Johansson *et al. Secur Inform (2015) 4:7*
DOI 10.1186/s13388-015-0022-z

Security Informatics
a SpringerOpen Journal

RESEARCH Open Access

Timeprints for identifying social media users with multiple aliases

Fredrik Johansson[1*†], Lisa Kaati[1,2†] and Amendra Shrestha[2†]

Abstract

Many people who discuss sensitive or private issues on social media services are using pseudonyms or aliases in order to not reveal their true identity, while using their usual, non-private accounts when posting messages on less sensitive issues. Previous research has shown that if those individuals post large amounts of user-generated content, stylometric techniques can be used to identify the author based on the characteristics of the textual content. In this article we show how an author's identity can be unmasked in a similar way using various time features (e.g., period of the day and the day of the week when a user's posts have been published). We combine several different time features into a *timeprint*, which can be seen as a type of fingerprint when identifying users on social media. We use supervised machine learning (i.e., author identification) and unsupervised alias matching (similarity detection) in a number of different experiments with forum data to get an understanding of to what extent timeprints can be used for identifying users in social media, both in isolation and when combined with stylometric features. The obtained results show that timeprints indeed can be a very powerful tool for both author identification and alias matching in social media.

Keywords: Alias detection, Time profile, Machine learning , Social media

Introduction

An increasing amount of many people's life is spent online. People are using Internet and social media in order to communicate, express their opinions and beliefs, discuss topics of interest to them, etc. While much of the information is expressed publicly, there is also more sensitive information available in web forums and other social media services that potentially could be harmful to the author if it became widely known who the physical person behind the user that is posting information is in reality. There are many examples from the Security Informatics research community related to the analysis of terrorist activities on the Web (see e.g., [1–3]), such as the spreading of extremism propaganda and discussions on how to make improvised explosive devices. In such settings, it can be of fundamental importance to intelligence analysts to find out what a person writes and who the physical person behind some pieces of texts really is. Similarly, it can be highly relevant for the police to find out who the author behind anonymous posts in a cybercrime investigation really is by linking the anonymized social media account to non-anonymized postings or accounts. This is the main motivation and driving factor for the research presented in this article. However, the Web is, fortunately, not only used for activities related to terrorism or crime. Ordinary citizens may also want to preserve their anonymity when discussing private issues such as religion, sexual preferences, political ideas, diseases, etc. in public. Obviously, what is considered as private and sensitive information varies from country to country and individual to individual. Many people would like to be able to freely express their ideas and beliefs, while at the same time avoid revealing their true identity to e.g., friends, employers, commercial companies, police, or intelligence services.

A quite common approach to reach some kind of anonymity when discussing sensitive issues online is to make use of "anonymous" or "alter ego" social media accounts when posting user generated content of private nature, while using other accounts for non-sensitive postings. A rather obvious problem with such an approach is that the

*Correspondence: fredrik.johansson@foi.se
†Fredrik Johansson, Lisa Kaati and Amendra Shrestha have contributed equally to this work
[1] Swedish Defence Research Agency (FOI), Stockholm, Sweden
Full list of author information is available at the end of the article

Internet service provider and social media service can log the IP-address and help identifying the user from this information, unless the providers can be fully trusted by the user or if extra counter-measures are applied, such as logging in from various Internet cafes or making use of tools such as Tor[1] [4]. There are, however, also less obvious problems with such an approach. It has for a long time been known that stylometric techniques can be used to identify an author among a small set of candidate authors given a large enough data material, but more recent research experiments presented in [5] suggest that this can be accomplished with reasonable accuracy also on large-scale datasets. A user who is aware of such techniques can in theory obfuscate their writing style intentionally, e.g., by using methods similar to those suggested in [6–8], but this is probably quite unusual.

In our previous work [9], we implemented a subset of the features suggested in [5] and used them for alias matching (i.e., the problem of identifying multiple aliases belonging to the same individual in an unsupervised fashion). In addition to the use of stylometric features we also used time as a feature to increase the possibility to detect users with multiple aliases. As the use of stylometric features sometimes is referred to as a user's "writeprint" in the sense that it can be compared to a fingerprint when it comes to identifying a person, we are here using the term "timeprint" when referring to how various time features can be used to identify a person. A timeprint can be seen as a property that reflect something about the characteristics of an individual's activity and her habits. In our previous work the timeprints were based on the publishing times when a user post messages, capturing the distribution of messages over the hour-of-the-day. By using timeprints in combination with stylometry, the detection rate of finding multiple aliases has been shown to increase significantly [9]. In this article, we are exploring various time features in more depth in order to increase the quality of timeprints and in order to find out how successfully they can be used for author identification and alias matching. One of the reasons why we are investigating to what extent timeprints can be used to identify a user is due to the fact that in some cases the publishing time is the only thing that is present, in particular in cases where users post images or videos with illegal content such as child pornography. Another reason is that a person's timeprint can be expected to be quite uncorrelated to her writeprint, making it possible to combine the two into a more powerful feature set.

We explore time features by explorative studies and experiments using the ICWSM forum dataset, containing data from an Irish forum site.[2] We show that a set of time features can be powerful for unmasking an author's identity in both a supervised (author identification) and an unsupervised (alias matching or similarity detection) setting. Additionally, we make experiments to find out which impact the amount of available posts and the way we split the posts into sub-users have on the resulting accuracy.

The rest of this article is structured as follows: in "Author identification and alias matching", we define the problems of author identification and alias matching, and present related work. In "Timeprints and activity profiles", we present various time features which potentially can be useful components of a timeprint, and show how the time features are varying among different users and over time for the Irish forum data. Moreover, we explain the concept of "circadian topology" or "chronotype" as a motivation for why people can be expected to have timeprints which are different from other individuals' timeprints. Next, we describe the experimental setup that has been used in our experiments on synthetically generated data in "Experimental setup". This section is followed up with the actual machine learning experiments, presented in "Experiments on author identification". In these experiments, we evaluate how well a classifier can learn to predict the correct author or user among a larger set of potential candidates by using time features. Hence, this is an example of how the classic problem of author identification can be tackled using non-textual features only. In "Experiment on alias matching", the usefulness of using timeprints for alias matching is evaluated and compared to stylometric-based features. In the case of alias matching we compare each user identity to all other identities and group together users (aliases) which are more similar than a certain threshold. In "Discussion", we briefly discuss under which circumstances the obtained results can be expected to hold in "the wild" and which implications our experimental results are likely to have. Finally, we present some conclusions and directions for future work in "Conclusions and future work".

