USENIX Association

Proceedings of the 9th USENIX Security Symposium

Denver, Colorado, USA
August 14–17, 2000

Can Pseudonymity Really Guarantee Privacy?

Josyula R. Rao Pankaj Rohatgi
I.B.M. T. J. Watson Research Center
P. O. Box 704
Yorktown Heights, NY 10598
{jrrao, rohatgi}@watson.ibm.com

Abstract

One of the core challenges facing the Internet today is the problem of ensuring privacy for its users. It is believed that mechanisms such as anonymity and pseudonymity are essential building blocks in formulating solutions to address these challenges and considerable effort has been devoted towards realizing these primitives in practice. The focus of this effort, however, has mostly been on hiding explicit identity information (such as source addresses) by employing a combination of anonymizing proxies, cryptographic techniques to distribute trust among them and traffic shaping techniques to defeat traffic analysis. We claim that such approaches ignore a significant amount of identifying information about the source that leaks from the contents of web traffic itself. In this paper, we demonstrate the significance and value of such information by showing how techniques from linguistics and stylometry can use this information to compromise pseudonymity in several important settings. We discuss the severity of this problem and suggest possible countermeasures.

1 Introduction

1.1 The Problem

Privacy in the computer age is the right of individuals to protect their ability to selectively reveal information about themselves so as to negotiate social relationships most advantageous to them. The complete absence of privacy protection in the basic Web infrastructure coupled with the increased migration of human activity from real–word interactions to web based interactions poses a grave threat to the privacy of the entire human population. Increasingly sophisticated databases of profiles of individuals based on their web interactions are being built and exploited. Clearly, assuring privacy has become one of the fundamental challenges facing the Internet today.

1.2 Background

A number of tools and techniques have been proposed and deployed to address the privacy concerns of Internet users. To place these in perspective, it is important to understand two primitives, *anonymity* and *pseudonymity*, that are considered to be the core building blocks of most privacy solutions.

Anonymity refers to the ability of an individual to perform a *single* interaction with another entity (or set of entities), without leaking any information about his/her identity. While anonymity is very effective at protecting the identity of an individual, its ephemeral nature makes it ill–suited for most kinds of web interactions.

This shortcoming is addressed by the concept of pseudonymity, which enables an individual to participate in a series of web interactions, all linkable to a single identifier (also known as a *pseudonym*), with the guarantee that the pseudonym cannot be linked back to the individual's identity. The persistence of pseudonyms permits the establishment of long term web–relationships. The ability of an individual to choose different pseudonyms for different activities enables an individual to further protect his/her privacy by partitioning his/her interactions into unlinkable activities.

Broadly speaking, technologies for realizing privacy fall into four different categories [5].

- Anonymizing agents ensure that packets from a user cannot be linked to his/her IP address. A simple anonymizing agent is just a single trusted third party (e.g. the Anonymizer [1]) that decrypts and redirects requests from users. Typically, the third party disassociates the initiator's identity from the recipient's identity by replacing the user identifying headers with its own headers. This approach is deficient in that it presents a single point of failure and is vulnerable to traffic analysis. Therefore, more sophisticated schemes use cryptographic techniques to distribute trust in a *chain* of third parties and employ traffic shaping techniques to defeat traffic analysis. This ensures that anonymity/pseudonymity is not compromised even in the presence of a few compromised third parties.

  Examples of such technologies include Anonymizer[1], Crowds [17], Onion Routing [8] and the IP Wormhole technology used in the Freedom network. [9].

- Application level filters to eliminate obvious identity information from an individual's web traffic, such as name, e–mail address and affiliation etc. Examples include the Freedom word scanner [10].

- Pseudonym agents which manage pseudonyms to develop persistent relationships. Examples include Lucent Personalized Web Assistant (LPWA) [7], P3P [16] and Freedom Network [9].

- Negotiation Agents/Trust Engines which negotiate on a user's behalf and determine when a user's privacy policies are satisfied. Examples include software tools for cookie management and P3P [16].

In practice, a service providing privacy would incorporate elements of some or all of these technologies. It has been the general belief that appropriate use of these tools and technologies should be adequate to counter threats to privacy, provided that the cryptography used is sufficiently strong, sufficient numbers of trusted anonymizing parties remain uncompromised and adequate traffic analysis countermeasures are in place. In fact, considerable energy has been devoted to perfecting and commercializing these technologies.

