Detecting Multiple Aliases in Social Media

Fredrik Johansson
Swedish Defence Research Agency (FOI)
Stockholm, Sweden
Email: fredrik.johansson@foi.se

Lisa Kaati
Uppsala University
Uppsala, Sweden
Email: lisa.kaati@it.uu.se

Amendra Shrestha
Uppsala University
Uppsala, Sweden
Email: amendra.shrestha@it.uu.se

***Abstract*—Monitoring and analysis of web forums is becoming important for intelligence analysts around the globe since terrorists and extremists are using forums for spreading propaganda and communicating with each other. Various tools for analyzing the content of forum postings and identifying aliases that need further inspection by analysts have been proposed throughout literature, but a problem related to this is that individuals can make use of several aliases. In this paper we propose a number of matching techniques for detecting forum users who make use of multiple aliases. By combining different techniques such as time profiling and stylometric analysis of messages the accuracy of recognizing users with multiple aliases increases, as shown in experiments conducted on the ICWSM dataset boards.ie.**

***Index Terms*—alias matching, multiple aliases, time profiling**

I. Introduction

Internet is a platform for individuals who want to express and share ideas and personal judgments relating to any subject matter. The downside is that many extremist groups and terrorists are using the Internet as a vital motivator for spreading their ideology and for exchanging and reinforcing their beliefs [1], which increases the risk of individuals committing violent acts against the society. A couple of years ago extremist groups mainly used printed magazines and centralized websites for spreading their views and information, but this has to a large degree been replaced by interactive discussion forums (such as the Ansar Al-Mujahidin Jihadist Forum) and social media platforms such as YouTube, Twitter and Facebook [2]. Monitoring and analysis of web forums (also called discussion boards, discussion forums, message boards, Internet forums, etc.) is therefore becoming an important task for analysts in order to detect individuals that might pose a threat towards society.

Terrorist activities on the Internet can potentially be detected by monitoring the traffic to websites and forums associated with terrorist organizations under surveillance. In this manner, users accessing these websites can be identified based on their unique IP addresses. Unfortunately, it is difficult to monitor those sites since they do not use single fixed IP addresses and URLs [3]. In addition, these sites frequently change their geographical location of Web hosting servers in order to prevent the use of such techniques. Similarly, anonymization techniques like Onion Routing [4] and Crowds [5] have made it easier for users to hide their identity and activity on the Internet.

Because of the difficulties in identifying individuals based on the network traffic (and since such monitoring raises privacy concerns since users may access such sites for legitimate reasons) it makes sense to instead analyze the content of postings on extremist forums for identifying individuals who are worth investigating more closely by other means. This is an approach previously suggested in e.g., [6], [7], [8], [9]. A problem with such a content analysis is that it is not unusual that individuals make use of several aliases on a single web forum or on different social media sites, making it harder to make correct assessments. As an example, the Norwegian right-wing extremist and lone-wolf terrorist Anders Behring Breivik made use of several aliases on various social media sites before his attacks in Norway 2011 [10]. The use of several aliases can be perfectly normal, but can become a problematic issue when utilizing content-based analysis. To overcome this problem, we propose a number of matching techniques that can be used to identify users with multiple aliases. The obtained experimental results suggest that the combination of matching techniques can give significantly better results than if the techniques are applied individually. We also show that the achieved accuracy is largely dependent upon the number of aliases under consideration.

The rest of this paper is structured as follows. In Section II we present related work. In Section III we present various cases in which people may use several aliases and suggest techniques for how the use of multiple aliases can be detected in those cases. The suggested matching techniques have been implemented into a testbed which is used for the experiments with the ICWSM dataset boards.ie presented in Section IV. A discussion is provided in Section V and the article is concluded in Section VI, together with thoughts for future work.

II. Related work

In [11], the problem of ”anti-aliasing” is studied, i.e. to link multiple aliases to known individuals based on their postings in public fora such as bulletin boards, weblogs, and web pages. More specifically, the technique used for matching the aliases is based on the used vocabulary (i.e., which words that are used in the postings). The results are promising, but the similarity of aliases relies heavily on the topic that they write about since the users vocabulary is used as discriminating features. For this reason the method is not as suitable for people writing about heterogeneous topics.