Author identification and alias matching

Author identification, also known as authorship attribution, can be defined as the problem of assigning a text of unknown authorship to one candidate author, given a set of candidate authors for whom texts of undisputed authorship are available [10]. Authorship identification is a fairly well-studied problem, where algorithms and various features have been extensively described in, e.g., [5, 11–13]. However, existing approaches rely on linguistic/stylometric features (lexical, syntactic, idiosyncratic, etc.), while we in this article mainly study the usefulness of time features based on when texts have been written or published. To the best of our knowledge, time features

[1] https://www.torproject.org/.

[2] https://www.boards.ie/.

have not previously been used for author identification purposes except for in [14], which the work presented in this article is an extension of. Clearly, information about time is not always available, but when analyzing posts from social media (e.g., Twitter, web forums, etc.), such information can often be extracted.

In the author identification problem we compare each anonymous user to a fixed set of pre-defined known entities. In this way, we assume that the anonymous user is one of the exhaustive list of candidate authors present in the training set. This kind of problem setting can for example be of interest in a criminal investigation in which threatening messages have been received from a specific computer or IP-address to which only a limited number of suspects have access. Now, given that we can retreive the publishing times of the threatening messages, we can construct timeprints and writeprints from the threatening messages and compare them to timeprints and writeprints extracted from blog posts, tweets, forum posts, etc. written in user-generated content on social media using the suspects' known social media accounts.

In an alias matching setting (described in more detail in [9]), we cannot assume that we have knowledge of all potential authors. The (intra-platform) alias matching problem is instead to compare each anonymous identity to all other identities and group together users (aliases) which are more similar than a certain threshold. Hence, while author identification can be seen as a supervised machine learning problem, alias matching is an unsupervised problem where the same supervised algorithms cannot be used. Instead, we are for the alias matching problem making use of a vector space-representation for the various aliases, where each feature corresponds to a single dimension in the vector space. Next, we make use of a distance function (in our case cosine similarity, although other alternatives such as Manhattan distance could be used) to calculate the similarity among the various aliases. In addition to publishing time, i.e., timeprints, there are also other kinds of features which can be used for alias matching. One obvious candidate is the use of stylometric features. In this work we compare the usefulness of stylometric features and timeprints, as well as their combination.

Usernames is another feature that can be considered when linking multiple aliases to each other. In [15] an attempt to link user accounts across different social media services is presented. The authors show empirically that in almost 60 % of the cases, people use the same username in all social media platforms they are part of. A more sophisticated study is presented by Perito et al. [16]. They also consider usernames for linking user accounts but have instead developed a model for estimating that two usernames from two separate social media services belong to the same individual. In their approach a Markov chain model trained on approximately 10 million usernames gathered from Google and eBay is used to estimate a probability of how common tho usernames are. Although both these studies are very interesting, usernames would hardly be as useful for situations where people actively would try to hide who they are when making sensitive postings, which is the main reason for not including them in this study. In [17], a supervised method for finding mappings among identities of individuals across social media sites is introduced. While the methods suggested in [15, 16] make use of only one and two features, respectively, extracted from the usernames, the MOBIUS approach in [17] builds upon a large set of features extracted from the usernames (including the ones suggested in [15, 16]). Many classification techniques are evaluated in their experiments, but a logistic regression classifier is shown to perform best (with an accuracy of 93.8 %). This classifier performs much better than the methods suggested in [15, 16] as well as a number of baselines (such as exact username matching and substring matching). Overall, the most important features in MOBIUS are the standard deviations of the normalized edit distance and normalized longest common substring between the candidate username and prior usernames, and the username observation likelihood, but in total more than 400 features extracted from the candidate usernames, observed prior usernames, and relations between the candidate username and the prior usernames are used by the classifier. The results are impressive, but are not likely to work well in situations where people actively are trying to avoid their usernames being linked to their physical identity.

Another type of approach to detect "split identities" of web authors is presented in [18]. They argue that feature extraction and machine learning on Web scale is very costly and does not scale well, since pages or postings written by the same author can be similar in many different ways (demanding large feature sets). Instead, they suggest using (open source) compression software for extracting the compression distance for pairs of web pages, hypothesizing "that every author has a unique compression signature that is similar across all the pages of the same author". Technically speaking, they make use of a two-sided normalized compressor distance (NCD) which measures how much the compression of each of two pairs of web pages is improved by using information in the other web page. The resulting distances are then used to cluster web pages, where the aim is to group all pages written by the same author in a common cluster. As shown in our experiments, our method can be applied to datasets containing quite a large number of users

(although it obviously is a difference between a few thousand users and "Web scale").

Another approach to identify users with multiple aliases on social media services that has been suggested in existing literature is to make use of social network analysis (SNA). In [19], a social network based on aliases (email addresses collocated on the same web pages) is constructed. Depending on the number of web pages that the aliases are co-occurring on, the network is weighted and the the geodesic distance is computed. Social networks can potentially be used in other ways for linking multiple aliases as well, for example to construct a social network based on communication patterns or what topics an alias is writing about, as suggested in [9, 20].