1.3 Contributions

The goal of this paper is to argue that the current focus on hiding explicit identity information is not sufficient to guarantee privacy. There is still a considerable wealth of identifying information in the body of the communication messages, which to the best of our knowledge, is only minimally altered by the existing tools and privacy infrastructure, before it is delivered to the recipient. At best, what may be removed is the obvious identity information such as header blocks, signature blocks, addresses, names, etc. However, the body of each communication leaks identifying information in many subtle ways. Although these leaks may be subtle, with a sufficient amount of communication, it becomes feasible to distill useful identifying information. This is especially true for pseudonymous communications where the recipient is a newsgroup, a chat–room, a mailing list or any server that stores the communication in an accessible manner. This information is usually publically archived and available for analysis for all eternity.

We can classify the information available in such pseudonymous communications into two categories:

- *Syntactic*: This term is intended to encompass information dealing with the components of the message and their structural layout. This could include information such as vocabulary, sizes of sentences and paragraphs, use of punctuation, frequency of blank lines, common misspellings, etc.

- *Semantic*: This term is intended to encompass the concepts used in the communication and the language used to express these concepts.

There is field of linguistics, known as stylometry, which uses both these categories of information to ascribe identity or authorship to literary works. This field has had remarkable success in authenticating the authorship of such diverse works such as the disputed *Federalist* papers [14, 15, 2], the Junius Letters [6], Greek prose [13], Shakespeare's works[3, 12, 18]. More recently, stylometry was used by Prof. Donald Foster at Vassar College to attribute the authorship of the novel "Primary Colors" to Joe Klein, an attribution subsequently confirmed by handwriting analysis and the author's own admission.

We believe (and demonstrate) that recent advances in stylometry pose a significant threat to privacy that merits the serious and immediate attention of the privacy community. For instance, using stylometry, one can link the multiple pseudonyms of a person and if one such pseudonym happens to be his/her identity, then the protection afforded by the other pseudonyms is compromised.

For instance, given a pseudonym for an identity and a moderately sized list (say around one hundred entries) of other pseudonyms, at least one of which corresponds to the same identity, we show that it is easy to perform semantic stylometric tests to correctly link the pseudonyms with good probability, provided enough data (a few thousand words per pseudonym) is available. In cases where, the list is very long, perhaps with a million or more entries, the same technique can be adapted to cut down the possibilities by a large factor.

In fact, the two parameters of the technique are (a) the data requirements per pseudonym and (b) the discriminating power of the technique, that is, the factor by which the number of possibilities is reduced by the technique. We believe that the discriminating power improves with increased data requirements. Our experience with this technique, leads us to conclude that the values of these parameters are such that the privacy of users of anonymous short e–mail messages or short newsgroup postings is not yet at risk but authors (anonymous or pseudonymous) of at least few thousand words are susceptible to our techniques and analysis.

Note that these results and estimates are based on a technique borrowed from stylometry that does not make any use of the wealth of other types of identifying information (such as misspellings, structural layout, etc) that is present in most types of web traffic. This is because, linguists have focussed mostly on proof-read and typeset documents such as books and journals. We believe that approaches that also incorporate this additional information could be even more effective.

While it may seem that the leakage of identifying information from the syntax of messages can be easily remedied by an application specific agent run on the user's behalf, the problem with the leakage from semantics is much more insidious and difficult to eliminate satisfactorily. Particularly egregious is the fact that this line of attack cannot be fixed without changing the meaning of the message. Perhaps with collaboration and close cooperation between linguistics, statisticians and the privacy community, it may be possible to develop a much more intelligent and usable agent to minimize semantic leakage without degrading the message quality and meaning.

1.4 Plan of Paper

In section 2, we begin describing in detail the types of identifying information available in the specific setting of newsgroup postings. In the following section, we describe a specific stylometric technique, called principal component analysis of function words which can be applied for attacking pseudonymity. In section 4, we describe our experiments with this technique on data collected from newsgroup archives and mailing lists and the results obtained. In section 5, we discuss several other promising avenues to further improve the basic attack by incorporating other types of leaked information. We conclude by presenting a discussion on possible countermeasures.

2 Exposures in Newsgroup Postings

Newsgroup postings are frequently archived and are freely available over the Internet. Typically, each valid (RFC-1036 compliant) posting, consists of "RFC-822 style" headers (including the *From:*, *To:*, and *Subject:* fields), a body and (optionally) a signature. Pseudonymous postings would at best hide all identifying information (except the pseudonym) in the header fields and would be unlikely to contain any obvious identifying information in the signature or the body.

