2010 Second Cybercrime and Trustworthy Computing Workshop

Authorship Attribution for Twitter in 140 characters or less

Robert Layton
Internet Commerce Security Laboratory
University of Ballarat
Email: r.layton@icsl.ballarat.edu.au

Paul Watters
Internet Commerce Security Laboratory
University of Ballarat
Email: p.watters@icsl.ballarat.edu.au

Richard Dazeley
Data Mining and Informatics Research Group
University of Ballarat
Email: r.dazeley@ballarat.edu.au

***Abstract*—Authorship attribution is a growing field, moving from beginnings in linguistics to recent advances in text mining. Through this change came an increase in the capability of authorship attribution methods both in their accuracy and the ability to consider more difficult problems. Research into authorship attribution in the 19^{th} century considered it difficult to determine the authorship of a document of fewer than 1000 words. By the 1990s this values had decreased to less than 500 words and in the early 21^{st} century it was considered possible to determine the authorship of a document in 250 words. The need for this ever decreasing limit is exemplified by the trend towards many shorter communications rather than fewer longer communications, such as the move from traditional multi-page handwritten letters to shorter, more focused emails. This trend has also been shown in online crime, where many attacks such as phishing or bullying are performed using very concise language. Cybercrime messages have long been hosted on Internet Relay Chats (IRCs) which have allowed members to hide behind screen names and connect anonymously. More recently, Twitter and other short message based web services have been used as a hosting ground for online crimes. This paper presents some evaluations of current techniques and identifies some new preprocessing methods that can be used to enable authorship to be determined at rates significantly better than chance for documents of 140 characters or less, a format popularised by the micro-blogging website Twitter[1]. We show that the SCAP methodology performs extremely well on twitter messages and even with restrictions on the types of information allowed, such as the recipient of directed messages, still perform significantly higher than chance. Further to this, we show that 120 tweets per user is an important threshold, at which point adding more tweets per user gives a small but non-significant increase in accuracy.**

I. Introduction

The Internet has typically facilitated shorter forms of communication more easily than traditionally longer forms such as handwritten letters and essays. One example of a shorter form of communication online is Internet Relay Chat (IRC) rooms, which provides a text based 'chatting' service, where users post short messages to a 'chatroom' which is then readable by all users in this chatroom. A typical message in IRC is very short and it is common for individual messages to be a single sentence or less. Twitter has surged in popularity in recent years and now reports that it receives over 50 million messages (called *tweets*) per day [34]. Twitter is a micro-blogging website and allows users to post messages with the restriction that posts must be 140 characters or less in length. Another social network based website, Facebook[2] has many different forms of communication between users such as instant messaging, internal messages and wall posts, with most of these forms of focusing on shorter messages. Many other websites include comments sections, such as YouTube[3] and blogs, which are typically focused on shorter messages. There is a clear trend overall towards shorter messages on the Internet, which is also shown in other technologies, where short message services (SMS) have become a very popular use of mobile phones.

This trend of shorter online messages is also seen in cybercrime where crimes such as phishing and cyberscams usually occur with shorter messages such as fraudulent emails [27], forum posts [2], underground IRC rooms [33], on Facebook and Twitter, as well as many other websites [1]. Cybercriminals attempt to use these websites and web services to gain information that can lead to identity theft or identity fraud [1]. Internet based chat rooms and forums have been known to be a tool used by cybercriminals to sell stolen identity information; sell and buy malware and botnet access; and, also to trade in illegal pornography and copyrighted materials [2]. Cybercrime is a growing area of crime and has been recognised as a priority by many governments, such as the Australian Federal Government[10].

As law enforcement agencies (LEAs) attempt to track down and monitor cybercriminals using these technologies, it is becoming increasingly important to determine the authorship of a message as technologies such as fast flux make it increasingly difficult to track down offenders using network based tools[13]. Accurate authorship attribution of such shorter messages help LEAs to prosecute these criminals [9]. Illegal resources have been shared on Internet forums for many years with very little description and some criminals write about their crimes under anonymous authorship either to gain notoriety or to profit from their crimes [2]. To be able to track and determine the authorship of these messages would provide a large leap in the ability of LEAs to prosecute criminals based on Internet postings.

[1]Twitter: `http://www.twitter.com/`

[2]Facebook: `http://www.facebook.com/`

[3]YouTube: `http://www.youtube.com/`

978-0-7695-4186-0/10 $26.00 © 2010 IEEE
DOI 10.1109/CTC.2010.17

It is widely considered in the literature that people exhibit particular trends in their writing which in turn reveals facts about them, such as their age, gender and personality traits [4]. An example of a stylistic marker is the use of 'txting' shorthand such as `2nite`[4] and the use of emoticons such as `:-)` [6]. It has been shown repeatedly in the literature that determining the particular author from a set of candidate authors is possible by looking at the documents that each author has written and matching a new document of unknown origin to a profile built of each author [14], [23], [26], [29], [32], [35]. This process is known as authorship attribution and is part of the field of authorship analysis which includes author profiling [4], similarity detection [22] and authorship intent determination [17].