In [21], the framework HYDRA is proposed for enabling large-scale social identity linkage across social media platforms. According to the authors, such linkage would allow for more complete and consistent user information when profiling users. HYDRA consists of two main components: one for measuring the heterogeneous behavior similarity between users and one for leveraging users' core social network structures. These components are combined using multiobjective optimization. In the behavior similarity calculation, many features are taken into account including a comparison of user profile images, various textual attributes from the users' profiles, and a rudimentary modeling of writing style (based on extraction of the most unique words of each user). The authors have been able to verify their suggested method on impressively large datasets (several million Chinese users with accounts on several social networks obtained from a third-party data provider). The results show that HYDRA outperforms the other approaches, which is unsurprisingly since the other approaches take fewer sets of features (mostly usernames) into account.

Although there are many different approaches to author identification and social media linkage suggested in existing research literature, there are to the best of our knowledge no previous attempts to make use of publishing times, except for our previous papers [9, 14], which this article is an extension of.

Timeprints and activity profiles

A chronotype or a circadian typology is an individual difference in personality, which is believed to be the cause of why some individuals prefer to work and exercise in the morning hours while others prefer evening hours. In 1976, Horne and Östberg published a 19-item morningness-eveningness questionnaire [22] that was used to measure if a person was a morning or evening person. The questionnaire has been validated in many countries with regards to local cultures and ages. Apart from the morningness-eveningness questionnaire, circadian typology has been described and measured using different questionnaires in many studies and several countries [23].

Circadian preferences are based on genetic influence. The circadian typology classifies individuals according to three different types: morning-type, evening-type, and neither-type. Most people are neither-types and are positioned somewhere between evening or morning types while half of the population are either morning-types or evening-types. It has been noted that people who have the same chronotype have similar activity pattern timing: they prefer to sleep, eat and exercise during more or less the same hours. The circadian typology seems to have an impact on the behavior of an individual and various studies have for example suggested that evening types spend more time in front of the screen [24] or that evening types have a higher tendency for cigarette craving and alcohol usage [25].

When a person is active or not on social media services may also be correlated to the person's chronotype. In a discussion forum such as boards.ie, it is possible that a morning-type individual post most of her messages during the morning hours, while an evening-type prefers posting messages during the evening. In addition to chronotypes, there are other aspects that affects when a user is active on social media. These factors are related to the living situation of a person, for example factors such as during what hours a person is at work, what time zone the person lives in, when a person has vacation and what kind of occupation the person has all matter when it comes to usage of social media. Altogether there are several distinguished features that are characteristic for a user's social media usage.

To obtain an understanding on how users behave, and investigate if there are some features that seem to be more characteristic than others when it comes to activity of a user, we studied the time-based activity patterns of a set of randomly selected users that were active in the discussion board. Figure 1 shows a randomly selected user's distribution of posts throughout a day. As can be noted in the figure, it seems to be the case that this user has a quite similar behavior of when he/she is active and not throughout all 5 years that are compared. When analyzing users' activity during various time intervals we noted that the activity pattern or time profile of a user seemed to be quite specific for each of the selected users. We also noted that the activity of a user seemed to be consistent over time (using data from different years). This might have to do with the fact most people are creature of habit and unconsciously prefer to do the same things during the same hours and periods. Some of the features that we considered in our manual analysis were:

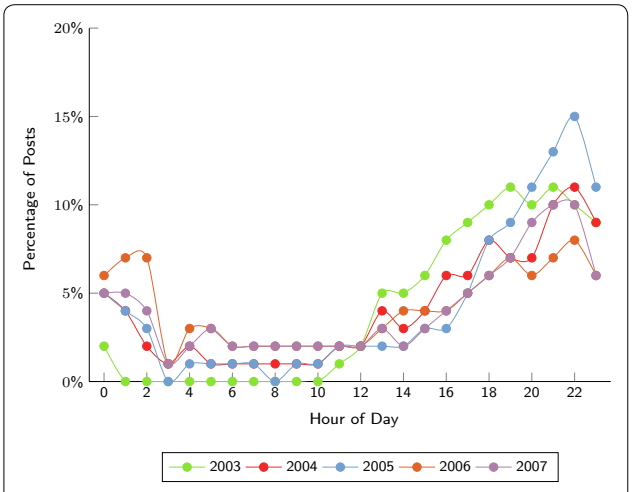

Fig. 1 Distribution of messages for a single user throughout the day during 5 years

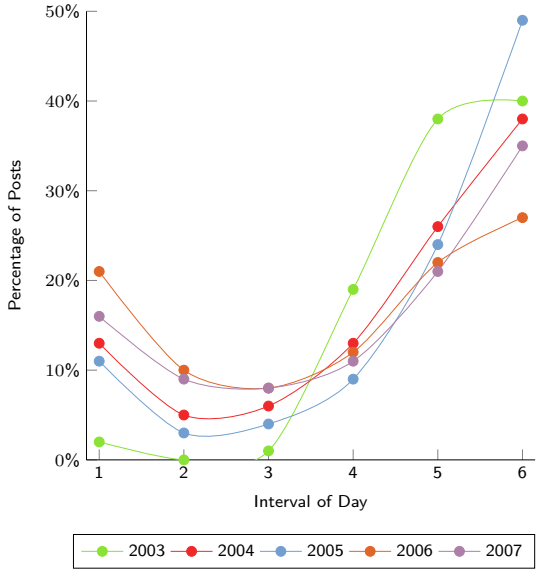

Fig. 2 Distribution of messages for a single user in 4 h intervals

- Activity during each month
- Activity during each hour of the day
- Activity during each day of week
- Activity during weekdays and weekends
- Activity during 4-h intervals (early morning, morning, midday, evening, night, midnight).

Figure 2 shows the distribution of messages (posts in the discussion board) in 4-h intervals for a single user in the boards.ie dataset. Each number on the x-axis represents a 4 h interval where 1, 00.00–03.59; 2, 04.00–07.59; and so on. The same user has been active during 2003–2007, and the distribution of messages is shown for each year. As can be noted in the figure, this user seems to have a consistent behavior when it comes to distribution of messages over the years. The fact that the behavior seems to be consistent over the years and that the pattern for each user seems to be somewhat unique is something that has been noted previously. How the behaviors and patterns can be used for author identification and alias matching is studied in more detail in later sections, but let us first see how such activity patterns have been utilized in related literature.

