

# Detection of Violent Extremists in Social Media

Hamidreza Alvari, Soumajyoti Sarkar, Paulo Shakarian  
Arizona State University  
Tempe, USA  
{halvari, ssarka18, shak}@asu.edu

**Abstract**—The ease of use of the Internet has enabled violent extremists such as the Islamic State of Iraq and Syria (ISIS) to easily reach large audience, build personal relationships and increase recruitment. Social media are primarily based on the reports they receive from their own users to mitigate the problem. Despite efforts of social media in suspending many accounts, this solution is not guaranteed to be effective, because not all extremists are caught this way, or they can simply return with another account or migrate to other social networks. In this paper, we design an automatic detection scheme that using as little as three groups of information related to usernames, profile, and textual content of users, determines whether or not a given username belongs to an extremist user. We first demonstrate that extremists are inclined to adopt usernames that are similar to the ones that their like-minded have adopted in the past. We then propose a detection framework that deploys features which are highly indicative of potential online extremism. Results on a real-world ISIS-related dataset from Twitter demonstrate the effectiveness of the methodology in identifying extremist users.

**Index Terms**—Extremists, Social media, Feature Engineering

## I. INTRODUCTION

Last years have witnessed a huge rise in the threat of radicalized extremists groups who seek to commit insurgent attacks around the globe. While new technologies such as the Internet and social media are now being used by many users [3], [20], they have also been leveraged extensively by radicalized groups to make relationships with their audience and recruit new members [15]. Social media platforms such as Twitter are now being utilized by terrorist organizations, to directly communicate with their worldwide audience [11]. The free and unregulated nature of these tools, have helped extremists to easily form online communities and disseminate their beliefs and training materials, without a fear of getting censored on traditional media outlets.

Recently, social networks have begun to actively fight against these groups. In august 2016, Twitter which has been long believed to be the main propaganda for ISIS, finally took serious actions by shutting down over 36,000 ISIS related accounts [1]. Social media are primarily based on the reports they receive from their own users, or from specific teams assigned to mitigate the problem. Recently they began to explore more effective ways such as using algorithms (e.g. spam-fighting algorithms) to automatically detect any violent content as supplements to these reports and boost performance [16]. Despite the huge effort of these social networks in shutting down many accounts, this solution is not guaranteed to be effective, because not all extremists are caught this way, and the owners of suspended accounts can

simply return with another account or migrate to other social networks. Consequently, extra efforts need to be dedicated to proposing capabilities that could be deployed by authorities to combat radicalized extremists and mitigate their threats, regardless of the underlying social network platform.

**Present work.** In this paper, we design a detection scheme that using as little as three groups of information (inspired from the literature [8], [11], [19]), can determine whether or not a given username<sup>1</sup> belongs to an extremist user. Specifically, we use a dataset from Twitter [25] and first show that *extremist users on Twitter tend to adopt handles that follow similar patterns, in contrast to the normal users.* Then, a detection framework is proposed to identify if a given Twitter handle belongs to an extremist given its proximity to an existing set of extremist-related handles. We compare different supervised and semi-supervised approaches using the features from Twitter handle, profile information and content which are highly indicative of online extremism. To further show the significance of the features we conduct significance analysis of the features using the labeled instances and feature selection measure  $\chi^2$  and compare our results against char-LSTM which automatically extracts features.

**Contributions.** Our main contributions are thus summarized in below:

- We first demonstrate that extremists on Twitter are inclined towards adopting handles with similar patterns. To that end, we used the well-known Lavenstein ratio as a measure of distance between two Twitter handles and performed two-sample t-test to demonstrate that compared to normal users, extremists tend to choose similar handles.
- We propose three main groups of features, related to the Twitter handles, profile information and tweet-level content. Overall, our feature engineering scheme has 13 features which are then fed into different supervised and semi-supervised learners.
- Results on a real-world ISIS-related dataset demonstrate that the introduced features are effective in detecting online violent extremists in social media.

**Observations.** We make the following observations:

- The highest precision of 0.96 in identification of the extremists belongs to SVM. This is in line with the previous research that SVM performs well on textual data.