Authorship attribution has its roots in stylometry, before much of the work in the field moved to simple statistical analysis [28]. With the exponential increase in computing power, the 20^{th} century saw a drastic rise in the complexity of the statistical analysis of documents and also a gradual shift towards machine learning methods [31]. By the 21^{st} century, a majority of the work in authorship analysis is now performed using complex statistical analysis that would not have been computationally feasible just 50 years earlier [18]. This increase in complexity also saw a decrease in the required length of a document to achieve good accuracies in classification. As an example, 19^{th} century authorship studies considered blocks of 1000 words to be a lower limit on the size of a block of text to analyse [25], and that even larger blocks were needed to removing accidental irregularities in writing style. By the 1990s 250 words, a quarter of the previous limit, was considered a limitation on the length of a document that could be attributed accurately [11], however more recent work has been able to break this barrier and achieve reliable authorship attribution in 250 word documents [3].

The reduction in the length of a document required has increased the scope of available applications. Where as early work in authorship attribution focused on documents of longer length such as the *Federalist papers* [28], more recent work is able to look to online documents such as blog posts [26] and at even shorter length forum postings [2]. The shorter required length has increased the viability of authorship attribution in an important area of online activity, cybercrime related documents.

The work presented here is closely related to an area of text mining called 'chat mining' [19], in which Internet based discussions are mined for certain information such as authorship. This area is motivated strongly, as this research is, by investigations into cybercrime [24] due to the need for information to help LEAs to conduct and prosecute offenders online, where direct attribution of attacks is often obfuscated using online anonymity tools such as the use of proxies or botnets.

[4]Shorthand for 'tonight'

A. Research Questions

In the presented research, we extend the field of authorship analysis towards determining the authorship to one of the shortest forms of communication currently in use - a tweet. A tweet is the name given to a post from the website Twitter, which is an example of a micro-blogging website in which users post messages about whatever topic they wish, but are limited in the number of characters they can use for a single tweet. This limit is 140 characters for Twitter, which is the limit used in this research. We are motivated by the continued use of shorter messages in cybercrime and aim to determine the viability of authorship analysis on these shorter messages in order to help investigations into these crimes. To those goals, this research aims to answer the following questions:

1) How effective is an existing leading authorship attribution technique (SCAP) at attributing tweets to a given author?
2) What properties of tweets enable or prohibit effective authorship attribution in tweets with respect to a cybercrime investigation?
3) Does splitting individual author's authorship profiles into a set of sub-profiles provide a significant benefit over using a 'complete' author profile?
4) How many tweets per author are needed for an accurate profile of an author is there a threshold in which increasing the number of tweets provide a non-significant accuracy gain?

The rest of this paper will follow this outline. The next section will provide an overview of Local n-grams and the SCAP methodology, which is one of the current leading methods in authorship analysis on structured text. The methodology that was used for the experiments will then be presented in section III, followed by the results from those experiments in section IV. The outcomes from those results will be discussed in section V along with the conclusions which will outline the contributions made in this paper in detail.

II. Local n-grams

Modern authorship analysis typically uses machine learning algorithms to investigate multiple variables and their relationships to the authorship of the documents in the training corpus. This type of learning closely follows other machine learning methods, such as a classification framework for authorship attribution [15] or a data clustering framework for similarity detection [21]. Differences to many other data mining applications are usually in the first stages of the data mining process, such as feature extraction from the text of the document [35], document preprocessing methods [20] and specific distance metrics [16]. These differences relate directly to the method of calculating the distance or similarity between two documents in the corpus. This is necessary as authorship analysis is performed on text documents and many machine learning algorithms deal specifically with numbers and vectors. Once distance is able to be measured between documents, this limit is overcome and a large range of data mining methods,

such as classification or clustering algorithms, are able to be used to generate models of the data.

Using character level n-grams to develop author profiles has proven to be a successful method of translating a corpus of documents into a set of models for authors [16]. Once the models are generated as n-gram distributions, the best matching author is decided by finding the nearest profile, calculated using a distance metric that accounts for the L most frequent n-grams in a document and the frequency with which they occur. These frequency lists are compared on the assumption that documents written by the same author use the same n-grams with a similar overall frequency. This work is an extension of a previously derived method [7] which was, at least computationally, ahead of its time. The results in [16] are significantly above chance rates, with many experiments achieving results above 80% authorship attribution accuracy and all experiments outperforming previous results on the same authorship attribution problems.