In [26], a temporal analysis of the blogosphere was done. The assumption was that each blogger has a different preference for posting. A dataset consisting of nearly 700,000 blog articles was analyzed according to two factors: (1) day of the week and (2) time of the day. One of the conclusions in the paper is that each blogger has a different temporal preference for posting which supports our thoughts that different discussion board users have different preferences for posting, and therefore will have timeprints that differ from each other. In [27], a double-chain hidden Markov model is used to characterize individuals' behavior in e-mail communication. The results show that users fall into two well-separated clusters: "day laborers" and "e-mailaholics". Given these clusters it is shown that a vast majority of the users retain their routines over an extended period of time.

The above studies indicate that timeprints may be useful for separating users to some degree, but the experiments presented in the following sections investigate this possibility much more thoroughly.

Experimental setup

We have conducted a number of different experiment to understand how well timeprints perform on the two problems of author identification and alias matching. In our experiments we have made use of the following sets of features when constructing our timeprints:

- *Hour of Day* Hour1, Hour2, ..., Hour24
- *Period of Day* MidNight, EarlyMorning, Morning, MidDay, Evening, Night
- *Month* Jan, Feb, ..., Dec
- *Day* Sunday, Monday, ..., Saturday
- *Type of Day* WeekDay, WeekEnd.

In the construction phase we go through each post made by an alias/user. We count the number of occurrences of each attribute and then express the values as relative frequencies, so that the values of each set of attributes sums to 1 (e.g., $WeekDay = 0.65$ and

WeekEnd = 0.35). The relative frequencies obtained from individual posts are then averaged into a single timeprint for each alias.

In some of our experiments we also combine our timeprint with stylometric features which are calculated in the same way. The stylometic features that we use are summarized in Table 1 and consist of various features which are supposed to be able to help distinguishing between various users without being sensitive to the actual topic of the text. We have in our stylometric implementation included a subset of the features used in the article by Narayanan et al. [5] plus an extra feature: the relative frequency of various smileys. As for the timeprints, the values for each set of stylometric features sum to 1. A lot of other features could have been used, including lexical features such as vocabulary richness [e.g., using frequency of hapax legomena (once-occuring words) or Yule's K measure], syntactic features such as part-of-speech tag n-grams, and idiosyncratic features such as misspelled words. We are not arguing that we have used the richest set of features possible, but rather that we have incorporated a lot of useful features that reasonably fast can be extracted from forum posts. The present features can be extended in the future to allow for even better stylometric "writeprints".

While we ultimately would have preferred to run our experiments on real-world data in which a subset of the users were known to make use of multiple aliases, such data is very hard to get hold of for research purposes and also raise ethical concerns. Instead, we have synthetically created data based on posts extracted from the ICWSM boards.ie forum dataset. First of all, we have identified and extracted the posts for the 4000 users who have posted most posts in the forum during year 2007. The reason for choosing those users is that we wanted to have as large data material as possible, since a reasonable assumption is that the amount of data will have an impact on the achieved results. This assumption is tested further in one of our experiments described below.

Table 1 The stylometric features

Category	Description	Count
Word length	Relative frequency of words with 1–20 characters	20
Letters	Relative frequency of *a–z* (ignoring case)	26
Digits	Relative frequency of 0–9	10
Punctuation	Relative frequency of characters . ? ! , ; : ( ) " - ′	11
Function words	Relative frequency of various function words	293
Smileys	Relative frequency of various smileys :') :-) ;-) :P :D :X <3 :) ;) :@ :* :| :$ %)	14

In our alias matching experiment we have from the set of top-posters first selected a smaller set of users ($n = 500$) (where the selection is based on the descending order of the users' amount of posts). Each of these users have been split into two separate users u_{ia} and u_{ib}, where $1 \leq i \leq 500$ and posts are assigned randomly among user u_{ia} and user u_{ib}. The intention of this split is to simulate a user who make use of two separate aliases, without assuming too much about the patterns in which the user will switch among the two aliases. Now, each user in the set $\{u_{1a}, u_{2a}, u_{3a}, \ldots, u_{na}\}$ is compared, one at a time, with all the users in the set $B = \{u_{1b}, u_{2b}, u_{3b}, \ldots, u_{nb}\}$. Based on the results from the time-based matching we rank the members of set B according to how similar they are to the selected user (where the similarity among two vectors is calculated using cosine distance). The most similar member of the set B is ranked as number one, the next most similar as number two and so on. The reported accuracy is calculated as the fraction of times the index of the selected alias is found within the top-N rankings (where the results for $N = 1$ and $N = 3$ are reported). This kind of experiment has then been conducted for increasing values of the number of users n, where we have varied n from 500 to 4000 in steps of 500.

In the author identification experiments, we have only made use of the top-1000 users, since a larger set of potential classes would have been hard to cope with for the more complex classifiers that have been used. Each user u_i has been split into five "sub-users" $u_{i1}, u_{i2}, \ldots, u_{i5}$. The posts are in most of our experiments divided randomly among the sub-users, but we have also made a separate experiment on how the results differ if the posts are divided sequentially rather than randomly (i.e., where the user's first post has been assigned to u_{i1}, the second post to u_{i2}, etc.). The reason for using five sub-users is that we in this way construct several training instances for each user in order to facilitate the learning phase in our supervised learning experiments. Based on the extracted posts, timeprint vectors have been constructed (one for each sub-user). In the author identification experiments we incorporate the UserID u_i as the target class. Hence, we have five (different) data instances for each UserID. In our supervised learning experiments we compare the accuracy for two popular supervised learning algorithms: a Naive Bayes (NB) classifier and a support vector machine (SVM) classifier [28]. For the author identification experiments we have made use of the Waikato Environment for Knowledge Analysis (WEKA) [29]. For the SVM classifier, we have used the nu-SVC classifier from the libsvm package in WEKA. We have used a linear kernel with default parameter settings since this was shown to give better results than a radial basis function in our initial experiments. In each step we have

performed tenfold cross validation and the results from the ten folds have been averaged into a single accuracy value which is reported.