<sup>1</sup>In this work, we may use the terms usernames and handles interchangeably

- Among several approaches used in this work, char-LSTM and the semi-supervised approach LabelSpreading with RBF kernel achieve the equal and highest F1-score of 0.76. The fact that the LabelSpreading achieves comparable performance as char-LSTM, further demonstrates the effectiveness of the proposed feature engineering scheme, as LSTM has shown promising results in the literature while it does not use any hand-crafted features.
- Char-LSTM achieves a precision of 0.77 while maintaining a high recall of 0.76 on the positive class. This suggests that the memory module in LSTM can help in minimizing the number of false negatives.

## II. RELATED WORK

The explosive growth of the Web has raised numerous security and privacy issues. Mitigating these concerns has been studied from different aspects [6], [18], [23], [26]. For instance, several studies have focused on understanding extremism in social networks [4], [5], [8], [9], [11], [16], [17]. For instance, the work of [11], uses Twitter and proposes an approach to predict new extremists. They also determine if the newly created account belongs to a suspended extremist, and predict the ego-network of the suspended extremist upon creating her new account. Their approach integrates variants of the logistic regression with optimized search policies to detect the new accounts of returning suspended extremist users. They (1) use potential features of an account to predict if this account belongs to an extremist user, (2) determine if multiple accounts belong to the same extremist user, based on the fact that new account shall resemble the suspended account in different aspects, (3) predict whom the suspended extremist user is most likely to follow again, and finally (4) develop a network search policy to find the suspected users upon returning to a social network. Similar work of [8] uses tweets to build models to predict (1) if a pro-ISIS user's account will be suspended due to the extremist content, (2) which users will adopt and retweet ISIS content, and (3) which users will have interactions with pro-ISIS users. To do so, the authors use logistic regression and random forest classifiers for different types of prediction tasks. They deploy variety of features across different dimensions, such as user meta-data, network statistics and temporal patterns of activity. Two scenarios are then designed for each prediction task: a time independent (static) one which does not take into account the temporal dependencies, and a simulated real-time one by considering the timeline of content availability. The difference between these two studies is, authors of [11] also study other aspects including identifying multiple accounts for an extremist user, re-following suspended accounts' connections and searching for the suspended extremist users who might return to a social media.

Other works also seek to identify the extremist content in radicalized groups beyond ISIS. The work of [16] uses data from Jihadist website Ansar AlJihad Network to develop supervised learning and NLP techniques to automatically detect cyber-recruitment of extremist groups. A comparison

is done between classifiers naïve Bayes, logistic regression, classification trees, boosting and SVM, for labeling forum posts as either related or not related to the recruitment of extremist groups. They leverage the bag-of-words technique to convert the corpus into a term-document matrix, following the standard routine of the preprocessing techniques such as basic normalization and stemming. Similarly in [17], same Jihadist network along with their previously developed SVM classifier are used to automatically identify recruitment posts. Their previous work shall be served as pre-screening step to reduce the efforts made by human analysts to manually hand-label the documents. In their new study, the textual content of the dataset is analyzed with latent Dirichlet allocation (LDA) and fed into several time-series models to predict cyber-recruitment. This new research conducted by the same authors complemented their previous study, by applying current natural language processing and time series analysis techniques to forecast the recruitments.

Beyond these works, the work of [9] takes a different approach to track individual's behavioral indicators of home-grown extremism, using public and law enforcement data. The intuition is to use graph pattern matching to identify suspicious trajectories and potential radicalization over a dynamic heterogeneous graph associated with the fused data from public and law enforcement. The authors first develop a query pattern of radicalization and then run several graph pattern matching algorithms to detect and track the on-going radicalization. They develop the investigative simulation graph pattern matching technique, which is composed of required extension to the existing dual simulation graph pattern matching method to avoid over-matching. This approach provides analysts and law enforcement officials with the ability to find partial/full matches, given a query of radicalization, as well as the pace of the appearance of the radicalized extremists. As opposed to the above studies, in this paper, we make the first attempt on determining if a given Twitter handle belongs to an extremist user or not, using only little information gathered from the handle, profile and content.

## III. DATA PREPARATION

The dataset was collected from Twitter and consists of approximately 1.6M tweets that were posted using 25 extremism-related hashtags such as #AbuBakralBaghdadi, #ISIL, #ISIS, #Daesh, and #IslamicState. We construct our extremist labels (positive labels) by collecting a limited number of 150 suspended ISIS-related Twitter handles which were reported to the Twitter Safety account (@TwitterSafety) by normal users. To make a balanced labeled dataset, 150 random handles corresponding to normal users were also collected to serve as our negative labels.

