of de-anonymization from the content being a quasi-identifier. In addition, no exploration was made on any relationships between credibility of information and anonymity, although it was implied that users trusted their friends' shared content more than that of strangers.

The relationship between historical posted content and user identity was partially investigated by Milhail and Ilya as a side effect of their study [10]. They looked into the situation where the same person had several different accounts on the same web portal, potentially for manipulation of feedback, ratings and Sybil attacks on the web portal. The study proposed a solution to short messages text authorship determination using a naive Bayesian classifier. The classifier was trained using short messages from known users. This classifier was then used to determine if a new post belonged to an existing user. One drawback of the study was the low accuracy of 50%, which could be attributed to the selection of features, size of the training data, or the classifier used. Another similar study was conducted by Keretna et al. on whiteprint identification in Twitter to recognize multiple accounts being created by the same user [11].

The relationship between anonymity and trust has also been explored in peer-to-peer networks for providing ratings and feedback anonymously [12], which also has applications in e-governance and online voting [13]. The proposed approaches focus on the anonymized reporting of aggregated results. This relationship is also important to dematerialized money and cryptocurrencies, where the emphasis is on completing trustworthy transactions while maintaining anonymity of the agents involved [14]. In contrast with these domains, to the best of our knowledge, there has not been much direct work done on enhancing the relationship between trust and anonymity in SNS. There are various SNS that have either internal or external anonymity controls. The former, such as Quora, allow users to anonymously post content without revealing their actual registered account's user name. The latter includes SNS that let anyone with an internet connection post content without having to register an account. Pseudo-accounts are a third option in which users register fake accounts to hide their real identities, and subsequently do not require additional anonymity controls or options [15].

3 Methodology

The general workflow of anonymization is summarized in Figure 1, where the user is provided a choice of using anonymity. If the user selects in the affirmative, then the author of the posting is reported with the generic label of "anonymous", and no hyper-links or internal associations to the actual user are maintained. Consequently, anyone viewing the content will see the content's author as anonymous. Otherwise, the actual identity of the user is displayed. These two binary choices are available on SNS such as Quora. Our proposed approach using the Iron Mask algorithm provides an alternative route for anonymity.

Iron Mask assigns a pseudonym using a two-stage approach: firstly, the content to be posted is scrutinized to determine the probability of de-anonymization. Secondly, a trust-preserving pseudonym is assigned based on the SNS characteristics and the user's profile. For example, the TPP assigned could be "close friend", which would let the reader know that a close friend of theirs has made this posting, which could be better received than a posting from a stranger or casual acquaintance. The TPP could also be "competent" based on a combination of the Dreyfus model of skill acquisition and the user's reputation points on the SNS. This would let the reader know that the user' is knowledgeable or not based on other users' feedback on previous postings.

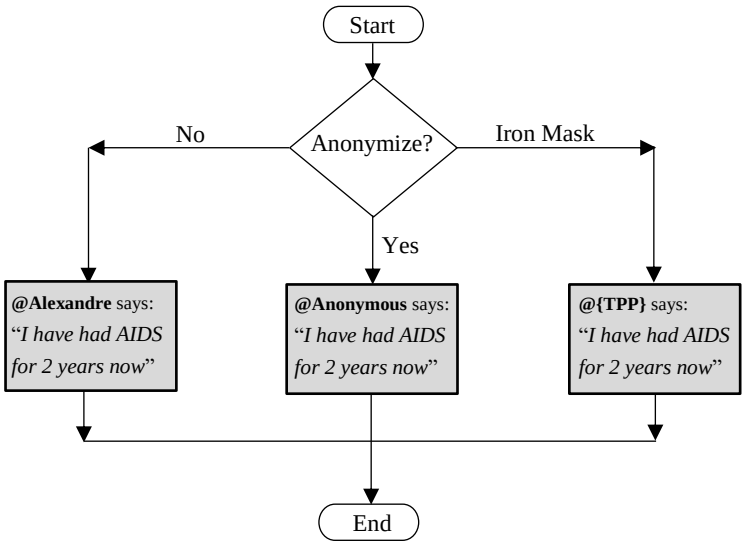

Fig. 1. Overview of anonymization approaches and trust-preserving pseudonyms.