The use of n-grams for authorship attribution was furthered in work by [12], which removed the complex distance metric in [16], replacing it with a simple set intersection based metric. The distance between a document and an author's profile is the size of the intersection between the set of the top L most frequently occurring n-grams for the document and profile. The Simplified Profile Intersection (SPI) was shown to be an effective distance metric for evaluating the authorship of the source code of computer programs, a highly structured form of written document. SPI either outperformed or equalled the relative distance (RD) given in [16] in all of the given experiments and is shown to be more robust than the RD when the profile size (L) increases for smaller values of n.

SPI is used as part of the Source Code Authorship Profile (SCAP) methodology [12], which proceeds as following:

1) Divide the known corpus into training and testing documents
2) For each author:
    a) Concatenate all training documents per author into a single document
    b) Calculate the top L most frequent n-grams for the combined document
    c) This list is the Simplified Profile for this author
3) Each testing document is assigned to the profile with the largest SPI similarity

To determine the best guess for attribution of testing documents, each testing document is profiled as a list of the L most frequently occurring n-grams. To calculate the similarity, the normalised size of the intersection of each user's profile and the testing document's n-gram list is used, the user profile with the highest similarity is declared the best match for the given document.

III. Methodology

This research presents an exploratory look into authorship attribution of tweets, aiming to investigate the viability of authorship attribution on these shorter messages. Using a collection of tweets collected from publicly available feeds (see subsection III-A) the SCAP method will be applied directly to the raw text of the messages. Other information, such the date a tweet is posted, will not be used while the author of a tweet will be used for the classification class only. All usernames are assumed to be for a single author, although this is not verified in the dataset collection. The accuracy will be tested as described in subsection III-B and then an investigation of the attributes of tweets relating to the semi-structured nature of tweets will be performed, described in subsection III-C. Once this has been performed, the impact of using sub-profiles will be investigated as outlined in subsection III-D. Finally, the impact of the number of tweets on the accuracy of SCAP will be in investigated using the methodology outlined in section III-E for both profiles and sub-profiles.

A. Tweet Dataset

The tweet dataset used is a collection of 14,000 Twitter users and their most recent tweets as of February 2010[5]. The most recent 200 tweets for each user were collected. The users were collected by searching twitter for a random function word from the list in [35] and collecting the usernames of each tweet that was returned by Twitter's search engine. This list of returned values is the most recent tweets posted to Twitter containing the search term.

Function words were search for over 4 days in regular 15 minute intervals, collecting over 56,000 usernames, of which 14,000 were selected at random for the dataset. For each of those usernames up to a maximum of 200 hundred tweets. The dataset contains the username, date and contents for each tweet, although the date is not used in these experiments. Only publicly available tweets were collected and any user with a private profile would not have been returned in the initial search and was therefore excluded from this dataset.

B. Applying SCAP

To determine the viability of performing authorship attribution on tweets, a preliminary experiment was performed where the SCAP methodology was applied directly onto the tweets dataset outlined in section III-A. This experiment aims to answer the first research question of this research; how effective is an existing authorship attribution technique (SCAP) at attributing tweets to a given author? For the test, 50 authors were selected at random from the dataset described in subsection III-A and their tweets comprise the sample used, giving an average chance rate accuracy of approximately 2%. Ten fold cross validation was used for dividing the corpus into training and testing documents The tweets were divided into ten random sub-samples over the entire sample and are therefore not normalised for author balance.

The SCAP method takes two parameters, the values for L and n. Values of L were selected by searching a wide range of values and narrowing searches around interesting values. Values for n were selected to be between 2 and 7 as this is comparable to values found in the literature [22]. The better

[5]A copy of the dataset can be obtained by contacting the authors.

values were determined by finding the highest mean values and testing if they are significantly better than other values for the parameters.

C. Structural Investigation

Tweets are a semi-structured form of text and while the structures are entirely optional, they are used regularly. There are two main types of structure in tweets [8], which are:

- @**replies**: To direct a message at a user with a given *username*, include `@username` in the tweet
- **#tag**: To give a message a tag, which can be used for searching and grouping similar messages across different people, include `#tagname` in the tweet

To investigate the impact of these structures on the accuracy of the system, an experiment was run that removes most of the information in these structures, in order to answer the second research question; what properties of tweets enable or prohibit effective authorship attribution in tweets with respect to a cybercrime investigation? It could be reasonably expected that in some cybercrime settings, a cybercriminal posting anonymously would be careful not to reply to regular contacts or apply their normal tagging to their tweets. For this reason, it is important to consider how effective authorship attribution can be when this information is not included.

Another reason to consider tweets without this information is the lack of authorship choice in using these structures. To direct a tweet at the user with username *example_user123*, the tweet will include `@example_user123`. The decision to use a directed reply is a decision made by the user, but 15 of the characters in the structure (the username) were used without any authorship decisions by the author of the tweet. For a 140 character tweet, there is a large portion of the message that the user did not actually 'author'. A similar problem occurs with the tagging syntax, where tag names can also account for over 10% of the message in some case. This could have an effect on the overall accuracy of the SCAP method which would be uncovered through this experiment.

