Authorship Analysis on Dark Marketplace Forums

Martijn Spitters
TNO
The Hague, Netherlands
martijn.spitters@tno.nl

Femke Klaver
Vrije Universiteit
Amsterdam, Netherlands
femke@holmes.nl

Gijs Koot
TNO
The Hague, Netherlands
gijs.koot@tno.nl

Mark van Staalduinen
TNO
The Hague, Netherlands
mark.vanstaalduinen@tno.nl

***Abstract*—Anonymity networks like Tor harbor many underground markets and discussion forums dedicated to the trade of illegal goods and services. As they are gaining in popularity, the analysis of their content and users is becoming increasingly urgent for many different parties, ranging from law enforcement and security agencies to financial institutions. A major issue in cyber forensics is that anonymization techniques like Tor's onion routing have made it very difficult to trace the identities of suspects. In this paper we propose classification set-ups for two tasks related to user identification, namely alias classification and authorship attribution. We apply our techniques to data from a Tor discussion forum mainly dedicated to drug trafficking, and show that for both tasks we achieve high accuracy using a combination of character-level n-grams, stylometric features and timestamp features of the user posts.**

I. Introduction

Over the past few years anonymizing darknets have found increasing interest of users who wish to stay anonymous when online. Even though such networks can be a vehicle for whistleblowers and political dissidents to exchange information, their downside is that they also provide a platform for users with malicious intentions. In a previous study we have shown that Tor [1], currently the most popular and well-known darknet, exhibits a wide range of illegal content and activity like drug and weapon trade, money laundering, hacking services, child pornography, and even assassination services [2]. Much of the trading activity on Tor has been organized in online marketplaces, most of which are accompanied by a discussion forum. Probably the most notorious example is Silk Road. In October 2013, the FBI shut down the first version of this drug market and arrested its owner Ross William Ulbricht, who operated under the pseudonym 'Dread Pirate Roberts'. Silk Road 2.0 was launched only a month later, and stayed online for a year, until also this version was seized, as part of Operation Onymous [3]. However, shortly after the Silk Road shutdown there has been a rapid activity increase on other marketplaces, especially on Evolution and Agora. Since the recent exit scam of the Evolution marketplace (March 2015), a similar migration wave can be observed, suggesting that at least a substantial part of the vendors on darknet markets are not easily discouraged and will simply continue their business elsewhere.

For law enforcement to get a grip on such flexible, large-scale and often professionally organized trade of illegal goods and services, the availability of smart tools for monitoring and analysis of these marketplaces and forums is starting to become an urgent matter. For instance, content and social network analysis techniques may be useful for identifying key members of Dark Web communities (e.g. highly active, reputable vendors) [4],[5]. Furthermore, a major issue in cyber forensic investigations is that the advanced anonymization techniques like Onion Routing [6], on which Tor is based, have made it very difficult to trace the identity of a targeted user. However, although identity cues are scarce on the Dark Web, users do leave behind traces of *textual identity* in the form of their writing style on discussion forums. Also, a user may be characterized by the times of day he shows activity, i.e. starts a thread or posts a message. In this paper we propose classification set-ups that use stylometric and time-based features to support certain stages of the user identification process. The first task we have studied is alias classification, in which the central question is: "given the forum posts of two users, A and B, are they or are they not written by the same individual?" In other words: is B an alias of A? In the second task of authorship attribution we try to assign a small set of forum posts to their author. We conduct our experiments on data from a Tor marketplace forum called Black Market Reloaded (BMR), which was taken offline by its owner shortly after the seizure of Silk Road 1.

The remainder of this paper is structured as follows. In Section II we describe the most important related literature. Section III describes the data set we used for our classification experiments. Then, in Section IV, we go into detail about the features we used, and the classification set-ups for both tasks mentioned earlier. Our experimental results are presented in Section V. Finally, we discuss the potential, as well as the possible drawbacks of our techniques in Section VI and conclude in Section VII.

II. Related work

Most of the older research on authorship analysis focuses on attribution tasks with few authors and a substantial amount of text per author available, e.g. the seminal work by Mosteller and Wallace [7]. However, with the rise of the internet and especially social media, the focus of authorship analysis has shifted to attribution problems with larger numbers of possible authors and less and shorter texts per author available to train on [8], [9], [10], [11], [12]. In the past few decades, authorship attribution has been explored for many different types of online communication, including email [13], forums and newsgroups [14],[9], and Twitter [15].

A. Authorship attribution on short texts

The work in [9] provides a framework for authorship attribution of forum posts. They experiment with sets of 5, 10, 15 and 20 authors and use different data subsets for training, ranging from 10 to 30 messages per author. Using lexical, syntactic, structural and content-based features, they report an impressive accuracy of 97.69%. Without structural features and content specific features, the accuracy drops to 90%. Unfortunately the paper lacks some clarity concerning the selection process of the messages, which cover a time span of 20 years. The fact that the digital writing conventions have changed considerably since the 1990's may have influenced their performance.