Writing style is also used in [12], but in this work stylometric features which are not topic dependent are used. Hence, the suggested method is more suitable when dealing

with multiple topics than the approach suggested in [11]. By using such stylometric features (e.g., function words and the use of syntactic category pairs) Internet-scale authorship identification is described. The experiments were made on a large collection of blog posts written by 100,000 different authors. By using a small sample of blog posts they tried to identify the rest of the posts written by the same author, mixed in with the 100,000 other blog posts. Their algorithms ranked the possible authors in descending order of probability and the top guess was correct about 20% of the time. In 35% of the cases, the correct author was in the top 20 guesses. The precision was improved to 80% by lowering the recall to 50%. The results in [12] indicate that the method is scalable and applicable to large amount of data from e.g the Internet. We have in this article used many of the stylometric features suggested in [12], but have also used other classifiers than stylometric maching.

In [13], methods for combining output from several matching techniques such as field matching, graph matching, and text-based matching are described. The combination of these methods is supposed to improve identification of multiple aliases. However, no experimental results are shown. The implementations and experiments presented in this paper can be seen as a continuation and validation of the framework we previously have suggested in [13].

III. Detecting the use of multiple aliases

When trying to detect individuals who are using multiple aliases on web forums, there are several kinds of features and techniques that may be considered. To be clear on the terminology used, we will in the following use the terms (matching) *techniques* or *classifiers* when referring to the algorithms used for identifying multiple aliases, while the term *features* will be used for the more low-level attributes which are used within the classifiers. To exemplify, we make use of a stylometric matching technique as one of several classifiers when matching aliases, and this technique relies on several features such as relative frequencies of function words and word length distributions. Another technique or classifier used is time profile matching, where the time profile can be represented as a feature vector consisting of the user's relative posting frequency distributed over a certain period of time.

When comparing various web forums it becomes obvious that their structure varies. Some forums contain additional information that can be used as features to build classifiers but which are not present in other web forums. In this work we focus on general information which can be extracted from most web forums. The techniques we have implemented are described in Section III-A, but first we present several reasons for why individuals may use several aliases. Based on these reasons we have identified two main cases that are particularly interesting for alias matching. Some of the techniques that we suggest are only suitable for one of the identified cases, while others can be applied in both cases. Below follows a list of potential reasons for using multiple aliases. This list is a mix of reasons identified in existing literature ([14], [15], [11]):

1) the old alias has been deleted due to inactivity
2) the old password has been forgotten
3) the old alias has been banned by a moderator
4) the old alias has lost the trust of other members
5) bad relationships have been developed with other members
6) the user wants an extra alias to support own arguments, cause debate or controversy (the extra alias used for purposes of deception is sometimes referred to as a sockpuppet)
7) the user wants to discuss immoral or illegal activities
8) the user wants anonymity due to privacy reasons

Based on these reasons, we have identified two main cases when alias matching may be useful. The first case is when the user creates a new alias without any attempt to disguise the fact that multiple aliases have been created by the same individual. This is the case for reasons $1-2$ and potentially also for 3 (depending on whether the user fears to be banned again or not if the moderators find out that a new account has been created). The second case is when the user does not want to reveal that several aliases belong to the same individual. This is in general true for reasons $4-8$ and also often for 3. The amount of effort put in for hiding that several aliases belong to the same individual will obviously vary, but it can be expected that the user will not use a very similar alias name (username) in the second case. In the following, we will therefore refer to these two major cases as the "non-concealed" case and the "concealed" or "alter ego" case respectively, although the level of concealment can vary within the second category.

A. Techniques for detecting multiple aliases

Most existing work on alias matching focus on techniques for finding similarities in usernames (see e.g., [16]). Such techniques may work well for non-concealed cases (e.g., when the user is using similar usernames on several social media services and is not deliberately trying to hide that several aliases belong to the same individual). However, an individual may choose very dissimilar aliases (deliberately or not) and for such cases techniques that simply rely on similarities in usernames will not be fruitful. Moreover, two usernames may be very similar without belonging to the same individual. Hence, string matching techniques will not always be enough. For those reasons, we have in addition to string matching techniques implemented several types of other techniques for alias matching and propose to combine the results in order to come up with better alias matching possibilities.