To determine the impact of these parts of the message on the overall accuracy of the profile based method, three preprocessors were used that perform the following actions:

- **At** preprocessor: Replace all instances of @**username** with @
- **Hash** preprocessor: Replace **#tag** with **#**
- **Both** preprocessor: Both **At** and **Hash** applied

These preprocessors aim to remove some of these extra clues that could be contained in tweets. For instances, by conversing with different people, the `@username` structure could contain more information about the choices a user's friends make about their Twitter username. These choices could be related to demographic information and convey information that could help the authorship analysis process. The difference in accuracy after applying these preprocessors determines the importance of this information in enabling or prohibiting authorship attribution in tweets.

D. Sub-profiles

The SCAP methodology, described in section II, works by finding a profile of an author that is ideally dense, in that documents by the same author are very similar to each other. Additionally, author profiles should be well separated from other profiles as ideally documents by different authors are less similar than each other. A set of profiles with this characteristic is likely to accurately classify future instances, as they will be classified to their nearest author. A measure of how well data is formed in this way is the silhouette coefficient [30] which is typically used in unsupervised learning. The silhouette coefficient is near 1.0 for dense, well separated clusters and near -1.0 for clusters that heavily overlap and is defined for each individual point (an overall score is calculated by taking the mean for each point in a dataset). The silhouette coefficient for profile p, where a_p is the mean distance between all documents in p and b_p is the mean pairwise distance to the nearest profile of another author, is defined in equation 1. The Silhouette Coefficient for a set of profiles is simply the mean of the silhouette coefficients for each profile within the set.

$$s_p = \frac{b_p - a_p}{max(a_p, b_p)} \tag{1}$$

To evaluate the quality of the profiles, the silhouette coefficient will be used to evaluate the profiles by measuring the internal profile distance compared to the distance between profiles. Profiles with a poor silhouette coefficient may be composed of separate 'sub-profiles', clusters within the profile of similar instances. These sub-profiles may have a high silhouette coefficient if considered separate from the other sub-profiles for a given user. If this is found to be the case then the accuracy of the overall system should improve, as the profiles and sub-profiles would have a higher silhouette coefficient and therefore a better chance at accurately classifying future documents.

To calculate sub-profiles, the k-means clustering algorithm is run for two clusters and the silhouette coefficient is calculated for the resulting sub-profiles within the authors profile. The mean intra-sub-profile distance (a), is compared against the mean distance between the two sub-profiles (b) and the mean silhouette coefficient for the author's sub-profile is calculated. A positive silhouette coefficient implies that the two sub-profiles are distinct and they are used to profile the author. If the resulting silhouette coefficient is negative or zero, then there is some overlap between the sub-profiles. The sub-profiles are discarded and the author is profiled using a single profile only.

To compensate for the randomness of the results obtained by the k-means algorithm, the k-means++ seeding algorithm was used, as was a number of trials for the sub-profiling of each author. The k-means++ seeding algorithm [5] was used to seed the initial clustering for the k-means algorithm. While the seeding from k-means++ reduces the need for a large number of trial runs, there is still an element of chance in arriving at the best partition, even with just two clusters. To compensate for this, 30 number iterations of k-means are

performed with randomised starting values for each user and the highest silhouette coefficient is used, given that it is a positive value.

The use of sub-profiles will be evaluated using the same process as described for the preliminary SCAP application in section III-B. Profiles will be generated for each author and then each profile will be tested to see if a number of sub-profiles can more accurately describe the user's writing style. Once the sub-profiles have been generated, testing instances are classified according to the closest profile or sub-profile and the accuracy of the system will then be evaluated using 10 fold cross validation.

E. Restricting the number of tweets

To answer the final research question, the number of tweets per author will be reduced to determine the impact of this number on the final accuracy. The SCAP methodology on the twitter dataset with no preprocessors will be used for this experiment. Values between 20 and 200 will be used in steps of 20, to account for the range up to the 200 tweets per user limit that was available in the original dataset. A consistent reduction is expected in the overall accuracy after applying the SCAP methodology as above using 10 fold cross validation. Further to this expectation, a t-test will be performed to determine whether adding more tweets significantly increases the accuracy. This test determines whether there is a critical threshold for the number of tweets that should be met, before adding tweets is less effective overall.

IV. Results

A. Benchmark SCAP Accuracy

To answer the first research question posed in subsection I-A, in the first experiment the SCAP methodology is applied directly to a sample of tweets from the collected dataset. The SCAP methodology was applied as outlined in subsection III-B and the results are listed here. The parameters to the SCAP methodology provide the largest issue, as two parameters must be searched. Results given in [12] suggest that the parameter space is probably smooth but non-linear. For this reason, a wide range of values were searched for L and ranges near interesting values were searched in more detail. For the values for n, previous results in [22] indicate that values between 2 and 7 inclusive should be sufficient to gain effective results, particularly for English-tending datasets such as the one used here[6].