Building upon the framework of [9], Abbasi and Chen developed a method called *Writeprints* [16]. Rather than using the same feature set for all authors, this method constructs a personal Writeprint for each author based on the author's key features. Features which are important for a particular author but not for another are used as pattern disruptor to decrease the level of stylometric similarity between two authors. Finally, they compare the Writeprint of a text to the Writeprint of each author to attribute the right author to a text. The Writeprints method is tested on four different corpora: email, Ebay comments, Java Forum and CyberWatch Chat. For each corpus, 100 authors were selected. The number of words per author ranged from around 1400 in the chat corpus to over 43000 words per author for the Java Forum[1]. Attribution accuracy ranged from 31.7% for the chat corpus to 91.3% for the eBay comments. With less authors, performance improved [16].

In [15] authorship attribution is applied to Twitter messages. Their dataset contains tweets of 14000 authors and experiments are conducted with different numbers of tweets per user, ranging from 20 to 200. Using character-level *n*-grams as features, they achieve an accuracy of over 70%. Including *@replies* of the messages is an important factor in achieving that result.

A particular problem in realistic authorship attribution situations – and more generally in text classification –, which is investigated by [17] is that of class imbalance, i.e. the number of available posts varies greatly between different authors. In [17] an approach is proposed which creates many short text samples for authors with few posts and less but longer text samples for authors with many posts. This method increases the accuracy for the minority authors with only a slight loss of accuracy for the majority authors.

The work in [12] uses topic-independent stylometric features and machine learning techniques for large-scale authorship attribution on a collection of blog posts written by 100,000 authors. Only a small sample of an author's posts was used for training. In over 20% of the cases the classifier was able to identify the correct author. By using confidence estimation, they were able to increase precision to 80% at half the recall.

B. Alias classification and detection

A specific problem which may arise when applying authorship attribution techniques to forum posts, is that individuals can operate under multiple user accounts, which are thus different *aliases* of the same person. A list of reasons for using multiple aliases is given in [18]. The most important reasons in our data set (a Tor marketplace forum dedicated to illegal trade, where most users are vendors or buyers) are:

1) the old alias has been banned by a moderator (e.g. because of abuse);
2) the old alias has lost the trust of other members (e.g. because of scamming);
3) a vendor creates an alias to positively review his own products or services;
4) the user wants anonymity because of illegal activities.

As also mentioned in [18], we should note here that in some cases, users will attempt to disguise the fact that they are using multiple aliases. Where some users will deem it sufficient to choose a username which is very different from their other one, the more cunning individuals may purposely alter their writing style to reduce the chance their aliases will be linked. The work in [19] addresses exactly this problem and tries to identify authors who intentionally obfuscate their writing style, but also authors who try to imitate the writing style of another author. Their conclusion is that the first task is harder than the latter.

In [20], an unsupervised clustering algorithm is used to detect users with multiple aliases. For each user in their data set two *pseudo users* are created by splitting the data of that user in two parts. With 50 messages per user on a set of 100 users they achieve an accuracy of 80%. When 125 messages are used, the performance is increased to 95%. An interesting observation in this research is that the attribution accuracy deteriorates significantly when they apply their technique to messages from multiple topics.

Besides authorship attribution, [16] also investigates alias detection. Rather than using an instance-level approach where each message is used separately to represent a user, they use an identity-level approach in which all messages from a user are combined to create a single profile. Similar to [20], the data from each user was first split up to create two pseudo users. On data from 100 different users, their *Writeprints* method resulted in an *F*-measure ranging from 49.91% for the CyberWatch Chat data to as high as 94.59% for the eBay comments.

The work of [18] does not only employ stylometric features to identify users with multiple aliases, but also includes time-based features, more specifically the average posting activity of a user per hour of day. Again, each user u in their corpus is split into two pseudo users u_{ia} and u_{ib}. The similarity of each user in the set $u_{1a}, u_{2a}, ..., u_{na}$ to each user in the set $u_{1b}, u_{2b}, ..., u_{nb}$ is computed, which yields a ranking of potential aliases for each u_{ia}. Accuracy is computed based on the position of the correct alias in the top N of these rankings. A combination of stylometric and time-based features yields the highest accuracy (on 50 users: 70% with correct alias at highest position; 85% with correct alias in top 3). In their follow-up work, the authors extend their time profiles with period of day, month, day and type of day, and evaluate these features on the tasks of author attribution and alias detection using data from a large web forum [21].

Finally, [22] try to detect sockpuppets[2] on Wikipedia using stylometric features. They found that a number of features also included in [9] were useful for sockpuppet detection. They also found that two additional features were helpful for similarity

[1] The texts extracted from the Java Forum contained programming code rather than natural language.

[2] A sockpuppet is an online identity, or alias, specifically created for the purpose of deception.