As mentioned earlier, in spite of identity hiding techniques, there still are two types of identifying information in the body. The first category, termed as *syntactic information* includes:

- *Length of sentences and paragraphs.*
- *Number of words and paragraphs.*
- *Paragraph style*: This includes the indentation style and the spacing between paragraphs.
- *Words, Phrases and their frequency*: This includes the vocabulary of the user.

- *Misspelled words and their frequency.*
- *Capitalized words and their frequency.*
- *Hyphenated words and their frequency.*
- *Punctuation usage*: including usage of hyphens: - vs. – vs. —.
- *Choice of acceptable uses of words or phrases*: That is, choices taken when there is more than one alternative such as, British vs. American spelling and "I'm" vs. "I am".

The second category termed as *semantic information*, captures the essence of concepts being conveyed and the manner in which language is being used to convey these concepts.

- *Function word usage and their frequency of occurrence*: Function words are specific English words that are commonly used to convey ideas and their usage is believed to be independent of the ideas being conveyed. The frequency of usage of these words has been used with remarkable success in authorship attribution studies [14, 4]. For example, the function words used in our experiments are given in Appendix A.
- *Concepts, topics and specialized areas of interest*: This encompasses the areas of expertise, the knowledge, the presentation style and the reasoning process of the author. This is the realm of traditional literature studies and it is unclear how computers can help in this arena.

A *profile* of a user consists of the some or all of the above syntactic and semantic information. An example profile of a user of the *sci.crypt* newsgroup who posted 554 articles over a period of 5 years is given in Appendix B. Intuitively, each of these attributes appears to be dependent on an author's style of writing and communication. One can analyze each of these attributes and study their effectiveness in determining the identity behind a pseudonym. Such an analysis falls within the purview of the field of stylometry. However, text collected from the Internet setting contains several additional attributes (such as misspellings, lines between paragraphs etc) typically not present in the type of text studied by classical stylometry.

3 Analysis of the Usage of Function Words

The current state of art in stylometry has achieved considerable success by using the *principal component analysis* technique from statistics on function word usage [4]. In particular, it has succeeded in attributing authorship in several disputed works.

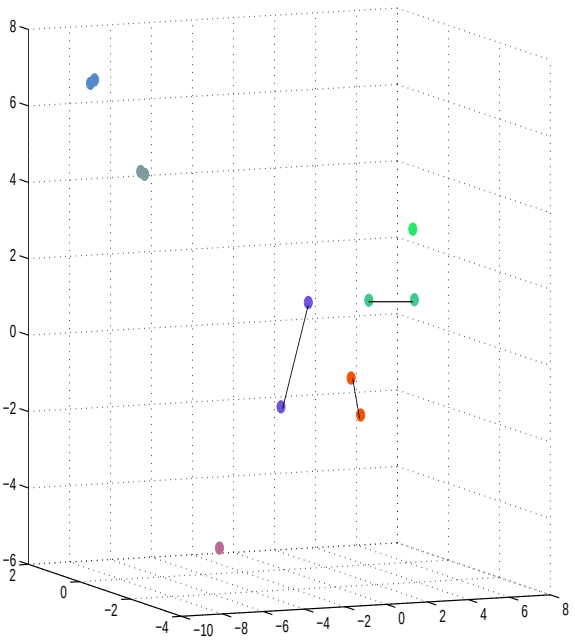
Our experiments show that this technique would be extremely effective in breaking pseudonymity in newsgroup postings. What was surprising to us was that even though the linguists have mostly used this technique to decide between two (or few) disputed authors, the technique is powerful enough to decompose a moderately sized (say of size 100) collection of pseudonyms into clusters where each cluster is highly likely to correspond to postings from a single identity.

The starting point for this technique is the computation of *frequency counts* of function words for each pseudonym. As mentioned earlier, function words are specific English words that are commonly used to convey ideas and are independent of the ideas being conveyed. They were defined and used with great success in one of the most celebrated works of statistical authorship attribution: the attribution of the 12 disputed *Federalist* papers to Madison [14, 15].

The frequency counts are then divided by the total number of words, to get an estimate of the probability of usage of each of these function words by the pseudonym. In our experiment, we also computed the estimate of the probability that the pseudonym uses a word outside the list. Traditionally, these probability estimates were used directly in authorship attribution studies [14], however direct analysis of function word usage is complicated by the fact that, from a statistical view, usage of different function words is not independent and this dependence should be accounted for in the analysis.