In the following, we outline a method for discovering multiple aliases created by a single author/individual by studying a number of classifiers. The classifiers we consider are:

- **String-based matching** (for matching based on alias names)
- **Stylometric matching** (for matching based on the written posts)
- **Time profile-based matching** (for matching based on the publishing time of the posts)

- **Social network-based matching** (for matching based on thread or friend information)

1) String-based matching: Aliases usually consist of text strings. As has been discussed above, the similarity of two aliases can be a useful feature to consider when trying to find users making use of multiple aliases, at least for the non-concealed case. Various edit distance measures have been proposed throughout literature, including Levenshtein distance [17] and Jaro-Winkler distance [18]. We have implemented the Jaro-Winkler distance measure since it has been specially designed to weight the first letters in a name higher than the ending of names, which makes sense for finding similarities in aliases such as JakeJ and JakeJ_3. More elaborate and non-standard methods such as the Markov Chain-based approach suggested in [19] can be of interest for more complex alias matching implementations, but will not be discussed further in this paper.

Basically, what the Jaro-Winkler distance does is that it returns a normalized score in the unit interval $[0, 1]$, where 0 means that there is no similarity at all among the strings and 1 is an exact match. Jaro-Winkler is an extension of the Jaro metric [20], meaning that it accounts for insertions, deletions and transpositions when comparing two strings. The exact implementation details are outside the scope of this paper, but Jaro-Winkler is a standard algorithm for string-based matching and has been described in detail elsewhere, see e.g., [18].

2) Time-based matching: Looking at the point in time when various aliases have created their forum posts can give important clues to whether two different aliases refer to one and the same individual or not. However, to compare the creation time of two posts is not reliable enough since it is likely that two individuals create their posts during the same time period without any other reason than pure chance or living in the same time zone. In our implementation we therefore create time profiles based on the relative distribution of the time of day when the postings have been made by the aliases, where the time of day is discretized into intervals of equal size (in this case each interval corresponds to one hour). To exemplify, assume $AliasX$ has written 8 posts in total, with the following times of posting: 07:16, 07:19, 07:27, 07:59, 09:18, 10:23, 12:16, and 12:42. The first step is now to construct a feature vector corresponding to the absolute frequency with how many posts that have been written each hour:

$$< 0, \ldots, 4, 0, 1, 1, 0, 2, 0, \ldots, 0 >$$

Since we are interested in how the number of posts are distributed throughout the day rather than in the exact numbers (some aliases will be used more frequently than others) we are in the next step normalizing the feature vectors, resulting in:

$$< 0, \ldots, 0.5, 0, 0.125, 0.125, 0, 0.25, 0, \ldots, 0 >$$

In this way, a normalized time profile is created for each alias. An example of two time profiles is shown in Figure 1. We have also experimented with binary feature vectors and different larger time intervals but this yielded slightly worse results in general. In addition to use the distribution of the number of posts per hour, it would also be interesting to see what happens if other time periods are used, e.g., the distribution of posts per week or month. This could show interesting information which does not become visible when looking at the selected time period. However, for the work presented in this paper the described time period has been used, so the use of longer periods remain as future work.

Once the time profiles have been constructed, the Euclidean distance is used for calculating how far away two time profiles are from each other. The smaller the distance between the time profiles for two aliases, the more likely it is that the two aliases belong to the same user. Formally, the Euclidean distance between two vectors p and q is given by:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (1)$$

The fact that two aliases have similar (dissimilar) time profiles does not mean that the individuals are necessarily the same (different) individuals, but it can be used as evidence for or against such hypotheses.

3) Stylometric matching: Another useful matching technique when trying to find out if multiple aliases belong to the same user is stylometric matching, where stylometry refers to the statistical analysis of writing style [14]. With this technique, the author's writing style is analyzed by constructing a "writeprint", which in many ways resemble how fingerprints can be used. A lot of algorithms and features for stylometry-based author identification have been proposed throughout the literature, see e.g., [21], [22], [23]. Most work has however been focused on closed-world problems with a small number of potential authors and a rather large quantity of text to build the stylometric profiles from (such as long literary books). Much less research has been devoted to problems with a large number of potential authors and smaller quantities of text material, or to the cyberspace domain in general [14]. We have in our implementation included a subset of the features used in the recent article by Narayanan et al. [12] and an extra feature: the frequency of sentence lengths. See Table I for a full list of the features we have used. The **count** column refers to the number of dimensions a certain feature demands. As an example, there are 26 dimensions for the $Letter$ feature (one for each letter).

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".