Inspired by the literature [8], [11], [19], we define 3 major groups of overall 13 features, which could be leveraged to filter out less likely extremists. This way, we obtain 300K highly extremism related tweets from which we randomly pick a smaller sample of 3K handles who posted the tweets. The description of the dataset is shown in Table I.

TABLE I  
DESCRIPTION OF THE DATASET.

Name	Value	
Raw	1.6M	
Filtered	300K	
Unlabeled (sampled)	3K	
Labeled	Positive	Negative
	150	150

#### IV. METHODOLOGY

Here, we first present the introduced feature groups used to filter out less likely extremists from the data. Next, we will pose our research questions and seek to answer them.

##### A. Feature Engineering

We categorize the features used in this work into the following 3 major groups:

- 1) **Twitter handle’s related features:** this group contains 3 features related to the given handle, namely, length of the handle, number of unique characters in the handle, and complexity of the handle. To compute the complexity, Kolmogorov complexity is used, which is defined as the length of the shortest program to reproduce the handle on a universal machine such as Turing machine [13]. Since Kolmogorov complexity is computationally intractable, we use the Entropy of the handle as a way to approximate its complexity.
- 2) **Profile related features:** this group contains 7 features related to the profile of the user who posted the tweet, including the number of her followers, friends and tweets, the existence of profile’s description and location. Also, for the last two features in this group, we check if the account is verified and geo-enabled.
- 3) **Content related features:** we have the following 3 features related to the content of the given tweet: the number of URLs, the number of hashtags and the sentiment of the content. For the sentiment, we check if the content has a higher negative score than its positive score.

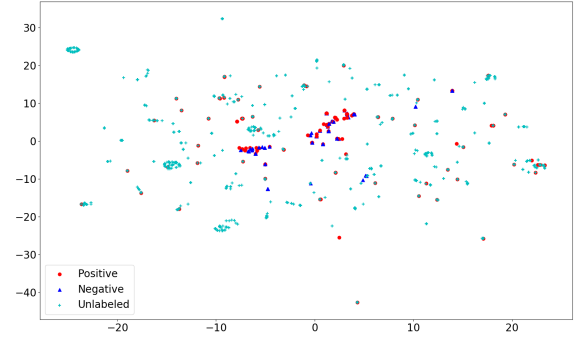
For the sake of visualization, a 2-D projection of the sample of the filtered dataset (using t-SNE transformation [22]) is depicted in Fig. 1. As it is seen, basic clustering techniques such as K-means will have difficulty to correctly assign labels to the unlabeled instances using only few existing labeled samples.

##### B. Research Questions

Having defined the feature engineering scheme in the previous section, here, we seek to answer the following research questions that will ultimately help identifying violent extremists in social media:

- **RQ1: Are extremists on Twitter inclined to adopt similar handles?**

Fig. 1. 2-D projection of the sampled filtered data using t-SNE transformation. Clustering techniques such as K-means will have difficulty to correctly assign labels to the unlabeled instances using only few existing labeled samples.



- **RQ2: Can we infer the labels (extremist vs. non-extremist) of unseen handles based on their proximity to the labeled instances?**

To answer the **first** question, for each pair of extremist users  $(i, j)$  we compute the similarity between their corresponding handles  $\langle s_i, s_j \rangle$  as follows.

$$Sim(s_i, s_j) = \frac{1 - L(s_i, s_j)}{\max(\text{len}(s_i), \text{len}(s_j))} \quad (1)$$

where  $L(s_i, s_j)$  is the well-known Levenshtein ratio [12] which is used as a measure of distance between the two strings  $s_1$  and  $s_2$ . We create a vector  $v_e$  whose elements are similarity scores between each pair of extremists. We repeat the procedure for each pair of extremist  $i$  and a normal user  $k$  and create a vector  $v_{en}$ . We conduct a two-sample t-test on  $v_e$  and  $v_{en}$  with the null and alternative hypotheses defined as  $H_0 : v_e \leq v_{en}$ ,  $H_1 : v_e > v_{en}$ . The null hypothesis is rejected at significance level  $\alpha = 0.01$  and p-value of  $p = 0.009$ , suggesting that extremists are biased towards adopting similar handles. Although this might seem a bit simplistic at the first sight, it has not been examined in the literature. Later, we will see how effective this simple idea could be in inferring the labels for unseen handles and help detecting the extremists by merely glancing at their handles.