One way to view these probability estimates is as a vector space over the reals, where the dimensionality of the vector space corresponds to the number of function words being analyzed. Each author can be identified by a *author–point* in this vector space, i.e., the i'th co-ordinate of the author-point is just the the probability estimate of the author using the i'th function word. For authorship attribution studies, a direct use of the author-points defined with respect

Figure 1: Plot of 12 author-points along the first 3 principal components



this co-ordinate system has problems since the observed distributions of author-points along different axes are not going to be independent, since there is dependence between the usage of different function words.

A major breakthrough in finding a better co-ordinate system for this problem was developed by Burrows [4], who applied the principal component analysis approach from statistics to the collection author–points themselves. The principal component analysis approach when applied to the given author–points yields a new set of orthogonal basis vectors (or axes) for the vector space, which have the property the distributions of author–points along the new axes are *statistically independent*. These axes are known as principal components and furthermore these principal components can ordered on the basis of decreasing discriminating capabilities, i.e., the variation within the author-points is maximum along the first axis (or component) and minimum along the last axis (or component). Thus, by transforming the author–points to a representation with respect to these new axes or principal components, we get a collection of transformed points for which we can use very simple measures to gauge similarity. Moreover, since most of the discriminating information or variation in the transformed author-point data set is in the first few components, similarity analysis can be further restricted to them. A very simple similarity measure could be the euclidean distance between the transformed author-points. Alternatively, we can treat these transformed points as vectors and define a similarity measure between two vectors as the cosine of the angle between them. This measure is fairly easy to compute since it is just the inner-product of the normalized form of the two vectors and this measure has been extensively used in text processing and information retrieval systems. Yet another similarity measure could be the correlation between the various co-ordinates of the transformed author points. In our experiments and analysis we chose the simplest similarity measure, i.e., euclidean distance, which yielded fairly good results. We also experimented with the other measures and the results were equally good.

The effectiveness of this approach is illustrated in Figure 1 which plots the representation of 12 author-points (the first 12 author points out of a collection of 126 authors in one of our experiments) along the top 3 principal components. These 12 points actually consisted of only 7 distinct authors with 5 of these authors writing under 2 distinct pseudonyms. Even with 3 components (or dimensions), it is easy to identify two of these multiple pseudonym authors since the corresponding author-points are in very close proximity. We have also marked the other 3 multiple pseudonym authors by connecting the author points of their pseudonyms by a line. At first glance, it appears that these pairs of author points are not sufficiently close to be immediately identified as pseudonyms. However, it should be noted that typically, the number of independent dimensions we consider is around 50 and such a high dimension space, the author-points corresponding same identity are likely to be close in most dimensions and author-points corresponding to two different identities are likely to be far apart in several dimensions and this difference in behavior can be easily distinguished by the similarity measures described above.

Technically, the principal component analysis approach begins by defining a $W \times A$ matrix M, where the rows correspond to each one of the W function words and the columns correspond to each of the A authors. A requirement is that $W < A$. Each (w, a) entry in M corresponds to the estimate of probability of usage of word w by author a. Note that an author–point, as defined above, corresponds to

a column of this matrix. The first step in the process is to normalize the matrix M by ensuring that each row has zero mean and unit variance. This can be done by subtracting the row average from each entry and dividing the result by the row's standard deviation. The reason for this normalization is to ensure that each dimension (or function word) is treated equally in the subsequent processing, since, a priori, we have no reason assign different weights to different function words. However, in more specialized scenarios, one may be able to obtain better results if all the dimensions are not equally weighed.

Next, compute the $W \times W$ co–variance matrix C as follows:

$$C = \frac{1}{A} M \times M^t$$

where M^t is the transpose matrix. Intuitively, the co–variance matrix is the "column–to–column" similarity matrix having high values c_{ij} if columns i and j are highly correlated. Note that since we had normalized the matrix M, C is also the correlation co-efficient matrix for the usage of the W function words by the A authors. In specialized scenarios M may not be normalized and the co-variance matrix C will not be the correlation matrix.