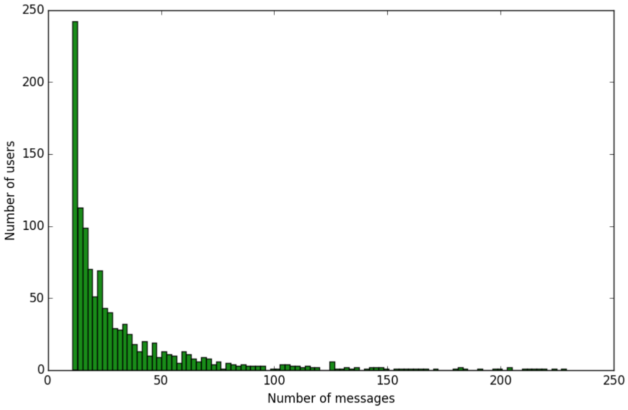

Fig. 1. The distribution of forum posts for all users with less than 250 messages.

detection: capitalization of the start of a sentence and white space between sentences. With their system, they were able to detect sockpuppets with an accuracy of 68%.

III. Description of the data

For our experiments we collected data from a multilingual marketplace forum called *Black Market Reloaded* (BMR). BMR was a Tor hidden service dedicated to the trade of illegal goods and services such as drugs, weapons, counterfeit money, stolen credit cards, and DDoS attacks. The takedown of Silk Road – BMR's largest competitor – by the FBI in October 2013, lead to an influx of new customers to BMR. By the end of November 2013, BMR was shut down by its owner because of a security breach that could compromise his identity.

A. Timespan and languages

Our full dataset spans over a year: from October 2012 to December 2013, and contains messages in 23 different languages, of which English (65%), German (9%), Dutch (3%), and French (2%) are best represented. The forum data was fully parsed, resulting in a structured database in which each post is linked to the thread it was posted in, the user who posted it, and the time it was created. The database contains a total of 92,333 posts from 8,348 different users, posted in 12,923 different threads.

B. Data selection

Because only the user-message distribution of the English subset is sufficient for attribution experiments with a substantial number of users, we only used this part of the data for this paper. The distribution of the forum posts (see Figure 1) show that in the English set more than half of all users posted less than 25 messages in total, and over 90% of all users posted less than 100. The average number of posts per user is 43. Because we need at least some minimum number of posts per user to train on, and also to create a more balanced data set, we chose to eliminate the outliers. Users who posted less than 50 or more than 250 messages in total were excluded from our experiments. This way we filter out users with a deviating role on the forum – e.g. the high-frequency posters are usually referral link spammers or administrators –, while still retaining enough users (177) to simulate a realistic application. A second restriction we applied is that a post has to contain at least five words. Finally, from each post, we removed all occurrences of the username of the user who posted it in order to minimize any influence of signatures.

TABLE I. The topic-independent features used for classification. The second column shows if and how the features are normalized; 'raw' means no normalization, 'chars' and 'words' mean normalization by total number of characters and words respectively.

Feature	**Norm.**
Global length	
- Number of characters	raw
- Number of words	raw
- Number of sentences	raw
Word/sentence length	
- Average word length	raw
- Average sentence length (in words)	raw
- Frequency of words with 1-20 characters (20 features)	words
Ratio features	
- Number of alphabetic characters	chars
- Number of uppercase characters	chars
- Number of digits	chars
- Number of function words	words
- Number of words containing both upper and lower case characters	words
Per-character/word frequencies	
- Frequency of alphabetic characters (26 features)	chars
- Frequency of punctuation marks (11 features)	chars
- Frequency of special characters (21 features)	chars
- Frequency of function words (293 features)	words
Richness	
- Total different words	words
- Number of short words (<4 characters)	words
- Hapax legomena (number of once-occurring words)	words
- Hapax dislegomena (number of twice-occurring words)	words

IV. Methodology

This section describes the details of our methodology. In IV-A we give an overview of the features we used for our classification tasks. Section IV-B describes our approach to alias classification, and IV-C goes into detail about our classification set-up for the authorship attribution problem.

A. Features

1) Topic-independent features: Most stylometric analysis is based on *topic-independent* features such as the length of the text and its words, the use of function words, interpunction and digits, and shallow syntactic patterns. These features are especially useful if authors write about many different topics, or if authorship attribution is applied to multi-domain data. As our baseline feature set, we selected a subset of the features also used in [9], [18] and [12], which we enriched with several additional features. Table I lists the topic-independent features we used in our experiments. For the English function words, we used the list as given in [12]. As shown in Table I, most of the features were normalized by the global length of the post (either number of characters or number of words).

2) Character n-grams: Besides these topic-independent stylometric features, we will also explore character n-grams as features to profile users. Character n-grams have been shown to perform very well on authorship attribution tasks in previous research [23], [24], [25], [15]. In [11] adding character n-grams improves the results of the extensive Writeprints framework previously proposed in [9]. Character-level n-grams are robust in many ways. First of all they can be applied language independently. In the study presented in [24], n-gram language models even outperform all other feature sets for authorship attribution in English, Greek and Chinese. Furthermore, character n-grams implicitly contain many different features. For example, they contain information about topic, punctuation use, occasional or typical spelling errors by users, but also about the use of function words. Without requiring NLP techniques like sentence splitting and morphological analysis, which usually perform very poorly on short, noisy texts like forum posts or tweets [15], character n-grams also encode typical behaviour concerning white space and capitalization use at sentence boundaries, and smooth out morphological variants of words. In our experiments we used n=3 (trigrams), and left the text unaltered, i.e. we keep capitalization, interpunction, and use of special characters intact.

