Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity

MICHAEL BRENNAN, SADIA AFROZ, and RACHEL GREENSTADT, Drexel University

The use of stylometry, authorship recognition through purely linguistic means, has contributed to literary, historical, and criminal investigation breakthroughs. Existing stylometry research assumes that authors have not attempted to disguise their linguistic writing style. We challenge this basic assumption of existing stylometry methodologies and present a new area of research: adversarial stylometry. Adversaries have a devastating effect on the robustness of existing classification methods. Our work presents a framework for creating adversarial passages including *obfuscation*, where a subject attempts to hide her identity, and *imitation*, where a subject attempts to frame another subject by imitating his writing style, and *translation* where original passages are obfuscated with machine translation services. This research demonstrates that manual circumvention methods work very well while automated translation methods are not effective. The obfuscation method reduces the techniques' effectiveness to the level of random guessing and the imitation attempts succeed up to 67% of the time depending on the stylometry technique used. These results are more significant given the fact that experimental subjects were unfamiliar with stylometry, were not professional writers, and spent little time on the attacks. This article also contributes to the field by using human subjects to empirically validate the claim of high accuracy for four current techniques (without adversaries). We have also compiled and released two corpora of adversarial stylometry texts to promote research in this field with a total of 57 unique authors. We argue that this field is important to a multidisciplinary approach to privacy, security, and anonymity.

Categories and Subject Descriptors: I.5.4 [**Pattern Recognition**]: Applications; I.2.7 [**Natural Language Processing**]: Text analysis; K.4 [**Computers and Society**]: Public Policy Issues—*Privacy*

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Stylometry, adversarial stylometry, authorship recognition, text mining, machine learning, privacy, anonymity

ACM Reference Format:
Brennan, M., Afroz, S., and Greenstadt, R. 2011. Adversarial stylometry: Circumventing authoeship recogition to preserve privacy and anonymity. ACM Trans. Info. Syst. Sec. 15, 3, Article 12 (November 2012), 22 pages.
DOI = 10.1145/2382448.2382450 http://doi.acm.org/10.1145/2382448.2382450

1. INTRODUCTION

Stylometry is a form of authorship recognition that relies on the linguistic information found in a document. While stylometry existed before computers and artificial intelligence, the field is currently dominated by AI techniques such as neural networks and statistical pattern recognition. Our work opens a new avenue of research in the field of authorship recognition: adversarial stylometry. We find that it is easy for an

This research was supported by Intel through the ISTC for Secure Computing and by DARPA through grant N10AP20014.

Authors' addresses: M. Brennan (corresponding author), S. Afroz, and R. Greenstadt, Computer Science Department, Drexel University; email: mb553@drexel.edu.

DOI 10.1145/2382448.2382450 http://doi.acm.org/10.1145/2382448.2382450

untrained individual to modify his or her writing style to protect her identity from being discovered through stylometric analysis. We demonstrate the effectiveness of multiple methods of stylometry in nonadversarial settings and show that authors attempting to modify their writing style can reduce the accuracy of these methods from over 95% to the level of random chance. With this we have demonstrated that current approaches to stylometry cannot be relied upon in an adversarial setting. It also demonstrated that for individuals seeking anonymity, manual attempts at modifying writing style are promising as a countermeasure against stylometry. We also show that automated attempts at circumventing stylometry using machine translation may not be as effective, often altering the meaning of text while providing only small drops in accuracy. We compiled and published the first two corpora of adversarial stylometry data (the Brennan-Greenstadt Adversarial Stylometry Corpus and the Extended-Brennan-Greenstadt Corpus) to allow others to carry out their own research on the impact of adversarial text on new and existing methods of stylometry.

Historians and literary detectives have used stylometry with great success to identify the authors of historical documents such as the Federalist Papers, Civil War letters, and Shakespeare's plays [Klarreich 2003; Oakes 2004]. The importance of stylometry can be seen through modern applications in the field of forensics, plagiarism, and anonymity. Stylometry is even used as evidence in courts of law in multiple countries including Britain and the United States [Morton and Michaelson 1996].