Experiments

As explained in the previous section, we have conducted a number of different experiments to understand to what extent timeprints can be used for author identification and for alias matching. In this section we describe the individual experiments in further detail and present the obtained results.

Experiments on author identification

In our first experiment, we have varied the number of potential authors from 200 to 1000 in steps of 200 as explained above. First, we have randomly divided a user's posts among its five sub-users. After this, we have used tenfold cross validation in which a Naive Bayes and a support vector machine have been trained and evaluated. We have then compared the obtained accuracies for the Naive Bayes and support vector machine classifiers. The results from the experiment are reported in Fig. 3 using the labels *NBRand* and *SVMRand*. Next, we have repeated the same experiment again, but this time we have distributed the posts sequentially among the five sub-users, rather than to distribute them randomly. These results are reported in Fig. 3 using the labels *NBSeq* and *SVMSeq*. Now, what is the reason for studying the effect of the way the sub-users are selected on the obtained results? The explanation to this is that it is hard to know exactly how someone would make use of several aliases.

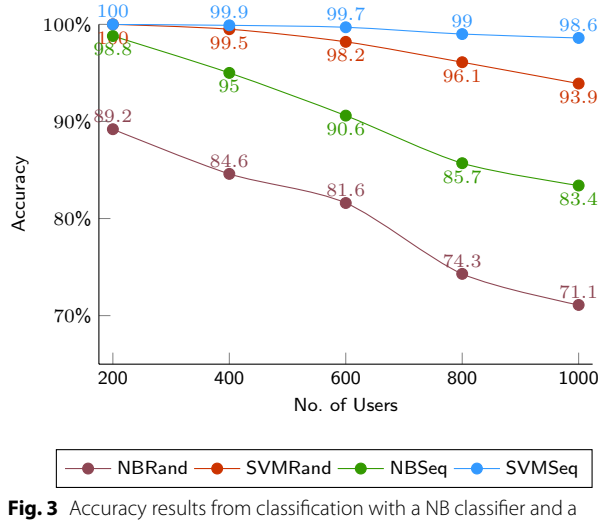

Fig. 3 Accuracy results from classification with a NB classifier and a SVM classifier. Experiments conducted on the 1000 users that made most posts in year 2007. Timeprints have been created by random and sequential division of users into sub-users

Would they first make a post using one of their aliases, then switch to a second, and so on? Would they first write a large number of posts using one account, then switch to the next, and so on? The answer is probably neither, and the exact behaviour is probably very different for different individuals and different purposes for why multiple aliases are used in the first place. For this reason, the sequential and random splits of posts are intended to be seen as two extreme points, where the sequential way of dividing the posts into sub-users is intended to give an upper bound on the classification accuracy that can be expected when using timeprints, while the random way of dividing posts is more challenging for the classifier and can be seen as a lower bound on what classification accuracy that can be obtained. In reality, we would expect the accuracy to be somewhere between these upper and lower bounds, depending on the individual and what she is using the multiple aliases for.

As can be seen when studying the results in Fig. 3, both classifiers perform well on the classification task, especially the SVM classifier. For 200 users the SVM classifier is achieving 100 % accuracy on the sequentially generated data and the NB classifier is not far away from this result either. On larger problem instances, the SVM classifier is consistently outperforming the NB classifier with approximately 5–20% higher accuracy, but this comes with a price. The training and evaluation phase of the NB classifier took a few minutes while the last steps took days to perform for the SVM classifier on the standard computer we used for the experiments, due to the large number of classes.

As expected, higher accuracies are achieved when distributing posts sequentially rather than randomly among sub-users. This is expected since features such as *Month* will have very similar relative frequencies among sub-users corresponding to the same user when dividing the posts sequentially. In a setting where people make use of several accounts sequentially (such as when having a "discussion" between two or more alter egos) the sequential approach make sense, while it probably is less realistic for more normal use of multiple aliases. For this reason we are in the rest of the experiments distributing posts among the sub-users in a random fashion, although this probably is closer to a lower bound on the accuracy.

In this experiment, the correct user is in more than nine out of ten cases selected when using the SVM, both when using the random and sequential approach to creating the timeprints. As can be seen, the accuracy is still over 90 % for the SVM classifier when increasing the number of users to 1000. The results are somewhat worse for the NB classifier, but are still impressive given the simple nature of the classifier. The achieved results imply that time features can be very useful for author

identification when having access to large amounts of data material. Those results are significantly higher than those obtained for author identification with textual (stylometric) features on a forum dataset reported in [11]. It should however be noted that it is not the same forum datasets that have been used in those experiments, making it hard to make a fair comparison between the results.

In order to find out which of the time-based features that are most important for the achieved classification performance, we have applied information gain, which is an entropy-based feature selection method [30]. The results vary somewhat depending on which number of users the measure is applied to, but in general we can see that attributes related to *Period of Day* (such as Night and Morning) receive highest average ranks, followed by the *Type of Day* (i.e., weekend or weekday). After this follows *Months* and *Days*, while the set of attributes which seem to be least useful is the *Hour of Day*.

An important part of the explanation to the decrease in accuracy when the number of potential authors is increased is obviously that there are more candidates to chose among for the classifiers, but a contributing factor may also be that there is less data material for the users further down in the list (since they are ordered based on their number of posts). To get a better understanding of what impact the amount of posts has on the results, we have in a second experiment modified the original dataset so that we start out with randomly selecting only 100 posts for each of the top-200 users. Since each user is decomposed into five sub-users, this means that the timeprint vectors are built from 20 posts each only. The number of users is kept fixed to 200 while we are increasing the number of randomly selected posts in steps of 100, until reaching 1400 posts (since the 200th user has written a total of 1484 posts). In effect, this experiment simulates a setting where only a restricted amount of data material is available for creating the timeprints. When adjusting the experiment in this manner, the results shown in Fig. 4 were obtained.