The next step is to compute the eigenvalues and the corresponding eigenvectors of the co–variance matrix C. One can easily verify that the distinct eigenvectors are orthogonal over the real space. In fact, these eigenvectors form the principal components for the given data set. Furthermore, the larger the eigenvalue, the larger the utility of the corresponding eigenvector in discriminating between the different author–points. This is because, the variance of author–points when projected along an eigenvector is proportional to the corresponding eigenvalue. The final step is to obtain the new representation of each of the author–points with respect to the principal components or eigenvectors. Note that it is enough to restrict the representation for each author–point to those eigenvectors that have significant variance and hence significant discriminatory information. This can be accomplished by pre–multiplying M by a matrix whose rows consist of only the significant eigenvectors.

Principal component analysis is a standard tool in several mathematical toolkits such as Matlab [11].

4 Experiments

To explore and illustrate our ideas, we began by collecting three types of easily accessible, commonly archived and singly authored documents, namely, the *sci.crypt* newsgroup, the IPSec mailing list and the RFC Database.

Each type of document has a standard author identifier field (e.g., *From:* header in e–mails and news articles). We observed that, in practice, many authors used somewhat different identifiers (such as slightly different names, different e–mail addresses etc) over time. As a first step, we used elementary heuristics to aggregate each of the slightly different identifiers (and the respective documents) as corresponding to an entity that we call a *persona*. This was done so that we could obtain a much larger collection of documents per persona, in order to get sufficient data for statistical analysis [1].

After extracting personas, we removed all identifying information (such as headers, signatures etc) from the documents. At this point, our database of documents and associated personas is no different from what we would have obtained if we had completely pseudonymized documents. Once we partitioned the document space on a persona basis, it is easy to simulate multiple pseudonyms per persona by sub–partitioning the document space of each persona.

In the following sub–sections, we discuss two experiments that we performed on this pool of documents and the results obtained. Out of respect for the privacy rights of the concerned personas we have chosen not to name them in this paper.

4.1 Clustering Newsgroup Persona

Our initial data set consisted of 53914 articles available from the USENET archives of *sci.crypt* (articles # 6871 to # 62874 at deja-news.com) covering the period from 5 Jan 1992 to 28 Feb 1997. There were over 10758 personas who posted articles over this period. As mentioned above, each article was cleaned by removing headers and sig-

[1] As will become evident later, even if we had not done so, our tools would have identified multiple identifiers as belonging to the same persona. In fact, our tools led us to discover multiple gaps in our persona aggregation heuristics.

natures. Further heuristics were used to remove quoted/referenced text and non–language words.

Suppose we had a set of newsgroup postings by multiple persons each of whom used a small set of different pseudonyms for their postings. For a technique to break pseudonymity, it should be effective at clustering pseudonyms based solely on the content of the postings. Clearly, it is reasonable to apply such a technique to pseudonyms with a minimum but reasonable number of articles/words, since even the best technique would fail on very little data.

Testing the effectiveness of the principal component analysis technique to break pseudonymity in newsgroup postings can be easily simulated with the data we have. We only keep personas which meet a minimum requirement on articles/words. We create multiple pseudonyms per persona (for instance, two pseudonyms for a persona) by partitioning the document set of those personas who have a large number of articles/words. We can then apply the principal components technique to the collection of postings corresponding to these pseudonyms to obtain a principal component representation of author–points corresponding to these pseudonyms. Clustering of pseudonyms could based on the euclidean distances between the principal component representations of author–points. For example, in the two way split, two pseudonyms are assumed to correspond to the same person, if both are mutually closest to one another in the principal component representation. For more than two pseudonyms per person more advanced clustering metrics can be used.

- *Experiment I*: As a first experiment, we considered only those personas who had posted more than 50 articles and more that 5000 words. There were 117 such personas accounting for 19415 articles. Of these, we could chronologically split postings from 68 personas into two pseudonyms, the "earlier postings" pseudonym the "later postings" pseudonym, while still maintaining the article/word requirements[2]. This resulted in 185 pseudonyms, in which 68 personas had two pseudonyms and the remaining 49 had a single pseudonyms.

  Results: The principal component analysis technique was able to correctly cluster 40 out of 68 multiple pseudonym personas and reported 2 false positives, i.e., 2 pairs wrongly classified as a single identity. This gives a success rate of 58.8% and a false positive rate [3] of 2.2%.

- *Experiment II*: As a second experiment we considered personas with more than 50 articles and more that 10000 words. There were 84 such personas accounting for 17001 articles. Of these, we were able to chronologically split 42 personas into two pseudonyms. This resulted in 126 pseudonyms, in which 42 personas had two pseudonyms and the remaining 42 had a single pseudonyms.