As pointed out by [12], a possible pitfall of using character n-grams is that they may be influenced by topic, and that a classifier learns similarities in the *content* of the text rather than in the actual *writing style* of the user. However, while this may be an issue for their type of data (blog posts), in our marketplace forum the topic of a post usually refers to the merchandise its poster wants to sell, review or buy, which we consider as just another aspect by which we can profile that particular individual. Also, at this stage, we are not trying to apply authorship attribution across different domains or social network platforms.

3) Time-based features: As we are applying authorship attribution on forum data, there is an additional feature dimension we can exploit for our user profiles, namely: *time*. The recent study presented in [18] shows that time profiles perform significantly better than their set of topic-independent features, and that a combination of the two feature categories increases accuracy even further. We follow their approach and create normalized time profiles consisting of 24 features, each of which represents the activity percentage of a certain user during that particular hour of day. We should note here that in their follow-up work the authors of [18] got better results using a combination of different abstractions of these features (period of day and month) on a large web forum [21]. However, using these versions of the features decreased the results on our data set, possibly because we have much less posts available per user and thus need the specificity of the hour of day for the time profile to become effective. Figure 2 illustrates the hour of day feature in a heatmap, where the intensity of a cell's color represents a user's activity at that time. For each user two *pseudo users* (A and B) are created by randomly splitting that user's set of posts in two parts. Each row in Figure 2 represents the activity of a pseudo user. The heatmap clearly shows the similarities between each pair of the same individual, for instance user 972813 tends to be active almost the entire day, while user 972889 mostly posts between 18:00 and 22:00, suggesting he has a daytime job.

We have experimented with different combinations of these three feature categories. The results of these experiments are presented in Section V. How exactly the feature vectors which are fed to our classification algorithms are created, depends on the task and will therefore be explained in detail in the following two sections: IV-B and IV-C.

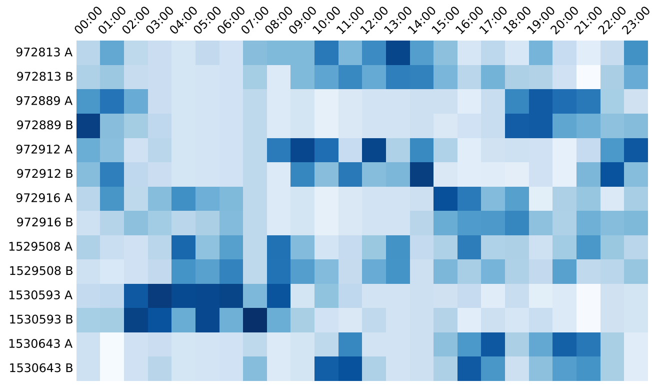

Fig. 2. Heatmap showing user activity per hour of day. For each user two *pseudo users* (A and B) are created by randomly splitting that user's data in two parts. Each row represents the activity of a pseudo user.

B. Alias classification

The first task we investigate in this paper is alias classification. The central question in this task is: "given the forum posts of two users, A and B, are they or are they not written by the same individual?" In other words: is B an alias of A? So, the classifier input at test time requires two instances, and the predicted class is either *yes* or *no*. Because we do not have the ground truth about user aliases in our data set, we follow [18] and simulate aliases by creating two pseudo users from the data of a single user. This means we randomly split the set of posts of a user u_i in two subsets, which represent pseudo user u_{ia} and u_{ib}. We then compute the feature vector for each post in the subset and use the mean of these vectors (or *centroid*) to represent the pseudo user. We should note here that there is a possibility that some of the different usernames in our data set are actually aliases used by the same individual. However, if this is the case, it will not have enhanced our results in any way.

Because the alias classifier has to learn from and predict classes for *pairs* of instances rather than for single instances, it is not straightforward to use a traditional machine learning algorithm such as support vector machines (SVMs) for this task. Where [18] use a similarity-based approach to *detect* potential aliases by *ranking* candidate pseudo users given a certain target pseudo user, we implement our classifier as a decision function based on some similarity threshold θ. Like [18], we compute the similarity of each pseudo user in the set $\{u_{1a}, u_{2a}, ..., u_{na}\}$ to each pseudo user in the set $\{u_{1b}, u_{2b}, ..., u_{nb}\}$. Because we have 177 users in our data set (see Section III), this means we compute a total of $177^2 = 31,329$ similarities, of which only 177 represent the positive class, i.e. the similarities between u_{1a} and u_{1b}, u_{2a} and u_{2b}, and so on. As a similarity measure we use the cosine of the angle between two pseudo user's feature vectors:

$$\cos(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\sum_{i=1}^{n}\mathbf{x}_i \times \mathbf{y}_i}{\sqrt{\sum_{i=1}^{n}(\mathbf{x}_i)^2} \times \sqrt{\sum_{i=1}^{n}(\mathbf{y}_i)^2}} \quad (1)$$

The cosine similarity is computed separately for each of our feature categories and then combined into a weighted average.