Many modern algorithms for author identification are based on machine learning, such as support vector machines (SVMs)

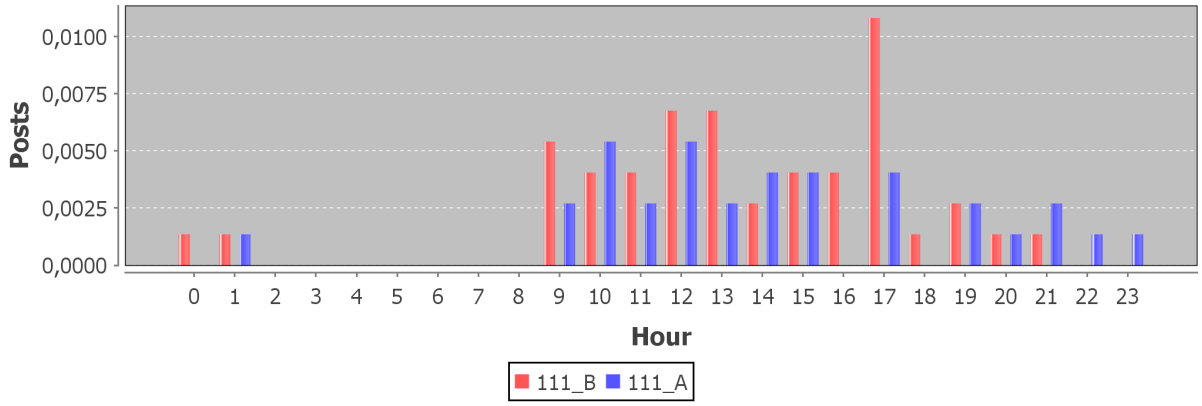

Fig. 1. Example of time profiles for two individuals.

TABLE I
THE FEATURES USED FOR STYLOMETRIC MATCHING (THE LIST OF FUNCTION WORDS USED CAN BE FOUND IN [12]).

Category	Description	Count
Word length	Frequency of words with 1-20 characters	20
Sentence length	Frequency of sentences with various lengths	6
Letters	Frequency of a to z (ignoring case)	26
Digits	Frequency of 0 to 9	10
Punctuation	Frequency of characters . ? ! , ; : ( ) ” - ´	11
Function words	Frequency of various function words	293

and decision trees. Such algorithms can be used for learning classifiers to generalize from training data in order to make good classifications on (previously unseen) test data, but are in general not appropriate for determining how similar the writeprints of two aliases are. We are therefore using the more basic approach to compare how similar the (normalized) stylometric feature vectors are for two aliases by simply calculating the cosine of the angle between them:

$$\cos(p,q) = \frac{p \cdot q}{\|p\|\|q\|} = \frac{\sum_{i=1}^{n} p_i \times q_i}{\sqrt{\sum_{i=1}^{n} (p_i)^2} \times \sqrt{\sum_{i=1}^{n} (q_i)^2}} \quad (2)$$

There are many other ways that also could be used to compare the similiarity between two stylometric feature vectors, but the use of cosine similarity is straightforward to implement and seems to work out well, as shown in our experiments.

4) Social-network based matching: The last type of matching technique we have implemented is what we have chosen to refer to as social-network based matching. The underlying idea of this is that a mapping and comparison of the social network of two aliases can reveal if those aliases are similar in the sense of whom they are connected to. The social network can be based on various information, depending on what the discussion forum look like. On some forums (such as the forums we have used in our experiments), there are friend or ”buddy” lists available, in which the user can mark other users as friends. On many forums such friend lists are lacking, but also other kinds of information can be used to create social networks, such as thread networks (connecting users who have made postings in the same thread) or topic networks (connecting users who have written about the same topic). In order to create topic networks, it is necessary to first extract the topics from the posts. This can be done with various topic detection and topic extraction methods such as the ones presented in [24], but is outside the scope of this paper.

To illustrate how social-network based matching can be used, consider the alter ego case discussed in Section III. For this case, it makes sense to measure how similar the thread networks are for two aliases when trying to determine if the aliases belong to the same user or not. In general, it is likely that both aliases will make postings in the same thread if they are alter egos, since the reason for creating an alter ego or sockpuppet often is to support one’s own arguments.

No matter if the constructed social network is based on friend-, thread- or topic information, we use vertex similarity to calculate how similar two aliases are in terms of their social network. The vertex similarity can be calculated as a function of the number of neighbors in common for two aliases. If the total number of neighbors should not impact the results too much, a normalization process in which the node degrees are taken into account is needed. Let Γ_p be the neighborhood of vertex (alias) p in the network and Γ_q be the neighborhood of vertex (alias) q. Now, the number of common neighbors is calculated as $|\Gamma_p \cap \Gamma_q|$. The normalization can be done in various ways (such as with dice or cosine similarity), but in our implementation we make use of the Jaccard similarity coefficient $J(p,q)$, where:

$$J(p,q) = \frac{|\Gamma_p \cap \Gamma_q|}{|\Gamma_p \cup \Gamma_q|.} \quad (3)$$

In Figure 2 we illustrate the ego networks of aliases A and C, where they have two neighbors in common (E and F).