Programmatically, the procedure for posting new content with options for anonymization with Iron Mask is abstracted in Algorithm 1. If there is a risk of re-identification, then the user is warned of this before proceeding. It is up to the user to take the risk or not. The user's pseudonym is determined by the `TPP()` function. The generic `SAVE()` procedure is dependent on the SNS to save the content to the appropriate persistent storage such as a database.

Algorithm 1. `POST` (*user*, *content*, *anon*)

Require: *user*: the posting's author, *content*: the content to be posted, *anon*: option to anonymize or not

```
1. if anon is False then
2.    SAVE (user, content)
3.    return
4. if IRONMASK(user, content) is False then
5.    WARN()
6. else
7.    pseudonym = TPP (user)
8.    SAVE (pseudonym, content)
9. Return
```

3.1 Whiteprint Identification using Probabilistic Classification

Algorithm 2 outlines the Iron Mask step-by-step procedure using a probabilistic classifier. Probabilistic classification is able to predict a probability distribution over a set of classes. In essence, probabilistic classifiers provide the degree of confidence of a sample belonging to a class [16]. To initialize, a probabilistic classifier is trained using existing users and their postings from Quora, where the user name is the class, and the content is converted to *n*-grams as features. Training computes a score for how strongly classes and attributes are associated, and the trained model can then be used for making predictions on new data, while probability calibration converts the scores to probabilities [17]. All possible combinations of adjacent words of length *n* within a posting are referred to as *n*-grams. For instance, a posting containing words [w_1, w_2, …, w_n] would yield bigrams as [w_1w_2, w_1w_3, …$w_{n-1}w_n$]. For our implementation, we use naïve Bayes with isotonic regression as the probability calibration in the Scikit-Learn library [18]. We used a combination of uni-, bi- and tri-grams as features.

Essentially, the probabilistic classifier is performing whiteprint identification by associating content and user identity [11]. The content is also being used as a quasi-identifier. More formally, user names and historically posted content can be expressed as the traditional database table defined in *k*-anonymization with *n* rows and *m* columns, with the rows representing each user's previously posted content, and the columns representing *n*-grams from the content, along with the user name. This database table also maps to the classification problem model, where each row comprises a complete tuple, and, in our case, the user name column is the identified class [19].

A new posting is input to the trained probabilistic classifier to get a set of predicted candidate users. On the trained probabilistic classifier, two thresholds are available for making a decision: top-*n* and τ. The top-*n* threshold returns the top candidate users based on the sorted degree of confidence. If the actual author is found within these top-*n* candidates, then the Iron Mask algorithm returns a warning status. On the other hand, if the confidence level for predicting the actual author is greater than a given threshold, τ, then Iron Mask also returns a warning status. The thresholds can be used concurrently or separately based on how they are configured. For instance, configuring $\tau = 1$ or $n = 0$ would disable either threshold.

Algorithm 2. `IRONMASK` (*user*, *content*)

Require: *user*: the posting's author, *content*: the content to be posted, τ: internal threshold for determining risk of re-identification, *n*: internal threshold for choosing number of predicted candidates

1. *candidates* = `PROBCLASSIFIER` (*content*)
2. *top_candidates* = *candidates*[:*n*]
3. *user_prob* = *candidates*.`FIND` (*user*).*probability*
4. **if** *user* **in** *top_candidates* **or** *user_prob* ≥ τ **then**
5. **return False**
6. **else**
7. **return True**

3.2 TPP Algorithm

In addition to the generic "anonymous" pseudonym, additional pseudonyms can be assigned to a user to preserve information about their credibility, based on level of expertise or level of relational closeness.