As can be seen in this figure the accuracy increases for both classifiers when the numbers of posts are increased and the SVM classifier consistently performs better than the NB classifier. When only 100 posts are randomly selected for 200 users, the accuracy is below 20 % for both classifiers. When the number of posts are increased to 500, the SVM classifier has an accuracy over 80 % while the NB classifier is just above 55 %. The SVM classifier reaches over 95 % accuracy when the number of posts is around 800 and slowly continues to increase. The figure gives a good idea of what amount of data that is needed to reach a specific accuracy level for the classifiers, but please note that the although the shape of the curves is similar also for other number of users, the exact values will differ.

Experiment on alias matching

In our alias matching experiment, we have selected the top-4000 users from year 2007 and randomly splitted a user's post into two equally sized users as explained in "Experimental setup". Now, we have for each sub-user created three sets of feature vectors from its posts: (1) a timeprint vector, (2) a stylometric-based vector, and (3)

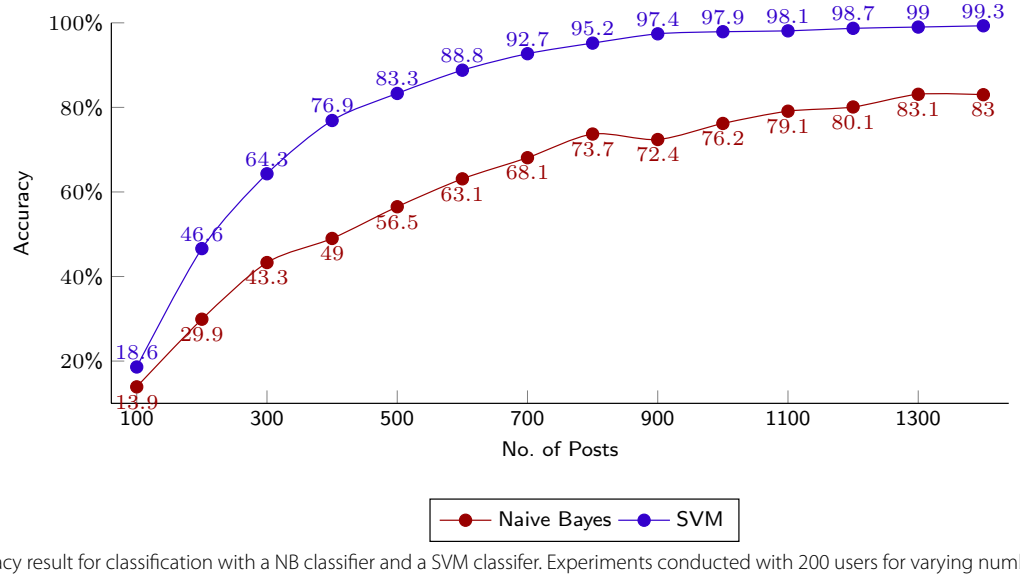

Fig. 4 Accuracy result for classification with a NB classifier and a SVM classifer. Experiments conducted with 200 users for varying number of posts on data from year 2007. The posts have been selected randomly

a combination of the timeprint feature vector and the stylometric-based feature vector. The results from this experiment are summarized in Fig. 5. As can be seen, both time-based and stylometric-based features work very well for a limited number of users, but as the number of users is increased the decrease in performance is less steep for the timeprints compared to the stylometric-based features. Looking at the top-1 rankings (with top-3 ranking within parantheses), time alone yields 100 % (100 %) for 500 users, 89.6 % (94.4 %) for 2000 users, and 70.6 % (79.0 %) for 4000 users. The corresponding results for stylometric features are 98.2 % (99.2 %) for 500 users, 71.4 % (78.5 %) for 2000 users, and 44.6 % (51.7 %) for 4000 users. To combine both time-based and stylometric-based features seem to give better performance than using the individual time-based or stylometric-based features on their own, and this effect seems to increase as the number of users is increased. More specifically, the corresponding results for the combination of time-based and stylometric-based features yield 100 % (100 %) for 500 users, 93.0 % (96.6 %) for 2000 users, and 75.4 % (82.8 %) for 4000 users. Based on these results we can conclude that time-based features are vey powerful on their own for alias matching and that combining time-based features with stylometric-based features allow for even (statistically significantly) better results.

Discussion

The experimental results presented in the previous section indicate that timeprints can be very useful for both author identification and alias matching. However, it should be noted that the results have been obtained in quite well-controlled experimental settings which does not necessarily hold true in a real-world environment.

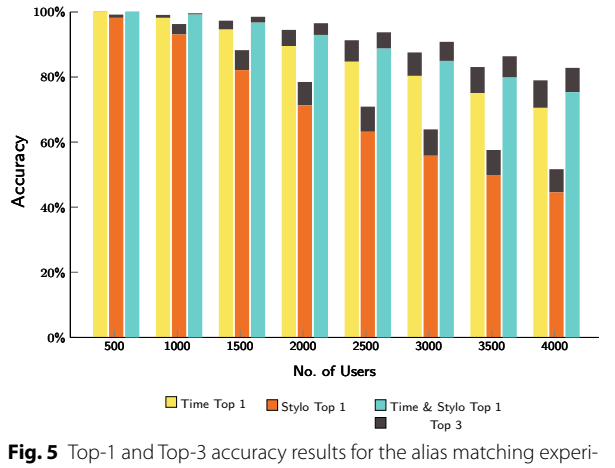

Fig. 5 Top-1 and Top-3 accuracy results for the alias matching experiment involving time-based features, stylometric-based features, and a combination of both

As have been shown in our experiment where we have used various ways to split users into sub-users (using random and sequential splits), the effect on the results can be quite large. The effect of splitting users into sub-users has been discussed before in existing literature on stylometric-based author identification [31, 32] but we are not aware of any previous studies on the effect on time-based author identification or alias matching.