To answer the **second** question, let us first obtain our feature spaces associated with the labeled and unlabeled instances, by converting each handle to a vector of 5 features. We use the following features: *length of the handle*, *maximum number of occurrence of a character in the handle*, *number of unique characters in the handle*, *number of digits that the handle starts with*, and *complexity of the handle*. Ultimately, these two feature spaces are fed as the inputs to the semi-supervised and supervised learners.

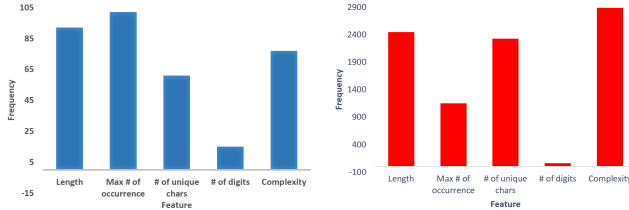
#### V. EXPERIMENTS

In this section, we first describe the learners used in this work and provide details on the parameters they use. Then, classification results are presented and finally significance of the features is discussed.

TABLE II  
COMPARISON OF THE METHODS ON THE LABELED DATA, FOR THE POSITIVE (EXTREMIST) CLASS.

Learner	Precision	Recall	F1-score
SVM	<b>0.96</b>	0.5	0.65
Char-LSTM	0.77	<b>0.76</b>	<b>0.76</b>
LabelSpreading (RBF)	0.85	0.69	<b>0.76</b>
Laplacian SVM	0.89	0.6	0.7
LabelSpreading (KNN)	0.83	0.67	0.73
Co-Training (SVM)	0.9	0.53	0.66
KNN	0.81	0.7	0.74
Gaussian NB	0.89	0.56	0.69
Logistic Regression	0.76	0.61	0.67
AdaBoost	0.88	0.58	0.69
Random Forest	0.79	0.71	0.74

Fig. 2. Frequency of each handle-related feature in the filtered dataset for (left) labeled and (right) unlabeled instances. The most frequent features for labeled and unlabeled instances are *Max # of occurrence of a character in a handle* and *Complexity of a handle*.



### A. Approaches

- **Semi-Supervised:** Laplacian support vector machines (SVM) [2], graph inference-based label spreading approach [21] with radial basis function (RBF) and K-nearest neighbor (KNN) kernels, and co-training learner [7] with two SVMs.
- **Supervised:** SVM, KNN, Gaussian naïve Bayes, logistic regression, adaboost, random forest, and Char-LSTM [24].

We note that supervised learners only use labeled instances for the training process, while semi-supervised algorithms use labeled and unlabeled instances [10].

For the sake of fair comparison, all algorithms were implemented and run in Python. Note for the methods that require special tuning of parameters, we performed grid search to choose the best set of parameters. Before going any further, we define the parameters used in each learner and then demonstrate their best picked values by our grid search.

- **SVM:** Tolerance for stopping criteria was set to the default value of 0.001. Penalty parameter  $C$  was set to 1 and linear kernel was used.
- **Char-LSTM:** This is similar to the character-aware models used for sequential word predictions. We adapt the neural network to a sequence classification problem where the inputs are the vector of one-hot encoding of each character of the handle and the output is the handle being classified as extremist or non-extremist. We set the maximum username length to 10, padding with zeros

where necessary. We use an embedding layer after the input layer to convert each username dimension to 16. This is fed to a single layer LSTM module having 30 units.

- **LabelSpreading (RBF):** RBF Kernel was used and  $\gamma$  was set to the default value of 20.
- **Laplacian SVM:** We used linear kernel and set the parameters  $C_l = 0.6$  and  $C_s = 0.6$ .
- **LabelSpreading (KNN):** KNN kernel was used and the number of neighbors was set to 5.
- **Co-training (SVM):** We followed the algorithm introduced in [7] and used two SVM as our classifiers. For both SVMs we set the tolerance for stopping criteria to 0.001 and the penalty parameter  $C = 1$ .
- **KNN:** The number of neighbors was set to 5.
- **Gaussian NB:** There were no specific parameter to tune.
- **Logistic regression:** We used the  $l_2$  penalty. We also set the parameter  $C = 1$  (the inverse of regularization strength) and tolerance for stopping criteria to 0.01.
- **Adaboost:** The number of estimators was set to 200 and we also set the learning rate to 0.01.
- **Random forest:** We used 200 estimators and the entropy criterion was used.