A preliminary search on values for L showed little difference between L values, which is considered due to the size of the individual tweets. This result is unsurprising when comparing using values of L between 200 and 3000, as reported in [12], which is well above the 140 character limit for an individual tweet. A search for L values less than 200 returned no significant differences in the resulting authorship profiles, as diversity within an author is not significant. For this reason, all n-grams are included, so L can be considered as any value more than the maximum number of distinct n-grams for any author.

n-value	μ	σ
2	0.531	0.015
3	0.708	0.013
4	0.729	0.022
5	0.719	0.015
6	0.706	0.017
7	0.682	0.018

TABLE I
OVERALL MEAN AND STANDARD DEVIATION OF THE CROSS FOLD VALIDATION ACCURACY PERFORMED USING SCAP FOR EACH VALUE OF n.

n-value	Both-μ	Both-σ	Hash-μ	Hash-σ	At-μ	At-σ
2	0.357	0.021	0.520	0.016	0.371	0.015
3	0.527	0.025	0.698	0.009	0.534	0.022
4	0.544	0.018	0.719	0.010	0.555	0.016
5	0.536	0.016	0.707	0.017	0.495	0.016
6	0.512	0.021	0.693	0.016	0.524	0.018
7	0.486	0.012	0.676	0.016	0.495	0.016

TABLE II
OVERALL MEAN AND STANDARD DEVIATION OF THE CROSS FOLD VALIDATION ACCURACY PERFORMED USING SCAP FOR EACH VALUE OF n, AFTER EACH PREPROCESSOR APPLIED.

Values for n were searched between 2 and 7 inclusive and at this point the scores achieved were progressively lowering. Table I show the overall mean and standard deviation of the cross fold validation performed using SCAP for each value of n. This highest mean is for $n = 4$, and is significantly different than $n = 5$ (difference of 0.01, p-value of 0.267) but not significantly different than $n = 3$ (difference of 0.021 and 0.021 as the p-value[7]). Overall, this accuracy of 0.729 is significantly higher than the chance rate of 0.02 and a high benchmark for future experiments.

B. Structural Clues

The results of the SCAP method after applying each of the three preprocessors (**At**, **Hash** and **Both**) are presented in table II. The results after applying **Both** preprocessors were on average 27% less accurate than their corresponding 'raw' results (without applying the preprocessor). For the **At** and **Hash** preprocessors the reduction was 26% and 1% respectively. This indicates that the tags applied to tweets do not contain much authorship information, while the replies contain quite a bit of information which is likely due to other users that an author frequently converses with. The network of other users that a given user converses with are therefore important clues for determining the authorship of tweets and removing these clues poses a significant decline in the accuracy of the SCAP methodology. Even without the network of people a user converses with, an accuracy of over

[6]The dataset is primarily English due to the use of English function words in the search, however there are other languages present in the dataset.

[7]A verified coincidence in the results.

Method	$\mu(S)$	$\sigma(S)$	p
Raw 2	-0.259	0.269	
Raw 3	0.682	0.293	0.000
Raw 4	0.850	0.293	0.005
Raw 5	0.893	0.293	0.461
Raw 6	0.910	0.292	0.779
Raw 7	0.918	0.291	0.890
At 2	-0.401	0.233	
At 3	0.652	0.289	0.000
At 4	0.850	0.293	0.001
At 5	0.897	0.294	0.422
At 6	0.914	0.293	0.770
At 7	0.922	0.292	0.892
Hash 2	-0.259	0.264	
Hash 3	0.685	0.289	0.000
Hash 4	0.851	0.292	0.005
Hash 5	0.894	0.293	0.463
Hash 6	0.911	0.292	0.782
Hash 7	0.918	0.291	0.894
Both 2	-0.418	0.231	
Both 3	0.653	0.285	0.000
Both 4	0.851	0.293	0.001
Both 5	0.898	0.294	0.421
Both 6	0.915	0.293	0.771
Both 7	0.923	0.292	0.896

TABLE III
ANALYSIS OF THE SIMPLIFIED PROFILES FROM SCAP, WHERE S IS THE MEAN OF THE SILHOUETTE COEFFICIENT FOR EACH USER'S PROFILE AND THE p-VALUE LISTED IS FOR THE TEST AGAINST THE NULL HYPOTHESIS THAT k CLUSTERS IS NOT SIGNIFICANTLY DIFFERENT FROM $k-1$ CLUSTERS.