  Results: The principal component analysis technique was able to correctly cluster 34 out of 42 multiple pseudonym personas and reported 1 false positive. This gives a success rate of 80.9% and a false positive rate of 1.6%.

We also experimented with varying the minimum requirements on the number of words and found that at around 6500 words the results are almost as good as for 10000 words. It appears that this is a reasonable threshold above which there is significant leakage of identity information.

In many false negative cases, the closest neighbor relationship held for one of the points and not for the other, and in most false negative cases the pairs were amongst the top few closest points to each other. This suggests that a slightly relaxed clustering criterion in conjunction with other, perhaps, syntactic criteria may yield better results.

Also note that the pair–clustering criterion we have chosen, i.e., mutually closest points in a high dimensional space, will usually find significant number of pairs in most cases, e.g., even when points are randomly chosen. Therefore, if there are very few multiple pseudonyms in the data set, we would get a significant false positive rate. In such cases a different approach would be better. For example, one could consider only the suspected multiple pseudonyms and look for author points closest to them. Another approach would be to do further processing on the clustered pairs, by looking at the nature of the proximity of the pair (e.g., the value of the distance vs. average distance when one of the identities in the pair is arbitrarily split, distribution

[2] We had several options on how the set of articles of a persona could be split. For example, we could have split the articles randomly. However, we chose to split chronologically since it is well known that an author's style evolves over time, and such a split should be slightly harder to reconcile

[3] Computed as the number of pseudonyms wrongly paired divided by the total number of pseudonyms.

of author points in the neighborhood of the pair etc) or perhaps by employing other syntactic and semantic criteria to weed out the false positives.

4.2 Linking E–mail Persona to Newsgroup Persona

One can argue that our results on breaking pseudonymity in newsgroup are not applicable, since people may not use more than one pseudonym for a single activity such as posting to newsgroups. For example, the FAQ from ZeroKnowledge.com states that one can use a different "nym" for each different activity. To show the generality of our approach, we choose a second domain, namely postings to mailing list and show how a persona in this domain can be linked to the corresponding newsgroup persona.

The first set of experiments described above, enabled us to build statistical profiles of frequent sci.crypt posters. To test whether a mailing list persona could be linked to a newsgroup persona, we looked for examples of personas who have posted material to both domains. In particular, we examined the list of frequent contributors to the IPSec mailing list archives from Aug 14, 1992 to Dec 31 1996 and found that there were 7 contributors who had posted sufficient material in both forums. As before, we preprocessed the IPSec archive to extract the e–mails of these 7 personas.

For each of these personas, we performed a principal component analysis test (as described earlier) on the frequent (> 10000 words) sci.crypt personas and the e–mail personas. We then looked at the proximity of the given e–mail persona to the sci.crypt personas. In 5 of the 7 cases, the corresponding newsgroup persona was the closest point to the e–mail persona. In one case, the corresponding newsgroup persona was second closest and in the remaining case, the corresponding personas were far apart.

In a sense, these results are not particularly surprising and merely reinforce the analysis for the newsgroup postings. It is also intuitive to expect this since e–mail postings are very similar in style to newsgroup postings. One could expect similar results to hold for chat–room transcripts.

4.3 A Comment on Styles: RFCs vs. Newsgroup Postings

We tried to extend the results of the previous section to other domains by comparing RFC personas with newsgroup personas. Preliminary experiments indicate that the linking of the personas in this case is quite poor. We suspect that the reason for this is that the writing style for newsgroup posting/e–mail is quite informal and direct, whereas RFC writing style is much more formal and indirect. We speculate that it may be feasible to find a transformation which would map the style of RFC personas to Newsgroup personas.

5 Future Directions

During the course of this work, we found several intriguing leads and potential directions for future research. Most promising amongst these is the wealth of identifying information that is available in the syntactic aspects of internet based communications. While we have already enumerated the various useful syntactic features before, it is important to note that we did not use any of this information in the analysis presented in the previous section. A number of these features have also not been studied in classical authorship studies since the latter have dealt mostly with proof-read texts that have appeared in well-scrutinized and standard formats traditionally followed by publishers. For example identifying features like misspellings, types of formatting (spacing, paragraph indentation style etc) and other stylistic idiosyncrasies such as hyphenation style, capitalization style, etc, are characteristic of internet based communications but rare in published literature.