### B. Classification Results

We use tenfold cross-validation on the labeled data as follows. We first divided the set of labeled instances into 10 different sets of equal size. Each time, we held one set out for validation (we did this by removing their labels and adding them to the unlabeled instances). For the supervised learners, this set along with a set of the existing unlabeled samples are only used for the purpose of testing whereas for the semi-supervised setting, we use both sets in the training and testing phases. This procedure is performed for all approaches for the sake of fair comparison. Finally, we report the average of 10 different runs, using various evaluation metrics including precision, recall and F1-score in Table II.

**Observations.** We make the following observations:

- SVM achieve the highest precision of 0.96 in identifying online violent extremists, which shows the significance of the proposed feature set.
- The semi-supervised LabelSpreading (RBF) was able to perform as good as Char-LSTM and they both achieve the highest F1-score on identification of extremists. This along with the fact that Char-LSTM has shown promising results in the literature while it does not use any of our hand-crafted features, further demonstrates the effectiveness of the introduced feature engineering scheme.
- For char-LSTM, we achieve a precision of 0.77 while maintaining a high recall of 0.76 on the positive class. This suggests that the memory module in LSTM can help in minimizing the number of false negatives.

Overall, the observations we make here suggest that the answer to the **second question** is positive— using an existing set of labeled examples could help inferring the labels of unseen usernames.

TABLE III  
SIGNIFICANCE OF THE FEATURES USING THE LABELED INSTANCES AND  $\chi^2$ . THE MOST SIGNIFICANT FEATURE FOR THE LABELED SET IS # of unique characters in a handle.

Feature	$\chi^2$
Length of the handle	22.73
Max # of occurrence of a character	0.24
# of unique characters	<b>37.2</b>
# of digits the handle starts with	12.57
Complexity of the handle	3.18

### C. Significance of Features

We conduct significance analysis of the features using the labeled instances and feature selection measure  $\chi^2$ . The results in Table III suggest that the most significant feature is the *number of the unique characters in the username* while the least important one is the *maximum number of occurrence of a character in the username*. This observation further demonstrates that frequency and importance of the features in the labeled dataset are not necessarily in line with each other and in fact are inversely related in our case. In other words, although *maximum number of occurrence of a character in the username* is the most frequent feature in the labeled dataset, it is the least important feature in identification of online violent extremists according to the Fig. 2 where we depict the frequency of each feature for both labeled and unlabeled examples.

## VI. CONCLUSION AND FUTURE WORK

In this work, we presented a scheme that using as little as three groups of information related to the Twitter handle, profile and textual content of users, can determine if a given handle could belong to an extremist. The framework first uses highly indicative patterns related to extremism to filter out less likely extremists. Ultimately, high likely extremist are identified using only features related to their usernames.

In future, we would like to replicate the work by deploying more features and investigate if incorporating those features to the framework can lead to performance boost. We also plan to incorporate the feature space designed in this work into a semi-supervised learner as regularization terms in order to further increase the classification performance in detecting online violent extremists. Finally, since hand-labeling unlabeled examples is expensive, a valuable research direction would be to deploy active learning to enable iterative supervised learning to actively query for labels.

## VII. ACKNOWLEDGMENTS

This work was supported through DoD Minerva program.