Method	μ	σ	Increase	p-value
Raw 2	0.547	0.016	0.016	0.027
Raw 3	0.716	0.014	0.008	0.182
Raw 4	0.731	0.019	0.002	0.854
Raw 5	0.724	0.013	0.005	0.411
Raw 6	0.707	0.009	0.001	0.888
Raw 7	0.681	0.015	-0.001	0.930
At 2	0.387	0.013	0.016	0.022
At 3	0.547	0.015	0.013	0.159
At 4	0.565	0.016	0.010	0.195
At 5	0.549	0.025	0.001	0.920
At 6	0.527	0.026	0.003	0.740
At 7	0.499	0.014	0.004	0.602
Hash 2	0.541	0.019	0.021	0.015
Hash 3	0.709	0.010	0.011	0.015
Hash 4	0.719	0.014	0.000	0.984
Hash 5	0.715	0.018	0.008	0.354
Hash 6	0.699	0.024	0.006	0.512
Hash 7	0.671	0.018	-0.005	0.533
Both 2	0.379	0.010	0.022	0.008
Both 3	0.535	0.017	0.008	0.403
Both 4	0.554	0.015	0.009	0.217
Both 5	0.542	0.015	0.006	0.392
Both 6	0.517	0.022	0.005	0.615
Both 7	0.485	0.023	-0.001	0.865

TABLE IV
ACCURACY FOR THE DATASET AFTER SUB-PROFILING APPLIED. $d\mu$ IS THE INCREASE OR DECREASE FROM FULL PROFILING AND THE p-VALUE IS THE SIGNIFICANCE OF THIS DIFFERENCE USING A TWO TAILED t-TEST.

0.555 was still achieved, significantly higher than chance rates for the 50 users in each test.

C. Sub-profiles

The results for the benchmark experiments showed a good deal of inter-author separation. Analysis of the twitter users with the silhouette coefficient [30] is given in table III and indicates that the ratio between authors is similar to the ratio within a given author's profile. This indicates that the profiles, while distinct enough to achieve the high results found in the benchmark experiments, still have some ambiguity in the boundaries between one user and another. Further to that, it shows that having values of n too high overfits the data, as higher n values lead to more distinct clusterings but not to higher accuracies as tested in subsections I and II. For this reason, it is important not to use the silhouette coefficients alone to justify the selection of a value for k, rather to test that the increase is significant over the smaller values.

The results after applying sub-profiling are given in table IV. Two clusters were chosen for each user using the above procedure There is a small but non-significant increase for most methods and values of n. Upon observation, there are two reasons that could cause this. The first is that each of the sub-profiles is generated using a lesser number of tweets than the full profile, which could be a factor in the overall accuracy. Secondly, the silhouette coefficients listed in table III are high, suggesting that sub-profiling is separating already dense author profiles.

D. Restricting the number of tweets

To answer the final research question, the number of tweets per author was limited to quantity between 20 and 200 inclusive, in steps of 20 tweets. The results of the accuracy, after applying SCAP on the tweets without any preprocessing, are given in table V. It can be shown that the accuracy increases with an increasing number of tweets, with the exception of 120 tweets per user. This increase over 140 tweets is not significant and also marks the lowest values where this is not the case. For less than 120 tweets, adding another 20 tweets increases the accuracy a significant amount, using a two tailed t-test against the null hypothesis that k tweets is not significantly different in accuracy than $k-20$ tweets.

V. CONCLUSIONS AND CONTRIBUTIONS

Four questions were posed for this research and the list of experiments has given answers to each along with a number of contributions. These are summarised below.

Firstly, the SCAP method is very accurate at determining the author for a given tweet. The accuracy of over 70% for n values of $3, 4, 5$ and 6 is significantly higher than the chance rate of 2% for the 50 authors in the sample used. This indicates that authorship of twitter messages is indeed possible at a much higher than chance rate, which is the first contribution of this work.

Secondly, up to 27% of this accuracy is lost when removing information about the users that a given author converses with through @**replies**. While **#tags** contain some information, it appears that there is a lot of information contained specifically in the @**replies** for tweets. The observation that the network

	N-value							
Restriction	2	3	4	5	6	7	Mean	p-value
20	0.509	0.564	0.602	0.599	0.603	0.589	0.578	NA
40	0.544	0.617	0.636	0.644	0.634	0.622	0.616	0.000
60	0.551	0.659	0.668	0.659	0.652	0.634	0.637	0.012
80	0.554	0.680	0.692	0.692	0.669	0.657	0.657	0.003
100	0.551	0.694	0.711	0.711	0.688	0.673	0.671	0.012
120	0.553	0.707	0.722	0.713	0.697	0.677	0.678	0.014
140	0.539	0.705	0.720	0.715	0.700	0.675	0.676	0.370
160	0.536	0.704	0.722	0.718	0.699	0.677	0.676	0.724
180	0.534	0.702	0.723	0.719	0.699	0.680	0.676	0.823
200	0.531	0.708	0.729	0.719	0.706	0.682	0.679	0.151

TABLE V
MEAN ACCURACY AFTER CROSS FOLD VALIDATION OF THE RAW TWEET DATASET USING SCAP. p-VALUE IS THE PROBABILITY FROM A t-TEST THAT THE RESULTS OBTAINED FROM HAVING k TWEETS IS DIFFERENT TO HAVING $(k-20)$ TWEETS.