B. Matching of aliases

In the previous section we have described a number of matching techniques, where each classifier outputs a similarity between two aliases. Which classifiers to include depends on the task at hand, e.g., if we are dealing with a concealed or

Fig. 2. Social networks where alias A and alias C have two neighbors in common (E and F).

non-concealed case (recall the list in Section III). If we are dealing with a non-concealed case all matching techniques may be used, while the string-based technique probably will be of little or no value for the concealed case.

Once the appropriate set of classifiers has been determined, the matching techniques can be combined in various ways. If we want to decide whether two aliases should be merged or not, a straightforward approach is to combine the results from the used classifiers into a (weighted) average. On other occasions we may want to find out which alias in a set of aliases $A = \{a_1, \ldots, a_n\}$ a certain selected alias a_0 is most similar to. Depending on the size of the set, various approaches can be used. If the set is reasonably small it makes sense to output a rank from each classifier (where the rank is based upon the computed similarities for each alias in the set). In this case a (weighted) average of the rankings can be calculated. This is the approach used in our experiments which are presented in Section IV. If there are many aliases to compare and the alias matching has to be performed in a near real-time application, it may take too long to apply all matching techniques in parallel. In such cases it is better to instead apply the matching techniques in sequence, starting with the computationally cheap techniques for making a first coarse filtering. After such a filtering where only the k best matches are kept, remaining techniques can be applied on the filtered subset of aliases.

IV. Experiments

A fundamental problem with algorithms for alias matching is that it is hard to find reasonable datasets to evaluate the suggested algorithms on. To the best of our knowledge, there are no standard datasets for alias matching available. We have not been able to find any datasets containing multiple aliases where the ground truth is known and where all features we have been suggesting are available. For this reason, we have limited the current experiments to only take into account stylometric and time-based matching. Hence, the social-network based matching techniques and the string-based matching techniques have not been utilized in the performed experiments. In this section, we describe the design of the conducted experiments and present the experimental results.

A. Experimental design

In our experiments we have used a dataset containing data from the Irish web forum site https://www.boards.ie/. The data in total consists of around 9 million documents in SIOC format and takes about 50 gigabytes of disk space. We have however limited our experiments to data from year 2008 which has been parsed and extracted into a SQL database. From this database we have extracted the posts from users who have written at least 60 messages in total.

From this set of users with sixty or more posts, we have first selected a small set of users ($n = 10$) (where the selection is based on their ID-number). Each of these users have been split into two separate users u_{ia} and u_{ib}, where $1 \leq i \leq 10$ and odd posts are assigned to user u_{ia} and even posts to u_{ib}. Now, each user in the set $\{u_{1a}, u_{2a}, u_{3a}, ..., u_{na}\}$ is compared, one at a time, with all the users in the set $B = \{u_{1b}, u_{2b}, u_{3b}, ..., u_{nb}\}$. Based on the results from the stylometric matching and time-based matching we rank the members of set B according to how similar they are to the selected user. We also combine the two techniques by computing a rank as an average of the ranks obtained from the stylometric and time-based classifiers (i.e., both classifiers are assigned equal weight). 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 50 to 1000 in steps of 50. The tests have been carried out on a computer with Mac OS X 10.8.2, 2.66 GHz Intel Core 2 Duo processor and 4 GB 1067 MHz memory.

The experimental design is intended to cover both the ”concealed” and ”non-concealed” cases described earlier, since no techniques are used that would not work for the other case (which would not hold true if string-based matching techniques would have been tested in the experiment).

B. Experimental results

The results from the experiments are shown in Figure 3 and Figure 4. Looking at the overall results, we can see that there is, as expected, a decrease in the accuracy when increasing the number of users. It can also been seen that the time-based matching consistently performs better than the stylometric matching for both top-1 and top-3 ranking. The combination of the classifiers consistently perform better than the two classifiers individually. Studying the results in further detail, we can see that the correct alias is ranked first with over 70% accuracy when there is up to 50 users. The accuracy drops as the number of users is increased further, but it is still higher than 60% for up to 150 users and 55% for 250 users. The accuracy for the combined results thereafter become more stable, remaining at 43% for 1000 users.