In our alias matching experiments, we have been able to control so that posts have been evenly distributed among sets *A* and *B*. In a realistic setting, posts from two aliases belonging to the same individual could potentially be more unevenly distributed, so that the posts for one alias could have been created during a completely other time period than the posts for the first alias. If this would be the case, this would have a negative impact on the obtained accuracy. Similarly, it can be expected that lower accuracy would be obtained for the author identification problem if the classifiers were applied on timeprints based on posts from a different time period than they were trained on. Furthermore, it is expected that timeprints from the same individual obtained from various social media platforms would look quite different from each other, even though they most likely would be correlated. For these reasons, the obvious next step is to attempt to evaluate the suggested algorithms on real-world data obtained from various social media platforms. Although it is not straightforward how to do this, there are various ways in which evaluation on real datasets could be made. One such way would be to make intra-platform experiments in which we attempt to find user accounts which are very similar to each other in terms of stylometric and time-based profiles and manually assess the matches to be either true or false positives, i.e., similar to the approach used in [33]. A drawback with such an evaluation is that only precision can be calculated, not recall. Another approach would be to try to get hold of a ground truth dataset based on uniquely identifying information such as e-mail addresses from a number of user accounts and use this as ground truth for whether two accounts should be matched. This would, however, only work for linking aliases across different social media services and would not be as suitable for evaluating how well we can detect users who actively try to remain anonymous.

In addition to the question of in which real-world settings the suggested approach is expected to work well from a technical perspective, a very related question is in which situations it would be ethically acceptable to make use of alias matching and author identification techniques? We have no definitive answer to this, but as future (and already ongoing) work we intend to discuss this issue with a combination of law enforcement

agencies, ethical philosophers, and IT law scholars. From our point of view, it would make sense from an ethical perspective to use this kind of techniques for author recognition purposes in a cyber crime investigation involving a few suspects, while it would be much more questionable to use it on large scale in an inter-platform setting which would allow for general dredging of possible matches. With this said, development of better and more exact guidelines for when this kind of techniques can and should be applied are left as future work.

For the author identification problem, we have split the available posts for a user into five separate training samples. This has proved to work quite well, but the number of training samples per user has been quite arbitrarily selected. The optimal value of training samples is probably dependent upon the number of potential authors as well as how much posts we have available for each user, but finding such an optimal value has been outside the scope of this article. However, as a rule of thumb, the more posts we have for a certain user, the more high-quality training samples we can create.

One positive interpretation of results that have been obtained (if transferable to the real-world) is that police and intelligence services around the world can become more effective in finding the author of large quantities of terrorist propaganda and other crime- or terrorism-related content. A more negative interpretation is that the online anonymity of ordinary citizens in worst case may be weakened if this kind of techniques would be used by, e.g., commercial companies or repressive regimes. This raises the question of whether the use of time-based features can be defended against by an individual who wants to preserve his or her anonymity. A potential solution could be to use software which does not publish posts directly as they are written, but rather delay the creation time of new posts randomly. However, it is unlikely that many people would make use of such techniques, which in practice will add the use of timeprints (combined with stylometric features) as a potential attack vector on online anonymity.

Conclusions and future work

In this article, we have presented the idea that a user's timeprint (which can be extracted from the publishing times of a large number of social media services) can be useful for identifying users who make use of multiple aliases. This idea has been motivated by arguments such as the existence of individual differences in personality preferences related to time (morning-type, evening-type, neither-type) and the fact that people have different working hours and sleeping hours. By selecting a few users and looking at their behavior over time we have noted that many users seem to have a quite stable activity behavior over time. Our initial manual analysis has indicated that there might be a possibility to tell individuals apart based on their timeprints. However, by just looking at a set of users' behavior over time we can not say much about how unique a timeprint is.

To get a better understanding of the uniqueness of individuals' timeprints, we have made supervised machine learning experiments where we have attempted to learn classifiers to tell users apart based on various time features. This can be thought of as author identification based on activity rather than textual style. The results suggest that high accuracy can be obtained also for large number of potential authors (over 90 % up to 1000 users), but that the accuracy is highly dependent upon the number of posts from which the timeprints are created. In a second set of experiments, we have tested the usefulness of time features for the unsupervised alias matching problem. We show that good performance can be achieved and that even better results are achieved when combining the time-based features with stylometric-based features.

The results in the article are encouraging from an intelligence and security perspective, but they might pose a threat towards privacy and online anonymity. If this kind of techniques can be used to reveal the true identity of a potential terrorist, there is a risk that the same techniques can be used also for other purposes, even though the usefulness of the technique decline as the number of users is increased. One way to defend against the use of "timeprint attacks" could be to use tools that automate the process of publishing. A more drastic defense could be that some individuals choose to stop posting sensitive information at all, but this would obviously have potentially severe consequences for democracy and individuals' right to freedom.

Future work

In this article we have only considered users in a discussion forum, but it is likely that the results can be transferred to other social media services as well. As future work we plan to test the usefulness of the developed timeprints on other social media services such as Twitter. We also aim at cross-platform experiments, in which correlations among discussion forums and other social media services can be explored. Moreoever, we would like to carry out large-scale experiments like those in [5], where the full set of their stylometric features are combined with the timeprint features developed in this article.

Another direction for future work is to move on and make real-world experiments using the proposed algorithms. Before this is done, it is however important to find out during which circumstances it is ethically

acceptable to use alias matching and author identification on real data, both for research purposes and for use by police and security agencies. For this reason, we are currently undertaking a project in which law enforcement agencies, ethical philosophers, and IT law scholars are involved. The results from this work will influence how and if the proposed methods will be evaluated on real-world data.

Authors' contributions
All authors have contributed equally much to the work reported in this article. All authors read and approved the final manuscript.

Author details
[1] Swedish Defence Research Agency (FOI), Stockholm, Sweden. [2] Uppsala University, Uppsala, Sweden.