## REFERENCES

- [1] Abutaleb, Y. "Twitter suspended 360,000 accounts for 'promotion of terrorism'." <http://www.reuters.com/article/us-twitter-terrorism-idUSKCN10T1ST>, August 2016.
- [2] M. Belkin, P. Niyogi, and V. Sindhwani. "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples." *JMLR*, vol. 7, no. Nov, pp. 2399-2434, 2006.
- [3] Alvari, Hamidreza, Sattar Hashemi, and Ali Hamzeh. "Detecting overlapping communities in social networks by game theory and structural equivalence concept." In *International Conference on Artificial Intelligence and Computational Intelligence*, pp. 620-630. Springer, Berlin, Heidelberg, 2011.
- [4] Benigni, Matthew C and Joseph, Kenneth and Carley, Kathleen M. "Online extremism and the communities that sustain it: Detecting the ISIS supporting community on Twitter." *PloS one*, 2017
- [5] Benigni, Matthew and Carley, Kathleen M. "From Tweets to Intelligence: Understanding the Islamic Jihad Supporting Community on Twitter." 2016 *International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation*
- [6] Beigi, Ghazaleh, and Huan Liu. "Privacy in social media: Identification, mitigation and applications." *arXiv preprint arXiv:1808.02191* (2018).
- [7] Blum, A., and Mitchell, T. "Combining labeled and unlabeled data with co-training." *COLT*, 1998.
- [8] Ferrara, E., Wang, W., Varol, O, Flammini, A., and Galstyan, A. "Predicting online extremism, content adopters, and interaction reciprocity." *SocInfo*, 2016.
- [9] Hung, B., Jayasumana, P., and Bandara, V. "Detecting Radicalization Trajectories Using Graph Pattern Matching Algorithms." *ISI*, 2016.
- [10] Alvari, Hamidreza, Paulo Shakarian, and JE Kelly Snyder. "Semi-supervised learning for detecting human trafficking." *Security Informatics* 6, no. 1 (2017): 1.
- [11] Klausen, J., Marks, C., and Zaman, T. "Finding Online Extremists in Social Networks." *INFORMS*, 2016.
- [12] "How python-Levenshtein.ratio is computed." URL <http://stackoverflow.com/questions/14260126/>
- [13] Li, Ming and Vitnyi, Paul M.B. "An Introduction to Kolmogorov Complexity and Its Applications." Springer, 2008.
- [14] Beigi, Ghazaleh and Jalili, Mahdi and Alvari, Hamidreza and Sukthankar, Gita. "Leveraging community detection for accurate trust prediction." *Academy of Science and Engineering (ASE)*, USA, 2014.
- [15] Overbey, L. A., McKoy, G., Gordon, J., and McKittrick, S. "Automated sensing and social network analysis in virtual worlds." *ISI*, 2010.
- [16] Scanlon, J., and Gerber, M. "Automatic detection of cyber-recruitment by violent extremists." *Security Informatics*, 2014.
- [17] Scanlon, J., and Gerber, M. "Forecasting Violent Extremist Cyber Recruitment." *IEEE Transaction on Information Forensics and Security*, 2015.
- [18] Alvari, Hamidreza, Paulo Shakarian, and JE Kelly Snyder. "A non-parametric learning approach to identify online human trafficking." In *Intelligence and Security Informatics (ISI)*, 2016 *IEEE Conference on*, pp. 133-138. IEEE, 2016.
- [19] Zafarani, R., Liu, H. "10 Bits of Surprise: Detecting Malicious Users with Minimum Information." *CIKM*, 2015.
- [20] Beigi, Ghazaleh, and Huan Liu. "Similar but Different: Exploiting Users' Congruity for Recommendation Systems." *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, Springer, 2018.
- [21] Zhou, D., Bousquet, O., and Lal, T., Weston, J., and Schlkopf, B. "Learning with local and global consistency." *NIPS* 2004.
- [22] L. van der Maaten and G. Hinton. "Visualizing High-Dimensional Data Using t-SNE." *JMLR*, 2008.
- [23] Beigi, Ghazaleh, Kai Shu, Yanchao Zhang, and Huan Liu. "Securing Social Media User Data-An Adversarial Approach." *Proceedings of the 29th on Hypertext and Social Media*, pp. 165-173, ACM, 2018.
- [24] Kim, Yoon, Yacine Jernite, David Sontag, and Alexander M. Rush. "Character-Aware Neural Language Models." In *AAAI*, pp. 2741-2749. 2016.
- [25] Alvari, Hamidreza and Shaabani, Elham and Shakarian, Paulo. "Early Identification of Pathogenic Social Media Accounts." In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 169-174, 2018.
- [26] Beigi, Ghazaleh, Ruocheng Guo, Alexander Nou, Yanchao Zhang, and Huan Liu. "Protecting User Privacy: An Approach for Untraceable Web Browsing History and Unambiguous User Profiles." In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 213-221. ACM, 2019.