We define our decision function for a similarity score σ given a threshold θ simply as:

$$prediction(\sigma) = \begin{cases} yes & \text{if } \sigma \geq \theta \\ no & \text{if } \sigma < \theta \end{cases} \tag{2}$$

The results of our alias classification experiments will be presented in section V-A.

C. Authorship attribution

Our authorship attribution task is set up as a regular multi-class classification task, in which each candidate author represents a class. However, rather than creating an instance for each post, we combine multiple posts into a single instance, which results in more appropriate representations of a user's writing style, as irregular or uninformative posts are ironed out. More specifically, an instance is created by taking the average of the feature vectors of n multiple randomly selected posts of a user u_i. We experimented with different numbers of posts per instance to investigate the trade-off between the advantage of these more robust representations and the total number of available instances per user.

For our classification experiments we use support vector machines (SVMs), one of the most widely used machine learning algorithms, which has successfully been applied for numerous text classification problems, including authorship attribution [9], [21], [26], [27], [28].

As we are using character trigrams as features, our data is very sparse, with a huge number of features (± 70,000). For such data, linear SVM classification has become one of the most promising and scalable learning techniques. We use LIBLINEAR, a library for large-scale linear classification, which was developed especially to deal with sparse data with huge numbers of instances and features [29].

Given a set of instance-label pairs $(x_i, y_i), i = 1, ..., l, x_i \in R^n, y_i \in \{1, +1\}$, linear SVM solves the following unconstrained optimization problem:

$$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi(w; x_i, y_i) \tag{3}$$

where $\xi(w; x_i, y_i)$ is a loss function, and $C > 0$ is a penalty parameter [29]. Our multi-class data is trained in a one-vs-the-rest fashion. Our authorship attribution results will be presented in section V-B.

V. Experimental Results

A. Alias classification results

As mentioned in Section IV-B, our alias classifier applies a simple decision rule based on some similarity threshold θ. We evaluate our classifier by performing a threshold sweep and computing classification performance at each threshold on our entire data set ($177^2 = 31,329$ similarities, see Section IV-B). If the similarity between the feature vectors of two pseudo users u_{ia} and u_{jb} is lower than the threshold, the classifier assigns the class *no* (u_{ia} is not an alias of u_{jb}), and otherwise it assigns the class *yes* (u_{ia} is an alias of u_{jb}).

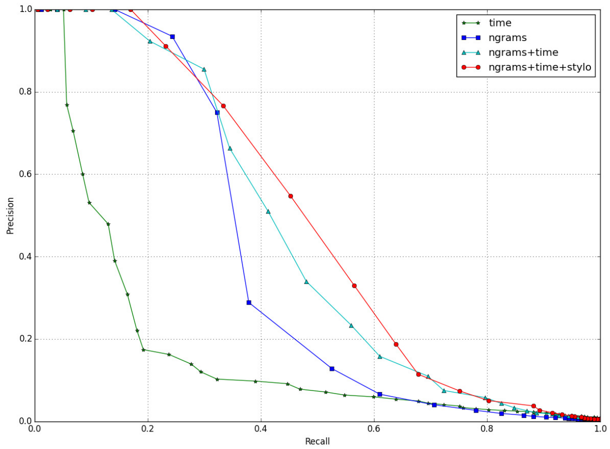

Fig. 3. Classification performance in terms of precision and recall at various thresholds. Each line corresponds to a different feature combination.

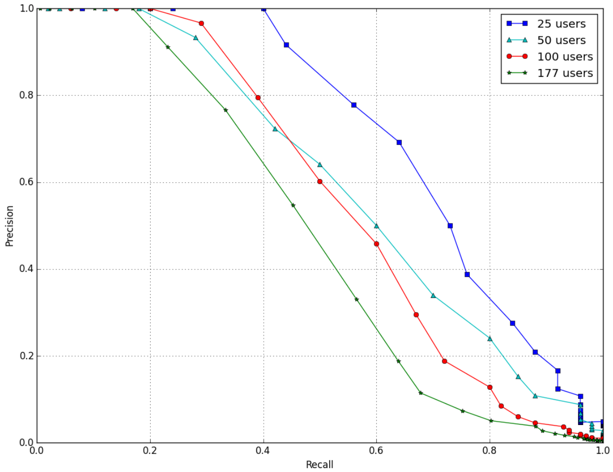

Fig. 4. Classification performance in terms of precision and recall at various thresholds. The performance is plotted for different numbers of users (25, 50, 100, and 177.

Figure 3 shows the classification performance in a Precision/Recall curve. Precision is the percentage of correctly classified positives (*true positives*) from all instances classified as positive. Recall (also called the *true positive rate* or *sensitivity*) is the percentage of true positives of all actual positives in the data set[3]. The graph shows the performance for different feature categories (time-based and character n-grams) and combinations thereof (n-grams + time-based, and n-grams + time-based + stylometric). Figure 3 clearly shows that the different feature categories reinforce each other. Using

[3] In our data, a pair of feature vectors from the same user (e.g. u_{1a} and u_{1b}) is a positive instance, and a pair of feature vectors from different users (e.g. u_{1a} and u_2b) is a negative instance. As noted before, we use the classes *yes* and *no* for positives and negatives respectively.