We now detail some of our initial findings in studying the discriminating potential of these syntactic features of internet based communications. A major discriminating feature appears to be the set of words misspelled by a persona. We observed that a majority of the documents we dealt with contained misspellings. Broadly speaking, there are three types of misspellings: The most important type for our purposes are words which are consistently misspelled by an author. We have observed that such a list of words can be used as a good discriminator for the author. In fact, very often this list could consist

of a few or even a single word. For example, three prominent sci.crypt posters could be easily identified by their use of the words "naiive", "cought & propietary" and "consistantly & insistance" respectively.

Closely related to this category of misspelling is the usage of British vs. American English spelling of words. Use of words such "authorise", "behaviour", "centre", "colour", for example, would clearly place the author in the British English camp.

The second type of misspelling are those words of which the author is somewhat unsure and are misspelled with some probability. These probabilities could potentially be used for discrimination purposes. The third type of misspelling are typos. These may be indicative of the author's typing style. More work needs to be done to establish the significance of these two categories.

Our preliminary analysis also indicates that formatting features such as spacing between paragraphs, paragraph indentation style, capitalized words, hyphenated words and words with an apostrophe, also have good discrimination potential.

In passing, we would like to mention the role of other syntactic features that have been studied in by classical stylometry such as vocabulary, choice of specific words to convey concepts and punctuation style in authorship attribution.

We speculate that a hybrid approach which uses the principal component analysis approach on function words and formatting features, coupled with misspelling lists and the classical stylometry measures mentioned above would be extremely effective in attacking pseudonymity.

6 Possible Countermeasures

In this paper, we have described two classes of information leakage: syntactic and semantic, both of which can be used for attacking pseudonymity.

Countermeasures should try to address both classes of leakage. For each class and for each type of leakage within the class, a countermeasure should either eliminate the leakage altogether or substantially reduce the amount of information leaked so that an adversary is unlikely to have access to the minimum amount of text needed to apply the attacks. On the other hand, drastic countermeasures run the risk of badly altering the meaning of the document thereby requiring extensive manual intervention and rendering them unusable.

It should be easy to design minimally disruptive countermeasures to tackle most syntactic leakage. However, fixing the problem of vocabulary requires more investigation. For instance a thesaurus tool could prompt the user to consider alternatives while composing messages, thereby reducing variations in vocabulary.

However, countermeasures to address semantic leakage appear to be hard to design. One drastic (and somewhat facetious) approach is to use natural language translation systems to translate a document from English to another language and back. While this may destroy the function word frequencies in the original document, it would considerably change the meaning of the message, given the primitive state of translation systems. Even this would not destroy information about the meta–level concepts that an author uses in his documents.

Apart from these automated countermeasures, another powerful countermeasure (as suggested to us by an anonymous referee) would be to educate users of pseudonymity services about the types of exposures and attacks discussed in this paper. A user community which is aware of these types of exposures may be the most successful way to minimize the effectiveness of these attacks.

7 Acknowledgements

The authors would like to thank Roy Byrd, Suresh Chari, Pau-Chen Cheng, Charanjit Jutla, Aaron Kerschenbaum and Charles Palmer for several useful insights and helpful discussions. We would also like to thank anonymous referees for their comments and suggestions on improving this paper.

References

[1] The Anonymizer Service, `www.anonymizer.com`.

[2] Robert A. Bosch and Jason A. Smith, "Separating Hyperplanes and the Authorship of Disputed Federalist Papers", The American Mathematical Monthly, Volume 105, Number 7, August–September 1998, pages 601–608.

[3] Barron Brainerd, "The Computer in Statistical Studies of William Shakespeare", Computer Studies in the Humanities and Verbal Behavior, Volume 4, Number 1, pages 9–15, 1973.

[4] J.F. Burrows, "Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style", Literary and Linguistic Computing, Volume 2, pages 61–70, 1987.

[5] Lorrie Faith Cranor, "Internet Privacy", Communications of the ACM, Volume 42, Number 2, February 1999, pages 29–31.

[6] Alvar Ellegard, "A Statistical Method for Determining Authorship: the Junius Letters 1769–1772", Gothenburg Studies in English No. 13, Acta Universitatis Gothenburgensis, Elanders Boktryckeri Aktiebolag, Goteborg.

[7] Evan Gabber, Phillip B. Gibbons, David M. Kristol, Yossi Matias and Alain Mayer, "Consistent, Yet Anonymous, Web Access with LPWA", Communications of the ACM, Volume 42, Number 2, February 1999, pages 42–47.