If we instead only demand that the correct user should be in the top-3, the combined classifiers yield accuracies over 80% for up to 100 users. The accuracy is still over 70% for up to 250 users. The accuracy is then slowly decreasing, resulting in an accuracy on 56% for 1000 users.

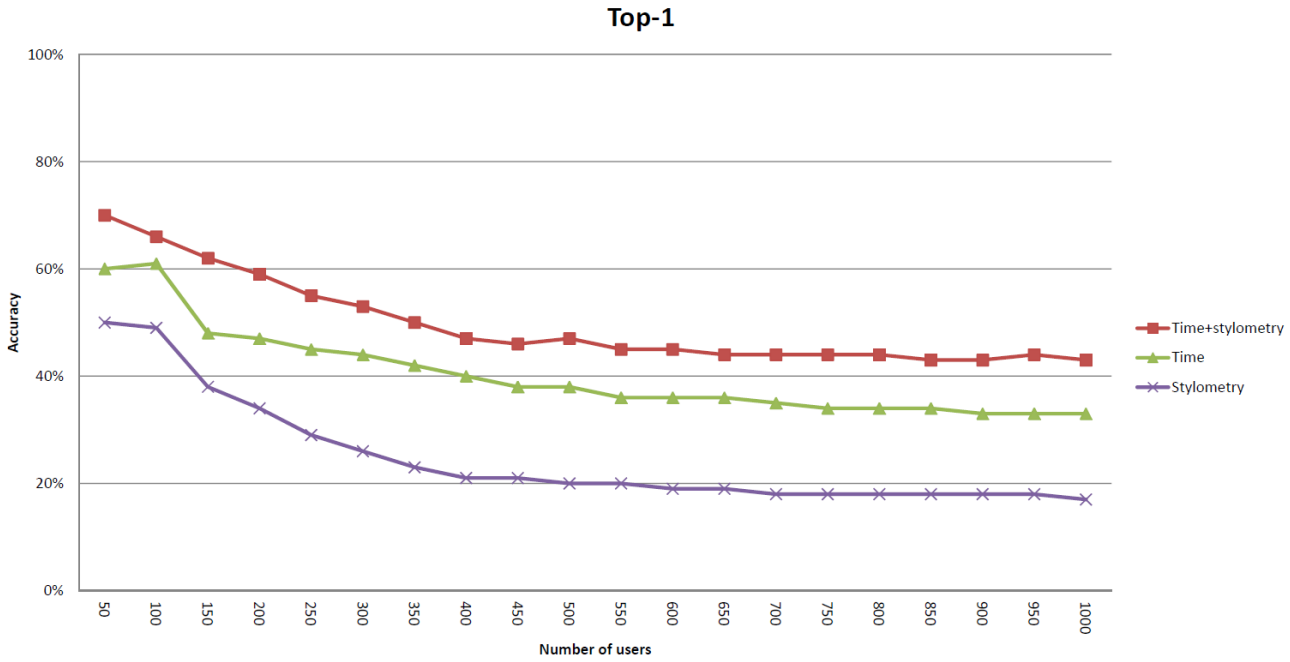

Fig. 3. Results for top-1.