TABLE II. CLASSIFICATION ACCURACY FOR DIFFERENT FEATURE COMBINATIONS. ACCURACY AT TOP n IS THE PERCENTAGE OF CASES IN WHICH THE CORRECT AUTHOR WAS IN THE TOP n MOST HIGHLY RANKED CANDIDATES.

Feature combination	**Accuracy at top n**		
	n=1	n=3	n=5
Time	0.06	0.14	0.21
Stylometric	0.61	0.76	0.81
Time + Stylometric	0.66	0.79	0.85
Character n-grams	0.84	0.92	0.95
Time + Character n-grams	0.85	0.94	0.96
Time + Stylometric + Character n-grams	0.88	0.95	0.97

just the time-based features results in a steep drop in precision at a low recall (0.05-2). Obviously, the time profile of a user's posting behaviour alone lacks sufficient distinctiveness. The curves of the feature combinations however show a much more gradual decrease of precision toward the increase of recall (which is achieved by lowering the threshold) and the overall best performance is achieved by using a weighted combination of all three feature categories. The optimal weights for combining the feature categories were found by performing a weight combination sweep on a smaller data set from another forum, and were as follows: n-grams: 0.7, time-based: 0.2, stylometric: 0.1.

As we create the negatives in our set by comparing each of our 177 users with all 176 other users, the class distribution is extremely unbalanced (177 positives and 31,152 negatives). The less users, the more even the positives-negatives ratio becomes in the data. Presumably, with less users there will be less confusion and therefore classification performance will increase. To test this, we plotted Precision/Recall curves for different numbers of users in Figure 4. For each number of users, we sampled the users randomly from the total set of 177 users. We did this three times and plotted the average performance over those three runs to smoothen out possible artifacts of the random draws. The results confirm that the classification performance increases when we use smaller numbers of users, even though the difference between 100 and 50 users is relatively small. At low recall, precision is even higher with 100 users than with 50 users.

B. Authorship attribution results

As described in Section IV-C, we train a linear SVM to perform authorship attribution. The classifier is trained on multi-class data, where the users represent the classes, and the instances are the average feature vectors of 5 randomly selected posts of a particular user. The results were produced using 5-fold cross validation on our set of 177 users with 50-250 posts (91 per user on average). Table II shows the classification accuracy for different combinations of our feature categories. The second column contains the regular accuracies (the percentage of correct authorship attributions), which is 88% if we use all available features. However, because in a practical forensic investigation it can already be useful to have a tool which suggests a few probable candidate authors for further inspection, we also evaluated the classification performance at the top 3 and 5 highest ranked users (columns three and four in Table II respectively). The highest accuracy is 0.97 (bottom-right of the table), which means that when all features are used, in 97% of the cases the correct author is among the five highest ranked users. In the subsequent experiments described below, all three feature categories are used, as this produces the best results.

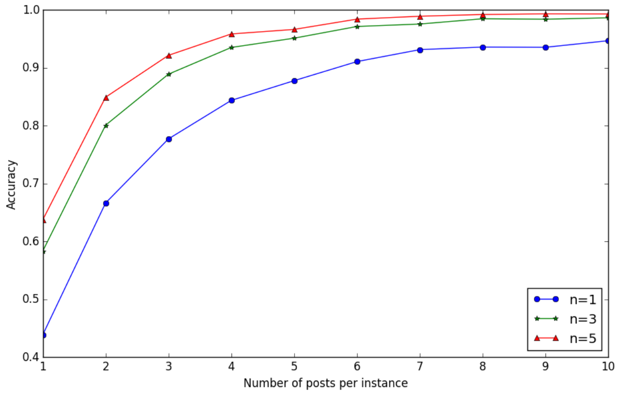

Fig. 5. Classification accuracy at top n with different numbers of posts per instance. Accuracy at top n is the percentage of cases in which the correct author was among the n most highly ranked candidates.

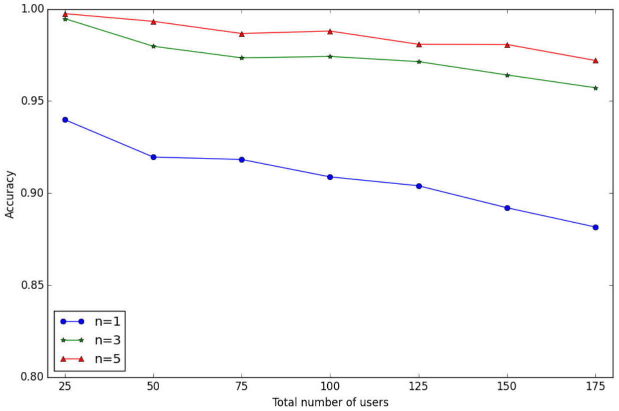

Fig. 6. Classification accuracy at top n with different numbers of users. Accuracy at top n is the percentage of cases in which the correct author was among the n most highly ranked candidates.