Level of Expertise. We use the Dreyfus model of skill acquisition as a reference for anonymizing a user's online reputation on the SNS [20]. The Dreyfus model specifies five categories of expertise: novice, advanced beginner, competent, proficient, and expert. Depending on the SNS, there are various reputation attributes available. Quora allows users to "Upvote" or "Downvote" postings based on the voter's perceptions of quality. This feedback, along with general interaction statistics such as number of postings and comments, can be aggregated as a reputation score for each user to determine the user's level of expertise on the Dreyfus hierarchical scale, with pre-configured mappings of reputation scores to each level. Algorithm 3 outlines this approach as an implementation of TPP using expertise and reputation. The scoring function incorporates the number of upvotes and downvotes received, as well as the total number of postings, while penalizing downvotes. The severity and effect of downvotes on reputation can be adjusted using a weighting factor. The reputation score aggregation formulation can be customized to fit the needs of the SNS. Moreover, if there is not enough data available to define level of expertise, the "anonymous" label can be used.

Algorithm 3. `TPP` (*user*)

Require: *user*: the posting's author, t_i: values for Dreyfus levels, where $t_{i-1} < t_i$, w: weighting factor to adjust severity effect of downvotes

1. *rep* = `GETREPUTATION`(*user*)
2. *score* = (*rep.upvotes*+*rep.num_postings*) / (*w***rep.downvotes*+1)
3. **if** *score* $\geq t_1$ **then**
4. **return** `EXPERT`
5. **else if** *score* $< t_1$ **and** *score* $\geq t_2$ **then**
6. **return** `PROFICIENT`
7. **else if** *score* $< t_2$ **and** *score* $\geq t_3$ **then**
8. **return** `COMPETENT`
9. **else if** *score* $< t_3$ **and** *score* $\geq t_4$ **then**
10. **return** `ADVANCED_BEGINNER`
11. **else if** *score* $< t_4$ **and** *score* $\geq t_5$ **then**
12. **return** `NOVICE`
13. **else if** *score* $< t_5$ **then**
14. **return** `ANONYMOUS`

Level of Relational Closeness. For SNS that do not use reputation metrics, the type of ties or connections and their relative perception of relational closeness can be used as an indicator of trustworthiness. For instance, Facebook is not designed as question answering community, and there is no explicit notion of reputation. Algorithm 4 outlines the approach for determining a TPP based on a hierarchy of relational closeness. At each level of the hierarchy, starting from the more intimate, for instance family, the connections of the user posting the new content are enumerated. Each connection's number of connections are then determined, and a probability of re-identification solely based on the number of connections is computed. As an example, a user named D'Artagnan has three "close friends" connections named, Aramis, Athos and Porthos. For each of the three connections, their number of "close friend" connections are computed in turn. Aramis may have only one close connection, D'Artagnan himself. Hence, Aramis will surely guess the identity of D'Artagnan if he were to be labelled with the pseudonym "close friend". In this situation, a pseudonym that is higher in the hierarchy is then attempted recursively. If no suitable pseudonym is found, then the "anonymous" label is used. Hence, the possibility of de-anonymization when using a TPP is also covered by incorporating the probability of re-identification at each stage of the TPP hierarchy. For social networks without hierarchical relationships, such as Twitter, a TPP would not capture trustworthiness due to lack of differentiation between connections.

The proposed approaches for TPP are meant to cover characteristics of different variations of SNS. For Quora, the level of expertise is more appropriate. Another aspect to consider when selecting either TPP approach is the nature of the posting. For instance, a posting about a user sharing their experiences on a sensitive topic does not necessarily require knowledge about their reputation, but their relational connection would help.

On the other hand, a user giving an anonymous answer or advise about a sensitive health topic would need some validation of credibility in order to ensure no harm is done.

Algorithm 4. `TPP` (*user*)

Require: *user*: the posting's author, *connection_hierarchy*: hierarchy of connection types based on level of relational closeness, τ: minimal connections threshold

```
1. pseudonym = ANONYMOUS
2. for each type in connection_hierarchy
3.     connections = GETCONNECTIONS (user, type)
4.     pseudonym = type
5.     for each connection in connections
6.         num_connections = GETCONNECTIONS(connection, type).COUNT()
7.         if num_connections ≤ τ then
8.             pseudonym = ANONYMOUS
9.             break
10. return pseudonym
```

4 Evaluation

For evaluation of the proposed methodology, we retrieved datasets from Quora via unofficial APIs that have been approved by Quora, and in line with Quora's terms of use on web scraping and rate limits. A summary of the number of subsets retrieved is given in Table 1, along with the topics used for filtering the postings. The topics were selected in line with the focus of our research on sensitive health content.