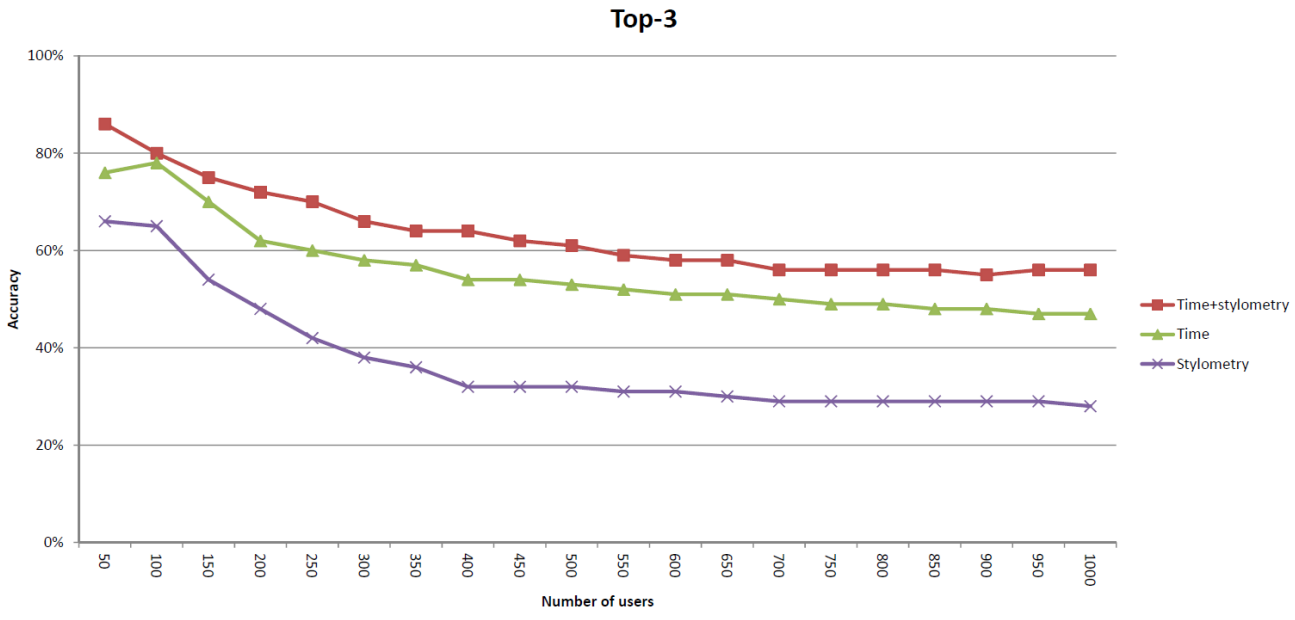

Fig. 4. Results for top-3.

V. Discussion

The presented results indicate that there is a possibility to use algorithms for detecting the use of multiple aliases on discussion forums using quite limited amounts of data. Although we have only tried the methods on data from a single discussion forum, there are no reasons for why this kind of methods cannot be used also for blogs or other kinds of social media services. There is also a possibility that the same methods could be use for linking user accounts from various social media services to each other, although this is not as obvious. It is likely that espcially the time-based matching technique would be affected if the used postings would have been acquired from several social media services. As an example, we do not necessarily expect a user's time profile on Twitter to match that of a web forum since such services are utilized in different ways. Similarly, it can be expected that the stylometric profiles from Twitter and web forums can look differently due to the constraints on the number of characters in tweets. Hence, the suggested techniques would probably have to be modified if they should be used for linking user accounts from many types of social media services.

The techniques that have been presented in this paper can also be generalized to other domains, such as record linkage of bibliographic data.When deciding whether two papers have been written by the same author or not, an obvious technique to use is string-based matching. However, to use only similarities in names is not enough, since two individual authors can have the same name, and since a middle name sometimes is included and sometimes not, a name can be misspelled, etc. For this reason it also becomes highly relevant to look at the social networks of the authors (who they have written paper with before) or even use author attribution or stylometric techniques to find out if the writing style is similar between various papers. Also time can be a relevant feature since it is unlikely that a paper written in the 60's has the same author as a paper written very recently, even though the author name may be the same.

Although the presented results show an interesting potential to be used for counter-terrorism purposes, they also raise privacy concerns since the same kind of techniques can be used also for more doubtful purposes. Applications that attack pseudonymity can pose a threat to the privacy of innocent people since the linking of anonymous postings made by, e.g., a dissident in a totalitarian regime to other pieces of text where the author reveals his or her identity could have severe consequences. Also commercial companies may have an interest in such techniques due to advertising campaigns and similar applications. These kinds of problems with alias matching techniques are discussed further in [12], [19], [25].

Our focus in the experiments have been on evaluating the proposed techniques on a controlled dataset where the ground truth is known. Obviously, what is more interesting in the long run is to see how well the algorithms perform "in the wild". However, before such tests can be made it is important to verify that the implemented algorithms work as expected, since it otherwise becomes hard to judge whether the found candidates actually should be merged or not. Hence, the presented experiments is a first attempt to evaluate some of the techniques which have been proposed in the paper.