As mentioned above, we use 5 randomly selected posts of a particular user to create a single instance. We do this to create better, more robust representations in which irregularities are ironed out. To see the influence of the number of posts we use to create an instance, Figure 5 plots the accuracy at 1-10 posts per instance. We can see that there is already a dramatic improvement if we use two instead of one post to create an instance. At 7 posts, the optimum seems to be reached. Interestingly the classifier can cope fine with only a few training instances per user: as our data contains 91 posts per user on average, and the train/test split ratio is 4:1, if we use 10 posts for a single instance, the SVM only gets 7 instances per user to train on. Evidently, it is better to train on a small number of very representative instances per class, than on a lot of less representative instances.

To see the influence of the total number of users in the training data, we trained classification models with user sets of different size. For each number of users, the user set was randomly drawn. We repeated this three times and computed the average performance over those three runs. The average accuracy at different numbers of users is shown in Figure 6. We see that accuracy decreases gradually from 94 to 88% for the n=1 case if we add more users to the model. In the top 3 and 5 cases, there is only a minor decrease (99.4-95.7% and 99.7-97.2%). The nearly perfect performance with 25 users is however dubious, as the a priori chance of the correct author being in the top 5 highest ranked authors in that case is 20%.

VI. Discussion

We have shown that, at least within an online trading community, it is possible to identify user aliases and attributing (small sets of) forum posts to the correct author. The achieved performance indicate that the classification techniques are probably already mature enough to be useful in practice, especially for the task of authorship attribution. These techniques could for instance be helpful in investigations into organized drug trafficking to find out which user accounts registered on a marketplace are potentially operated by the same individual, or which accounts are shared by multiple users. We should note, however, that they are unlikely to work if users intentionally alter their writing style to reduce the chance their aliases will be linked. An interesting study related to this issue is that in [30], which demonstrates how sensitive state-of-the-art authorship attribution techniques are to manual circumvention methods.

Even though we have only shown that our classifiers work well on a single marketplace forum, there is no reason why they would not perform as well across *multiple* marketplace forums, e.g. for identifying shared accounts or aliases. However, as our classifiers rely heavily on the character-level n-gram features (See Figure 3 and Table II), it is very uncertain how well these techniques will work across different domains and platforms. As mentioned in Section IV-A, character n-grams also encode information about the discussed topics, so besides a user's characteristic writing style, the classifier will also learn what topics this user typically talks about. For example, a drug vendor will discuss and promote his merchandise on the marketplace forums, but it is very unlikely that he would do the same on his Facebook account. Also, a user's typical activity profile on a darknet market does not have to be similar to his activity profile on clearweb social media channels, so the time-based features are probably also less useful for these purposes. Although cross-platform authorship attribution is beyond the scope of this paper, it is important to keep these limitations in mind. It would be very interesting to further investigate how authorship attribution and alias classification techniques can contribute to cross-platform user linking, e.g. in combination with additional features like biographical attributes, specific entities mentioned in the text (for instance geographical), gender markers, pointers for occupation or nationality, relational, and even opinion or emotion cues (see e.g. [31]).

Of course, using de-anonymization techniques has privacy implications. Where anonymity can be essential for e.g. activists, journalists, whistleblowers and political dissidents to safely communicate and exchange information, it is also increasingly abused by criminals to deploy their illegal activities. For instance, anonymity networks like Tor nowadays harbor many underground markets dedicated to the trade of illegal products and services. Tools providing more transparancy in cyberspace are necessary to counter this often large-scale and professionally organized illegal trade, but as with so many other tools and techniques, in the wrong hands they may be used for undesirable purposes. An in-depth discussion about the privacy consequences of authorship attribution techniques is beyond the scope of our research. The interested reader is reffered to e.g. [30], [32], in which the role of stylometry in privacy and anonimity is discussed to a higher extent.

VII. Conclusions and Next steps

In this paper we introduced classification set-ups that use stylometric, character-level n-gram, and time-based features to support certain stages of the de-anonymization process in online forensic investigations. We conducted our experiments on English posts from a Tor marketplace forum called Black Market Reloaded.

For the task of alias classification, in which the classifier has to decide whether or not two user accounts are aliases, we achieve a classification precision (of the positive class) of 91% at a recall of 25% for 177 users, and a precision of 92% at a recall of 45% for 25 users. We use a simple approach which makes its classification decisions based on whether or not the similarity between the posts of two user accounts exceeds a given threshold.

In the second task of authorship attribution, we train a support vector machine that assigns a small set of forum posts to their author. If we use all three feature categories, we achieve a classification accuracy of 88% when the classifier is trained on 177 users. If we train the classifier on less users, accuracy goes up to 94%. One of the main technical contributions of our work is that we create training instances by combining multiple randomly selected posts of a user. Our results show that this yields better, more robust representations in which irregularities are ironed out. If we use more posts to create an instance, accuracy goes up to 95% for 177 users.

Even though stylometric and time-based features have been combined in previous work for the task of alias detection [18], we are not aware of earlier attempts to use this combination for the task of author attribution. Furthermore, we have shown that adding a third feature category, namely character-level n-grams, increases classification accuracy hugely with 22%.