The retrieval process involved accessing a topic's list of questions, then retrieving the list of followers of the topic. Each follower's profile was then programmatically accessed, and questions they have posted were retrieved, as well as upvotes and downvotes on each question. Users are also allowed to post questions anonymously, in which case the questions do not appear on their profile's listing of questions asked. Next, for each question retrieved, the corresponding answers were also enumerated, including the associated upvotes and downvotes, as well as additional profiles of users who authored the answers.

Table 1. Quora dataset for evaluation, filtered by topics
A: Men's sexual health, B: Women's sexual health, C: Sexuality, D: HIV, E: Mental Health

Subset	A	B	C	D	E	**Total**
Initial profiles retrieved	58	48	110	56	122	394
Questions retrieved	179	122	300	151	300	1,072
Answers retrieved	895	488	1,500	302	960	4,145
Additional profiles	358	97	750	30	348	1,619

4.1 Accuracy of Content as Quasi-Identifier

In order to determine the accuracy of whiteprint identification, a sample of users were arbitrarily selected from the Quora dataset. Various iterations of this process were performed using different configurations of threshold, while the number of users selected was kept constant for all iterations. The sample dataset was then split into two parts for training and testing. The training set *Tr* was used for building the probabilistic classifier model. Next, the trained model was used with the other half of the sample dataset, i.e. the test set *Ts*, to predict user identity. Both the test and training sets were split such that the users in the test dataset were also in the training dataset. However, the content within the test dataset was not in the training counterpart. More formally, if u_i represents users and c represents content of the users, then $u_i \in Tr$, $c_i \in Tr$, $u_j \in Ts$, and $c_j \in Ts$, but $u_j \subset u_i$ and $c_j \not\subset c_i$.

The recall measure was used to determine the effectiveness of the trained model. Probabilistic classifiers are traditionally evaluated using root mean square error, but since we are evaluating the Iron Mask algorithm and various threshold configurations for top-n and τ, the recall metric is best to achieve our evaluation goals as well as capture the classifier's accuracy. As an illustrative example of our evaluation strategy, if top-n = 1, that implies that Iron Mask would only detect the user's correct identity and give a warning if the trained model ranked that identity with the highest probability. In other words, if a given user's identity was correctly predicted within the top-n, the recall score was recorded as 1, else it was recorded as 0. An average of the recall was taken for the various users selected for each iteration, shown in Figure 2 for different values of top-n. Similarly, for different values of τ, recall was recorded based on whether Iron Mask gave a warning or not, and the results are summarized in Figure 3.