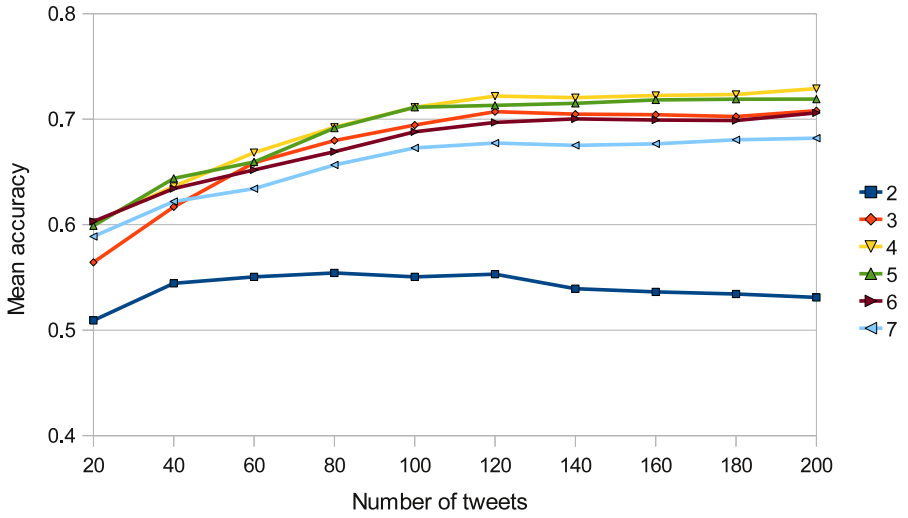

Fig. 1. Graph of results shown in table V with the number of tweets as the x-axis and the mean accuracy of SCAP shown on the y-axis

of communication of a particular author is very important in determining authorship is the second contribution of this paper.

Thirdly, creating sub-profiles of each author was shown to give a small but insignificant increase in the accuracy of the methodology, presenting the third contribution of this paper. It is possible that other methods may perform better, especially given the simple nature of the splitting used in this research.

Finally, it was shown that approximately 120 tweets per user is an important threshold for determining authorship. After this, increases in the accuracy for an additional 20 tweets are not significant. While 'more tweets are better' in all cases, significant increases can be made by just adding a small number of tweets below this threshold. This threshold is the final contribution of this paper, giving a guideline to future studies in this area.

With the above contributions from this work, it has been shown conclusively that authorship is possible for twitter messages at significantly higher than chance rates. Importantly, there are a few areas in which this accuracy might be improved further. Some of these possibilities are discussed in the next subsection.

A. Future Work

It was shown in this research that including the full usernames of the other Twitter users that are **@replied** to has a higher accuracy than removing this information. This information was removed due to the motivation behind this research; in a cybercrime setting, it is safe to assume that a person trying to be anonymous would not converse with their normal circle of friends. However in other settings, or in a situation where the cybercriminal is part of a known network of users collectively communicating, this assumption may not be needed and instead this communication network information could be leveraged to compliment the n-gram profiling performed in this work. A combination of authorship analysis above with a network analysis on the other users that a given user converses with could show a drastic improvement in overall accuracy.

The authors hypothesise that the main reason for the high accuracy for the raw dataset is based in the sampling method

used for selecting authors. Given that the authors were randomly selected, there is a low probability that any two authors converse with the same groups of people or even any particular person. If one author converses with @**fred** often while another user converses with @**JANE**, then the first author will have a high frequency of the tri-gram (n-gram with $n = 3$) `@fr` while the second author will have a high frequency of `@JA`. A future study in this area could perform a 'crawl-based' collection approach, where a single user is selected, then their communication network is collected, continuing outwards. This sampling technique could pose a more difficult problem, as there would be more authors conversing with @**fred**, reducing the impact of the `@fr` trigram.

Another possible avenue for future work is the application of these techniques into other short messages, such as IRC logs, instant messaging, blog comments and Facebook status updates. All of these areas have issues with types of cybercrime, such as spam, defamation and harassment.

Finally, the sub-profiling method shown in this work was very simple, as a profile was split into exactly two sub-profiles using the k-means algorithm. An improvement method of splitting the profile could generate better sub-profiling which could again lead to a higher classification accuracy. Together with the other suggested improvements, there is significant scope for more research in this area.

ACKNOWLEDGEMENT

This research was conducted at the Internet Commerce Security Laboratory and was funded by the State Government of Victoria, IBM, Westpac, the Australian Federal Police and the University of Ballarat. More information can be found at `http://www.icsl.com.au`

REFERENCES

[1] Web security report for 2009. Technical report, BlueCaot, 2010.