[8] David Goldschlag, Michael Reed and Payl Syverson, "Onion Routing for Anonymous and Private Internet Connections", Communications of the ACM, Volume 42, Number 2, February 1999, pages 39–41.

[9] `www.freedom.net`.

[10] `www.freedom.net/info/freedompapers/Freedom-Architecture.pdf`.

[11] `www.mathworks.com`.

[12] T.C. Mendenhall, "A Mechanical Solution to a Literary Problem", Popular Science Monthly, Volume 60, Number 2, pages 97–105, 1901.

[13] A.Q. Morton, "The Authorship of Greek Prose", Journal of the Royal Statistical Society (A), Volume 128, pages 169–233, 1965.

[14] Frederick Mosteller and David L. Wallace, "Inference and Disputed Authorship: The Federalist", Addison–Wesley, Reading, MA, 1964.

[15] Frederick Mosteller and David L. Wallace, "Applied Bayesian and Classical Inference: The Case of The Federalist Papers", Springer Series in Statistics, Springer–Verlag, 1984.

[16] Joseph Reagle and Lorrie Faith Cranor, "The Platform for Privacy Preferences", Communications of the ACM, Volume 42, Number 2, February 1999, pages 48–55.

[17] Michael K. Reiter and Aviel D. Rubin, "Anonymous Web Transactions With Crowds", Communications of the ACM, Volume 42, Number 2, February 1999, pages 32–38.

[18] C.B. Williams, "Mendenhall's Studies of Word-Length Distribution in the Works of Shakespeare and Bacon", Biometrika, Volume 62, pages 207–212, 1975.

A Function Words

a	*able*	*about*	*above*
after	*again*	*against*	*all*
allow	*almost*	*also*	*although*
am	*among*	*an*	*and*
another	*any*	*apply*	*are*
as	*ask*	*at*	*away*
be	*because*	*been*	*being*
between	*but*	*by*	*can*
could	*did*	*do*	*does*
down	*either*	*enough*	*even*
every	*for*	*from*	*get*
give	*had*	*has*	*have*
he	*her*	*his*	*how*
i	*if*	*in*	*into*
is	*it*	*make*	*may*
me	*might*	*more*	*most*
must	*my*	*no*	*nor*
not	*now*	*of*	*often*
on	*one*	*only*	*or*
other	*our*	*same*	*say*
see	*shall*	*she*	*should*
so	*some*	*someone*	*still*
such	*sure*	*take*	*than*
that	*the*	*their*	*them*
then	*there*	*they*	*this*
to	*under*	*until*	*up*
upon	*use*	*want*	*was*
we	*were*	*what*	*when*
where	*which*	*who*	*why*
will	*with*	*would*	*yes*
you	*your*		

B Profile of a Persona

```
Persona: 9069
Number of Articles: 554
Total Number of Words: 91257

1. Blank Line Distribution:

1: 2418
2: 768
.
.
.
123: 1
359: 1

Mean: 1.913 Variance: 39.980

2. Word Count Distribution:

0: 3
2: 1
.
.
.
4372: 1

Mean: 164.724 Variance: 180411.828

3. Number of Paragraphs Distribution:

1: 4
2: 102
.
.
.
252: 1

Mean:7.863 Variance: 406.642

4. Paragraph Size Distribution:

1: 5
2: 253
.
.
.
606: 1
776: 1

Mean: 20.855 Variance: 749.314

5. Paragraph Indentation Style:

Does not use tabs to indent
paragraphs:

Used spaces to indent paragraphs:

Mean: 5.023 Variance: 29.718

6. Hyphenation styles:

 - (283)  -- (21)  --- (2)

7. 9069 uses the following words
   commonly:

a: 1981
```

about: 174
.
.
.
with: 500
you: 292

8. 9069 uses apostrophes in the
   following words:

Alice's, Bamford's, Bidzos', ..., He'd,
..., hoaxer's, ..., you're, you've.

9. 9069 likes to completely capitalize
   the following words:

ACCESS, ACLU, ACM, ..., YOU, YOUR, ZIP

10. 9069's vocabulary (including
    misspelled words):

ACCESS, ACLU, ACM, ..., Zero-Knowledge,
..., ability, able, ..., zeroize,
zeros.

11. 9069's misspelled words:

ableto, accessdata, ...,
cought, ..., yeild