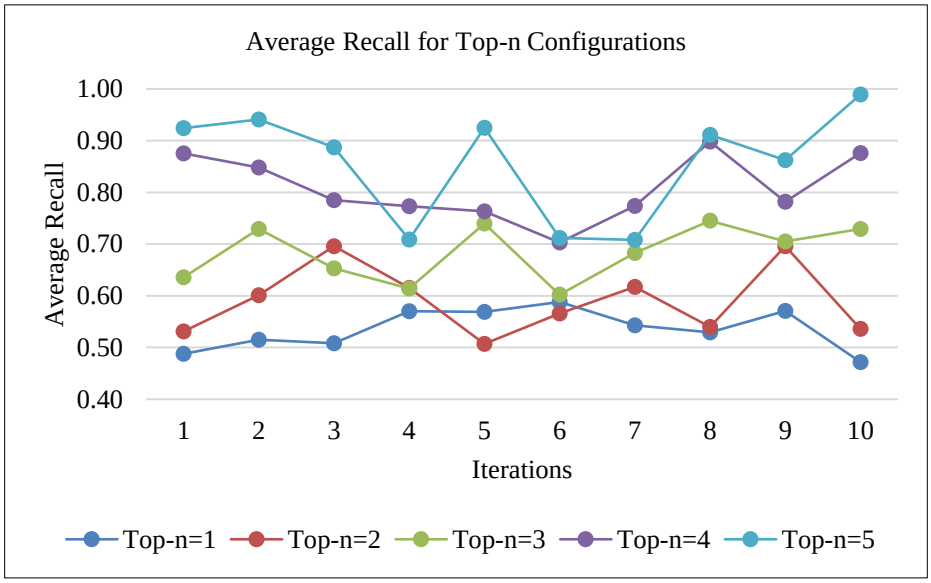

Fig. 2. Average recall for Top-n configurations

For top-*n*, the recall and hence the prediction of Iron Mask gets better with larger values of *n*. This is expected, because the larger the options to choose from, the higher the likelihood of discovering the item being searched. Similar results are also observed with τ, where lower values result in a much higher recall. These results demonstrate that Iron Mask is able correlate identity with historical postings to a fairly satisfactory level of performance. Even with tighter constraints of $n = 1$ or $\tau = 0.90$, the algorithm performs reasonably well and demonstrates that there is indeed a correlation between historical postings and user identity.

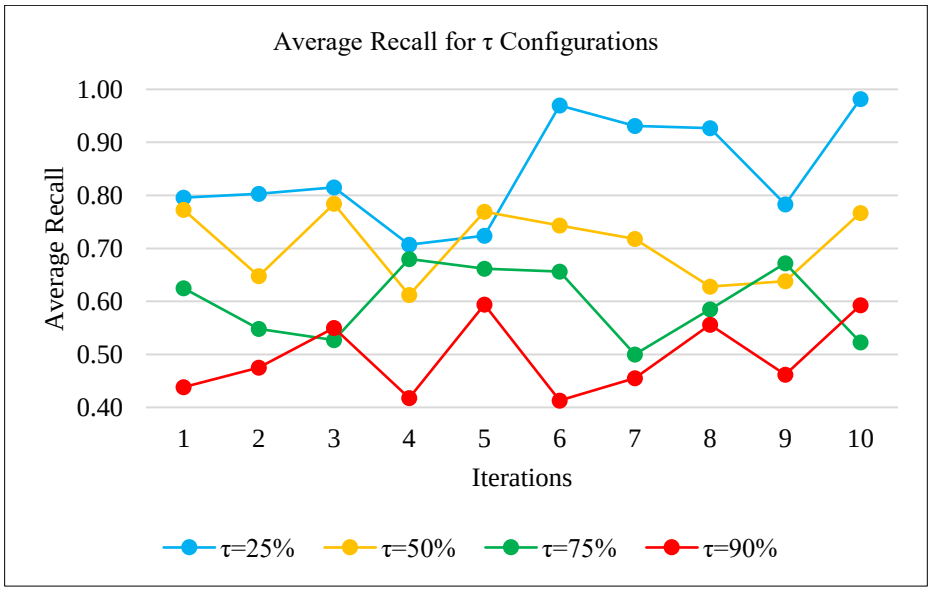

Fig. 3. Average recall for τ configurations

To reiterate, τ is used to control the level of confidence that Iron Mask can work with; false positives are prevented by setting a high level, which prevents Iron Mask from warning users of re-identification if the probability is low. On the other hand, top-*n* allows whiteprint identification without explicitly considering confidence; as long as the user's identity is among the most likely candidates, the Iron Mask algorithm warns the user. Ultimately, the Iron Mask algorithm can successfully predict the user's identity ahead of any re-identification attacks using either metric.

4.2 Effectiveness of TPP

To evaluate the effectiveness of TPP, we designed an online questionnaire-based survey. A total of 46 responses were recorded for the survey, and there were no specific user profile criteria for participation. The survey starts with displaying an arbitrarily selected question from the Quora database. Users are then asked to read the question, and then for the first input step, they are shown one of the answers for the question.