VI. Conclusions and future work

We have presented four different types of techniques for alias matching: string-based, stylometric-based, time profile-based, and social network-based matching. Several of those matching techniques have been proposed and used earlier, but there are no earlier attempts to use them in combination to find the use of multiple aliases within discussion forums. Moreover, we are not aware of any previous attempts to use time profile-based matching for alias matching purposes. In our experiments on forum data we have evaluated how accurate the stylometric and time-based techniques are on their own and in combination. The results suggest that our novel time-based matching technique yields better accuracy than stylometric matching, and that the combined result is always better than the individual classifiers alone. Furthermore, it is shown that quite good accuracy can be achieved also with limited amounts of posts and a large number of potential authors (e.g., 70% for 50 users, 55% for 250 users and 43% for 1000 users).

For future work, we attempt to improve the implemented matching techniques by adding more features and applying feature reduction techniques such as principal component analysis. We also hope to find non-synthetic data on which all the implemented classifiers can be tested. It would also be interesting to test the methods on a large scale, where there are thousands of potential authors. Finally, as also pointed out by one of the reviewers of this paper, it can be interesting for future work to look into other evaluation criteria than just accuracy to judge how well the algorithms work.

Acknowledgments

This research was financially supported by Vinnova through the Vinnmer-programme, and by the Swedish Armed Forces Research and Development Programme.

References

[1] E. Pressman, "Risk assessment decisions for violent political extremism 2009-02," *Public Safety Canada*, 2009.

[2] A. Y. Zelin and R. B. Fellow, "The state of global jihad online," *New America Foundation*, 2013.

[3] J. Corbin, *Al-Qaeda: In Search of the Terror Network that Threatens the World*. Thunders Mouth Press / Nation Books, New York, 2002.

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

[5] M. K. Reiter and A. D. Rubin, "Anonymous web transactions with crowds," *Communications of the ACM*, vol. 42, no. 2, pp. 32–48, Feb. 1999.

[6] J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Mårtenson, and P. Svenson, "Analysis of weak signals for detecting lone wolf terrorists," in *Proceedings of the 2012 European Intelligence and Security Informatics Conference*, 2012, pp. 197–204.

[7] C. Yang and T. Ng, "Terrorism and crime related weblog social network: Link, content analysis and information visualization," in *Proceedings of the 2007 IEEE Conference on Intelligence and Security Informatics*, may 2007, pp. 55 –58.

[8] M. Yang, M. Kiang, Y. Ku, C. Chiu, and Y. Li, "Social media analytics for radical opinion mining in hate group web forums," *Journal of Homeland Security and Emergency Management*, vol. 8, no. 1, 2011.

[9] A. Abbasi and H. Chen, "Affect intensity analysis of dark web forums," in *Proceedings of the 5th IEEE International Conference on Intelligence and Security Informatics*, 2007.

[10] J. Brynielsson, A. Horndahl, F. Johansson, L. Kaati, C. Mårtenson, and P. Svenson, "Harvesting and analysis of weak signals for detecting lone wolf terrorists," *Submitted to Security Informatics*, 2013.

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

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

[13] J. Dahlin, F. Johansson, L. Kaati, C. Mårtenson, and P. Svenson, "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 2012)*, 2012, pp. 850–857.

[14] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 378–393, 2006.

[15] H.-C. Chen, M. K. Goldberg, and M. Magdon-Ismail, *Intelligence and Security Informatics*. Springer, 2004, ch. Identifying Multi-ID Users in Open Forums, pp. 176–186.

[16] M. Shaikh, N. Memon, and U. Wiil, "Extended approximate string matching algorithms to detect name aliases," in *Proceedings of the 2011 IEEE International Conference on Intelligence and Security Informatics*, july 2011, pp. 216 –219.

[17] V. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, 1966.

[18] W. E. Winkler, "String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage," in *Proceedings of the Section on Survey Research Methods*, 1990, pp. 354–359.

[19] D. Perito, C. Castelluccia, M. Kaafar, and P. Manils, "How unique and traceable are usernames?" *Privacy Enhancing Technologies*, pp. 1–17, 2011.

[20] M. A. Jaro, *UNIMATCH: A Record Linkage System*. Bureau of the Census, Washington, 1978.

[21] 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.

[22] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Trans. Inf. Syst.*, vol. 26, no. 2, pp. 7:1–7:29, Apr. 2008.

[23] P. Juola, "Authorship attribution," *Found. Trends Inf. Retr.*, vol. 1, no. 3, pp. 233–334, 2006.

[24] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation," in *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, 2010.

[25] J. R. Rao and P. Rohatgi, "Can pseudonymity really guarantee privacy?" in *Proceedings of the 9th conference on USENIX Security Symposium*, vol. 9. Berkeley, CA, USA: USENIX Association, 2000.