Acknowledgements
This research was financially supported by Security Link and the Swedish Armed Forces Research and Development Programme.

Compliance with ethical guidelines

Competing interests
The authors declare that they have no competing interests.

Received: 16 March 2015 Accepted: 15 September 2015
Published online: 24 September 2015

References

1. Abbasi A, Chen H (2007) Affect intensity analysis of dark web forums. In: Proceedings of the 5th IEEE international conference on intelligence and security informatics
2. Brynielsson J, Horndahl A, Johansson F, Kaati L, Måartenson C, Svenson P (2012) Analysis of weak signals for detecting lone wolf terrorists. In: Proceedings of the 2012 European intelligence and security informatics conference, pp 197–204
3. Brynielsson J, Horndahl A, Johansson F, Kaati L, Måartenson C, Svenson P (2013) Harvesting and analysis of weak signals for detecting lone wolf terrorists. Secur Inform 2(11):1–15
4. Goldschlag D, Reed M, Syverson P (1999) Onion routing. Commun ACM 42(2):39–41
5. Narayanan A, Paskov H, Gong NZ, Bethencourt J, Stefanov E, Shin ECR, Song D (2012) On the feasibility of internet-scale author identification. In: 2012 IEEE symposium on security and privacy (SP), pp 300–314
6. Brennan M, Afroz S, Greenstadt R (2012) Adversarial stylometry: circumventing authorship recognition to preserve privacy and anonymity. ACM Trans Inf Syst Secur 15(3):12:1–12:22
7. Kacmarcik G, Gamon M (2006) Obfuscating document stylometry to preserve author anonymity. In: Proceedings of the 2006 COLING/ACL
8. Almishari M, Oguz E, Tsudik G (2014) Fighting authorship linkability with crowdsourcing. In: Proceedings of the second ACM conference on online social networks (COSN'14). ACM, New York, pp 69–82. doi:10.1145/2660460.2660486
9. Johansson F, Kaati L, Shrestha A (2013) Detecting multiple aliases in social media. In: Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM'13), pp 1004–1011
10. Stamatatos E (2009) A survey of modern authorship attribution methods. J Am Soc Inf Sci Technol 60(3):538–556
11. Abbasi A, Chen H (2008) Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans Inf Syst 26(2):7–1729
12. Zheng R, Li J, Chen H, Huang Z (2006) A framework for authorship identification of online messages: writing-style features and classification techniques. J Am Soc Inf Sci Technol 57(3):378–393
13. Juola P (2006) Authorship attribution. Found Trends Inf Retr 1(3):233–334
14. Johansson F, Kaati L, Shrestha A (2014) Time profiles for identifying users in online environments. In: Proceedings of the 2014 IEEE joint intelligence and security informatics conference (JISIC'14), pp 83–90
15. Zafarani R, Liu H (2009) Connecting corresponding identities across communities. In: Adar E, Hurst M, Finin T, Glance NS, Nicolov N, Tseng BL (eds) Proceedings of the third international conference on weblogs and social media
16. Perito D, Castelluccia C, Kaafar M, Manils P (2011) How unique and traceable are usernames? In: Fischer-Hübner S, Hopper N (eds) Privacy Enhancing Technologies. Lecture Notes in Computer Science, vol 6794. Springer, Berlin, Heidelberg, pp 1–17 (2011)
17. Zafarani R, Liu H (2013) Connecting users across social media sites: a behavioral-modeling approach. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, pp 41–49
18. Amitay E, Yogev S, Yom-Tov E (2007) Serial sharers: detecting split identities of web authors. In: ACM SIGIR 2007 workshop on plagiarism analysis, authorship identification, and near-duplicate detection. ACM, New York
19. Hölzer R, Malin B, Sweeney L (2005) Email alias detection using social network analysis. In: Proceedings of the 3rd international workshop on link discovery (LinkKDD'05). ACM, New York, pp 52–57
20. Dahlin J, Johansson F, Kaati L, Måartenson C, Svenson P (2012) Combining entity matching techniques for detecting extremist behavior on discussion boards. In: Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM'12), pp 850–857
21. Liu S, Wang S, Zhu F, Zhang J, Krishnan R (2014) Hydra: large-scale social identity linkage via heterogeneous behavior modeling. In: Proceedings of the 2014 ACM SIGMOD international conference on management of data (SIGMOD'14), pp 51–62
22. Horne JA, Östberg O (1976) A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. Int J Chronobiol 4(2):97–110
23. Adan A, Archer SN, Hidalgo MP, Di Milia L, Natale V, Randler C (2012) Circadian typology: a comprehensive review. Chronobiol Int 29(9):1153–1175
24. Urban R, Magyarodi T, Riga A (2011) Morningness-eveningness, chronotypes and health-impairing behaviors in adolescents. Chronobiol Int 28:238–247
25. Adan A (1994) Chronotype and personality factors in the daily consumption of alcohol and psychostimulants. Addiction 89:455–462
26. Lee B (2012) A temporal analysis of posting behavior in social media streams. In: International AAAI conference on weblogs and social media
27. Malmgren RD, Hofman JM, Amaral LAN, Watts DJ (2009) Characterizing individual communication patterns. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, pp 607–616
28. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
29. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. SIGKDD Explor Newsl 11(1):10–18. doi:10.1145/1656274.1656278
30. Lee C, Lee GG (2006) Information gain and divergence-based feature selection for machine learning-based text categorization. Form Methods Inf Retr 42:155–165
31. Layton R, Watters P, Dazeley R (2010) Authorship attribution for twitter in 140 characters or less. In: Proceedings of the second cybercrime and trustworthy computing workshop
32. Zechner N (2014) Effects of division of data in author identification. In: Proceedings of the fifth Swedish language technology conference
33. Novak J, Raghavan P, Tomkins A (2004) Anti-aliasing on the web. In: Proceedings of the 13th international conference on world wide web. ACM, New York, pp 30–39