Two versions of the answer are shown: one with the generic “anonymous” label, and the other with a TPP determined from level of expertise. Users are asked to select the answer format that they find more credible from the two choices; a binary comparative choice. In the second input step, users are shown a different answer to the question and asked to select if the answer is trustworthy or not; a binary affirmative yes/no selection. The user label for this step is either “anonymous” or TPP, so some users see the “anonymous” label while others are shown a TPP label based on level of expertise.

Figure 4 presents a summary of the results from step 1, showing the total number of labels presented over the course of the survey, the number of positive selections for each label, and the number of negative selections as well. At first glance, it may look like the “anonymous” label was selected as the majority but this is actually not the case. Relatively, the generic label was selected by 17 out of the 46 users, while 29 users selected one of the TPP labels. In the breakdown shown for TPP labels, negative selections imply the “anonymous” label was preferred. Likewise, for the “anonymous” label, non-selection implies that one of the TPP labels were preferred. Further analysis reveals that out of the 17 selections, 10 were when the “novice” label was presented alongside with “anonymous”. This might be due to the surveyors perceiving “novice” and “anonymous” being relatively similar in terms of trustworthiness.

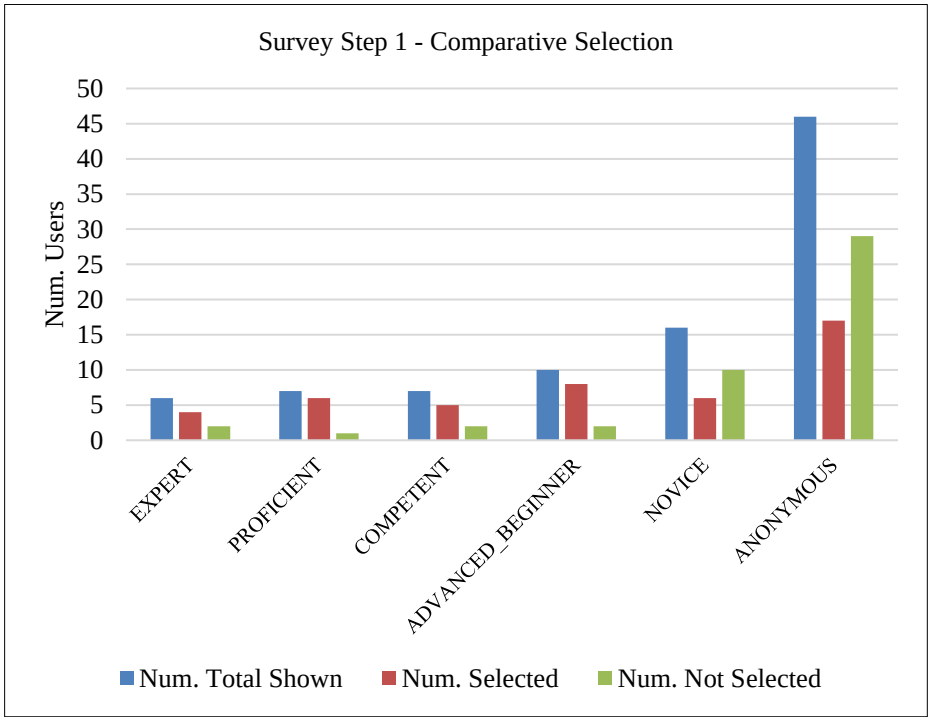

Fig. 4. Survey step 1 results

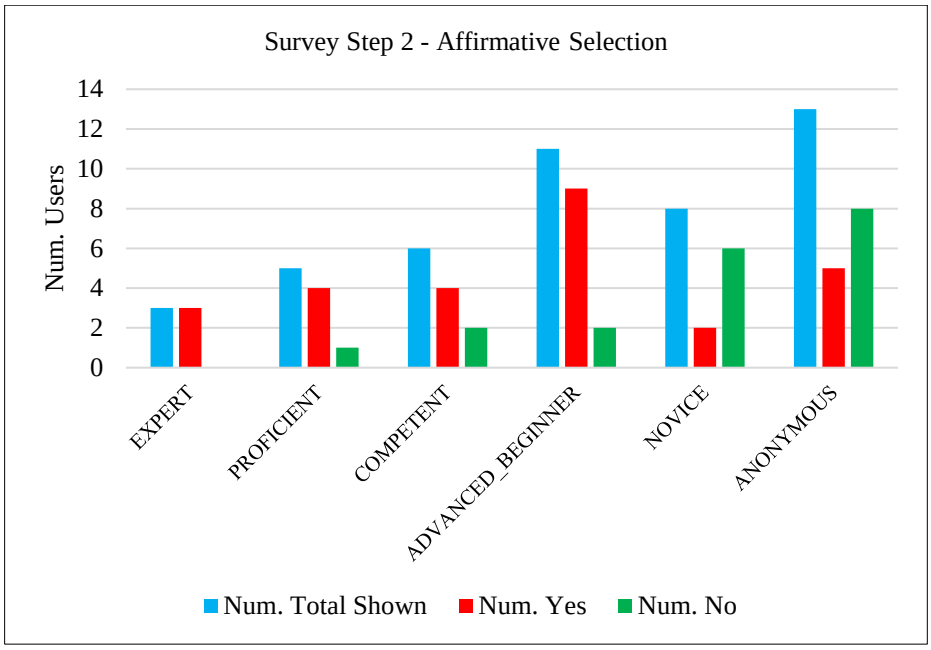

Fig. 5. Survey step 2 results

Figure 5 shows the results of step 2 of the survey, displaying the total number of instances of the labels presented, the number of "yes" selections implying the label was trustworthy, and the number of "no" selections when surveyors disagreed with the labels conveying trustworthiness. The results show that when the "novice" label was used, the users were more likely to disagree with the label conveying trustworthiness. As with step 1, the users seemed equally likely to select between "anonymous" and "novice". For the questions showing the higher-level expertise labels, the users agreed in the majority with the label being correlated with trustworthiness. This can be seen in both steps 1 and 2, implying there was a general consensus within the sample population about the effectiveness of the TPP labels.