[2] A. Abbasi and H. Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.

[3] A. Abbasi and H. Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2):1–29, 2008.

[4] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123, 2009.

[5] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. pages 1027–1035, 2007.

[6] R. Beasley. Short Message Service (SMS) Texting Symbols: A Functional Analysis of 10,000 Cellular Phone Text Messages. *Reading*, 9(2), 2009.

[7] W. R. Bennett. *Scientific and engineering problem-solving with the computer*. In [16], 1976.

[8] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. *Hawaii International Conference on System Sciences*, 0:1–10, 2010.

[9] C. E. Chaski. Whos at the keyboard? authorship attribution in digital evidence investigations. *International Journal of Digital Evidence*, 4(1):1–13, 2005.

[10] S. Department of Innovation, Industry and Research. National Research Priorities (NRPS) Fact Sheet, 2009. "`http://www.innovation.gov.au/Section/AboutDIISR/FactSheets/Pages/NationalResearchPrioritiesFactSheet.aspx`".

[11] R. Forsyth and D. Holmes. Feature-finding for test classification. *Literary and Linguistic Computing*, 11(4):163, 1996.

[12] G. Frantzeskou, E. Stamatatos, S. Gritzalis, and C. E. Chaski. Identifying authorship by byte-level n-grams: The source code author profile (scap) method. *Int. Journal of Digital Evidence*, 6(1), 2007.

[13] T. Holz, C. Gorecki, K. Rieck, and F. Freiling. Measuring and detecting fast-flux service networks. In *Network & Distributed System Security Symposium*, 2008.

[14] F. Iqbal, R. Hadjidj, B. C. M. Fung, and M. Debbabi. A novel approach of mining write-prints for authorship attribution in e-mail forensics. *digital investigation*, 5:42–51, 2008.

[15] P. Juola, J. Sofko, and P. Brennan. A prototype for authorship attribution studies. *Lit Linguist Computing*, 21(2):169–178, 2006.

[16] V. Keselj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. In *Proc. Pacific Association for Computational Linguistics*, 2003.

[17] R. Kilgour, A. Gray, P. Sallis, and S. MacDonell. A fuzzy logic approach to computer software source code authorship analysis. *International Conference on Neural Information Processing and Intelligent Information Systems*, pages 865–868, 1997.

[18] M. Koppel, J. Schler, and S. Argamon. Computational methods in authorship attribution. *J. Am. Soc. Inf. Sci. Technol.*, 60(1):9–26, 2009.

[19] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can. Chat mining: Predicting user and message attributes in computer-mediated communication. *Information Processing & Management*, 44(4):1448–1466, 2008.

[20] R. Layton, S. Brown, and P. Watters. Using differencing to increase distinctiveness for phishing website clustering. *Proceedings of the Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing*, 2009.

[21] R. Layton and P. Watters. Determining provenance in phishing websites using automated conceptual analysis. In *eCrime Researchers Summit 2009*, Tacoma, WA, USA, 10 2009.

[22] R. Layton, P. Watters, and R. Dazeley. Unsupervised authorship analysis. *submitted*, 2010.

[23] J. Li, R. Zheng, and H. Chen. From fingerprint to writeprint. *Commun. ACM*, 49(4):76–82, 2006.

[24] S. H. Marjuni, R. Mahmod, A. A. A. Ghani, A. B. M. Zain, and A. Mustapha. Lexical criminal identification for chatting corpus. *Computer Science and Information Technology, International Conference on*, 0:360–364, 2009.

[25] T. Mendenhall. The characteristic curves of composition. *Science*, (214S):237, 1887.

[26] H. Mohtasseb, U. Lincoln, and A. Ahmed. Mining Online Diaries for Blogger Identification. *Proceedings of the World Congress on Engineering*, 2009.

[27] T. Moore and R. Clayton. Examining the impact of website take-down on phishing. In *eCrime '07: Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, pages 1–13, New York, NY, USA, 2007. ACM.

[28] F. Mosteller and D. L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[29] A. Orebaugh and J. Allnutt. Classification of Instant Messaging Communications for Forensics Analysis. *The International Journal of Forensic Computer Science*, 1:22–28, 2009.

[30] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20(1):53–65, 1987.

[31] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Automatic authorship attribution. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 158–164, Morristown, NJ, USA, 1999. Association for Computational Linguistics.

[32] M. Tearle, K. Taylor, and H. Demuth. An algorithm for automated authorship attribution using neural networks. *Lit Linguist Computing*, 23(4):425–442, 2008.

[33] R. Thomas and J. Martin. The underground economy: priceless. *The USENIX Magazine, December*, 00:2006–12, 2006.

[34] K. Weil. Measuring tweets, Feb. 2010. "`http://blog.twitter.com/2010/02/measuring-tweets.html`".

[35] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3):378–393, 2005.