For future work, we plan to incorporate more features. One specific idea is to create features based on entities extracted from the text, like locations, names of other forum users, and products. Sentiment toward such entities could add an another interesting dimension. We would also like to explore ways to classify a user's native language and add that as a feature. On the longer term we want to look into cross-domain and -platform user linking. However, besides the privacy issues, a major hurdle for this is that there are no suitable labeled data sets available.

References

[1] The tor project. [Online]. Available: https://www.torproject.org

[2] M. Spitters, S. Verbruggen, and M. v. Staalduinen, "Towards a comprehensive insight into the thematic organization of the tor hidden services," in *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*. IEEE, 2014, pp. 220–223.

[3] Operation onymous. [Online]. Available: https://www.europol.europa.eu/content/global-action-against-dark-markets-tor-network

[4] G. L'Huillier, S. A. Ríos, H. Alvarez, and F. Aguilera, "Topic-based social network analysis for virtual communities of interests in the dark web," in *ACM SIGKDD Workshop on Intelligence and Security Informatics*. ACM, 2010, p. 9.

[5] S. A. Ríos and R. Muñoz, "Dark web portal overlapping community detection based on topic models," in *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*. ACM, 2012, p. 2.

[6] D. Goldschlag, M. Reed, and P. Syverson, "Onion routing," *Communications of the ACM*, vol. 42, no. 2, pp. 39–41, 1999.

[7] F. Mosteller and D. L. Wallace, "Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers," *Journal of the American Statistical Association*, vol. 58, no. 302, pp. 275–309, 1963.

[8] S. Argamon, M. Šarić, and S. S. Stein, "Style mining of electronic messages for multiple authorship discrimination: first results," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 475–480.

[9] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–393, 2006.

[10] K. Luyckx and W. Daelemans, "Authorship attribution and verification with many authors and limited data," in *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2008, pp. 513–520.

[11] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Transactions on Information Systems (TOIS)*, vol. 26, no. 2, p. 7, 2008.

[12] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song, "On the feasibility of internet-scale author identification," in *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 2012, pp. 300–314.

[13] O. De Vel, "Mining e-mail authorship," in *Proc. Workshop on Text Mining, ACM International Conference on Knowledge Discovery and Data Mining (KDD2000)*, 2000.

[14] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *Intelligent Systems, IEEE*, vol. 20, no. 5, pp. 67–75, 2005.

[15] R. Layton, P. Watters, and R. Dazeley, "Authorship attribution for twitter in 140 characters or less," in *Cybercrime and Trustworthy Computing Workshop (CTC), 2010 Second*. IEEE, 2010, pp. 1–8.

[16] A. Abbasi and H. Chen, "Visualizing authorship for identification," in *Intelligence and Security Informatics*. Springer, 2006, pp. 60–71.

[17] E. Stamatatos, "Author identification: Using text sampling to handle the class imbalance problem," *Information Processing & Management*, vol. 44, no. 2, pp. 790–799, 2008.

[18] F. Johansson, L. Kaati, and A. Shrestha, "Detecting multiple aliases in social media," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 1004–1011.

[19] S. Afroz, M. Brennan, and R. Greenstadt, "Detecting hoaxes, frauds, and deception in writing style online," in *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 2012, pp. 461–475.

[20] J. Novak, P. Raghavan, and A. Tomkins, "Anti-aliasing on the web," in *Proceedings of the 13th international conference on World Wide Web*. ACM, 2004, pp. 30–39.

[21] F. Johansson, L. Kaati, and A. Shrestha, "Time profiles for identifying users in online environments," in *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*. IEEE, 2014, pp. 83–90.

[22] Z. Bu, Z. Xia, and J. Wang, "A sock puppet detection algorithm on virtual spaces," *Knowledge-Based Systems*, vol. 37, pp. 366–377, 2013.

[23] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *Proceedings of the conference pacific association for computational linguistics, PACLING*, vol. 3, 2003, pp. 255–264.

[24] F. Peng, D. Schuurmans, S. Wang, and V. Keselj, "Language independent authorship attribution using character level language models," in *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 267–274.

[25] E. Stamatatos, "A survey of modern authorship attribution methods," *Journal of the American Society for information Science and Technology*, vol. 60, no. 3, pp. 538–556, 2009.

[26] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, "Authorship attribution with support vector machines," *Applied intelligence*, vol. 19, no. 1-2, pp. 109–123, 2003.

[27] G.-F. Teng, M.-S. Lai, J.-B. Ma, and Y. Li, "E-mail authorship mining based on svm for computer forensic," in *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*, vol. 2. IEEE, 2004, pp. 1204–1207.

[28] C. Sanderson and S. Guenter, "Short text authorship attribution via sequence kernels, markov chains and author unmasking: An investigation," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2006, pp. 482–491.

[29] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[30] M. Brennan, S. Afroz, and R. Greenstadt, "Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity," *ACM Transactions on Information and System Security (TISSEC)*, vol. 15, no. 3, p. 12, 2012.

[31] M. J. Edwards, A. Rashid, and P. Rayson, "A service-indepenent model for linking online user profile information," in *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*. IEEE, 2014, pp. 280–283.

[32] J. R. Rao and P. Rohatgi, "Can pseudonymity really guarantee privacy?" in *USENIX Security Symposium*, 2000.