5 Conclusion

The aim of this research was to improve upon existing anonymization options by investigating content as a quasi-identifier. In addition, this study explored inter-dependencies between identity, anonymity, and trust. The research questions were motivated by the need to provide anonymity for avoiding social stigmatization when users discuss about sensitive topics. Our results provide a satisfactory baseline for concluding that content created by users can reveal their identity, evaluated via machine learning methods. Moreover, our proposed trust-preserving pseudonyms have shown potential for providing a balance between credibility and anonymity based on user surveys.

For future work, there is room for improvement in the evaluation of trust-preserving pseudonyms within real social networks. Furthermore, one potential drawback of the whiteprint identification evaluation is the *cold-start problem*, where newly registered users may not have enough data to be classified using the trained probabilistic model. In this case, the naïve approach is to use the default "anonymous" label. Additional exploration can be done regarding how much data is necessary to tackle the cold-start issue and maintain the effectiveness of Iron Mask and TPPs. In other words, one research question we intend to explore in future research is how much data is too little. Moreover, we plan to incorporate Iron Mask into a health social network under construction, code named Cardea, which allows patients and medics to communicate with each other online within specialized, secure, private, and trusted areas for patient-patient, patient-medic and medic-medic conversations. Within Cardea, users can also create support groups based on mutual topics of interest and develop hierarchical connections with other users. We also plan to investigate alternative machine learning approaches to authorship identification, such as clustering and deep neural networks. Another area of interest is the contextualization of credibility by topic, whereby users' level of expertise could be granularized to topical expertise.

Acknowledgements

The authors would like to thank the Alberta Machine Intelligence Institute (Amii) for funding this research. Amii is a research lab at the University of Alberta, Edmonton, Canada, working to enhance understanding and innovation in machine intelligence, existing at the intersection of machine learning and artificial intelligence.

References

1. Hansen, D. L., Johnson, C.: Veiled Viral Marketing: Disseminating Information on Stigmatized Illnesses via Social Networking Sites. In: 2nd ACM SIGHIT International Health Informatics Symposium, pp. 247-254. ACM, Miami, Florida, USA (2012).
2. White, M., Dorman, S. M.: Receiving Social Support Online: Implications for Health Education. Health Education Research 16(6), 693-707 (2001).
3. Frost, D. M.: Social Stigma and its Consequences for the Socially Stigmatized. Social and Personality Psychology Compass 5(11): 824-839 (2011).
4. Hong, Y., Patrick, T. B., Gillis, R.: Protection of Patient's Privacy and Data Security in E-Health Services. In: 1st International Conference on Biomedical Engineering and Informatics, pp. 643-647. IEEE, Sanya, Hainan, China (2008).
5. Cofta, P.: Confidence, Trust and Identity. BT Technology Journal 25(2), 173-178 (2007).
6. Child, J.: Trust - The Fundamental Bond in Global Collaboration. Organizational Dynamics 29(4), 274-288 (2001).
7. Dwork, C.: Differential Privacy: A Survey of Results. In: Agrawal, M., Du, D., Duan, Z., Li, A. (eds.): TAMC 2008, LNCS 4978, vol. 4978, pp. 1-19. Springer Berlin Heidelberg (2008).

8. Narayanan, A., Shmatikov, V.: De-Anonymizing Social Networks. In: 30th IEEE Symposium on Security and Privacy, pp. 173-187. IEEE Computer Society, Washington, DC, USA (2009).
9. Beach, A., Gartrell, M., Han, R.: q-Anon: Rethinking Anonymity for Social Networks. In: 2nd International Conference on Social Computing, pp. 185-192. IEEE, Minneapolis, Minnesota, USA (2010).
10. Milhail S., Ilya L.: Methodologies of Internet Portals Users' Short Messages Texts Authorship Identification based on the Methods of Mathematical Linguistics. In: 8th International Conference on Application of Information and Communication Technologies, pp. 1-6. IEEE, Astana, Kazakhstan (2014).
11. Keretna, S., Hossny, A., Creighton, D.: Recognising User Identity in Twitter Social Networks via Text Mining. In: International Conference on Systems, Man, and Cybernetics, pp. 3079-3082. IEEE Computer Society, Washington, DC, USA (2013).
12. Johnson, A. M., Syverson, P., Dingledine, R., Mathewson, N.: Trust-Based Anonymous Communication: Adversary Models and Routing Algorithms. In: 18th Conference on Computer and Communications Security, pp. 175-186. ACM, Chicago, Illinois, USA (2011).
13. Sassone, V., Hamadou, S., Yang, M.: Trust in Anonymity Networks. In: Gastin P., Laroussinie F. (eds) CONCUR 2010, LNCS 6269, vol 6269, pp. 48-70. Springer Berlin Heidelberg (2010).
14. Maurer, F. K.: A Survey on Approaches to Anonymity in Bitcoin and other Cryptocurrencies. In: Mayr, H. C., Pinzger, M. (eds) INFORMATIK 2016, LNI, vol P-259, pp. 2145-2150. Gesellschaft für Informatik, Bonn (2016).
15. Bernstein, M. S., Monroy-Hernández, A., Harry, D., André, P., Panovich, K., Vargas, G. G.: 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. In: 5th International Conference on Weblogs and Social Media, pp. 50-57. AAAI, Barcelona, Spain (2011).
16. Taskar, B., Segal, E., Koller, D.: Probabilistic Classification and Clustering in Relational Data. In: International Joint Conference on Artificial Intelligence, pp. 870-878. Lawrence Erlbaum Associates Ltd., Seattle, Washington, USA (2001).
17. Zadrozny, B., Elkan, C.: Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In: 8th International Conference on Knowledge Discovery and Data Mining, pp. 694-699. ACM, New York, NY, USA (2002).
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825-2830 (2011).
19. Kameya Y., Hayashi K.: Bottom-Up Cell Suppression that Preserves the Missing-At-Random Condition. In: Katsikas S., Lambrinoudakis C., Furnell S. (eds) Trust, Privacy and Security in Digital Business. Lecture Notes in Computer Science, vol. 9830, pp. 65-78. Springer Cham (2016).
20. Dreyfus, S. E., Dreyfus, H. L.: A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition. California University Berkeley Operations Research Center (1980).