> "In some criminal, civil, and security matters, language can be evidence... When you are faced with a suspicious document, whether you need to know who wrote it, or if it is a real threat or a real suicide note, or if it is too close for comfort to some other document, you need reliable, validated methods." [The Institute for Linguistic Evidence 2008]

Stylometry has been a successful line of research but there is one underlying assumption that has not been widely challenged: that the author of an unknown document has been honest in his or her writing style. We define adversarial stylometry as the notion of applying deception to writing style to affect the outcome of stylometric analysis. This new problem space in the field of stylometry leads to new questions such as what happens when authorship recognition is applied to deceptive writing? Can effective privacy-preserving countermeasures to stylometry be developed? What are the implications of looking at stylometry in an adversarial context?

Dr. Patrick Juola, an expert in computer linguistics at Duquesne University, discussed the importance of research in this area in his 2008 monograph on authorship attribution, stating there is obviously great potential for further work here [Juola 2006]. Our research shows that nonexpert human subjects can defeat multiple stylometry methods simply by consciously hiding their writing style or imitating the style of another author.

Stylometry is also a necessary new ground for research in privacy and security. Current anonymity and circumvention systems focus strongly on location-based privacy but do not address many avenues for the leakage of identification through the content of data. Writing style as a marker of identity is not addressed in current circumvention tools, nor is it addressed in the security and privacy community at large. Given the high accuracy of even basic stylometry systems this is not a topic that can afford to be overlooked.

Our contributions include multiple methods of circumventing author recognition through stylometry and experimental results that show the efficacy of doing so through deceptive writing. We found that the accuracy of all the methods we examined dropped significantly when presented with adversarial passages created by inexperienced human adversaries using two of the three circumvention methods. This work demonstrates the need for testing stylometry techniques in adversarial contexts when the application domain warrants it and a framework (including a publicly available adversarial corpus) for doing so.

Section 2 presents background and related work in stylometry and discusses the implications of stylometry techniques for privacy and anonymity. Section 3 discusses our experimental methodology and presents our three methods for circumventing stylometry. Section 4 presents experimental results of classifying circumvention passages on three representative stylometry techniques and discusses the techniques our human subjects used to modify their writing style. Section 5 discusses implications for future work.

2. BACKGROUND, RELATED WORK AND MOTIVATION

Stylometry is based on the assumption that every individual has a unique style to his or her writing to some degree just as every individual has a unique fingerprint. The most famous early example of this assumption being put into practice with stylometry is the classic case of the Federalist Papers. Eighty-five papers were published anonymously in the late 18th century to persuade the people of New York to ratify the American Constitution. The authorship of 12 of these papers was heavily contested [Oakes 2004]. To discover who wrote the unknown papers, researchers have analyzed the writing styles of the known Federalist authors and compared them to the papers of unknown authorship. The features used to determine writing styles have been quite varied. Original attempts looked at the length of words, whereas later attempts looked at pairs of words, vocabulary usage, sentence structure, function words, and so on. Most studies show the author was James Madison.

Recent work in stylometry has been geared towards high accuracy in larger and more diverse datasets. Writeprints is a method that has demonstrated superiority over all other modern stylometry techniques because of its ability to categorize large corpora that include 25 to 100 unique authors in many different contexts including online chat messages, eBay comments, email messages [Abbasi and Chen 2008]. It is an unsupervised method that can be used for both authorship recognition of known authors and similarity detection among unknown authored documents. This approach is significant in the field of stylometry because of its use of individual author-level feature sets and pattern disruption and also because of its high accuracy in authorship recognition.

Stylometry has played a serious role in forensics over the years, especially in the 1990s with the introduction of the cumulative sum, or *cusum*, technique. This method measures the stability of a specific feature throughout multiple documents to establish authorship and was adopted in many court cases, specifically in England. But the accuracy of cusum came under fire, culminating in massive controversy about the use of the technique [Matthews 1993]. It has since been modified with better accuracy that has been supported by researchers but questions of accuracy are still considered by some to be the biggest problem faced by the field [Juola 2006].

A number of resources are available that give an overview of stylometry methods that exist [Malyutov 2006; Uzuner and Katz 2005], and focus on the state of the field as it relates to computer science and computer linguistics [Juola 2006] or digital forensics [Chaski 2005]. Artificial Intelligence has been embraced in the field of stylometry, leading to more robust classifiers using machine learning and other AI techniques [Tweedie et al. 1996; Holmes and Forsyth 1995; Uzuner and Katz 2005].

Early work in the field we now call adversarial stylometry includes research by Dr. Josyula Rao and Dr. Pankaj Rohatgi studying the impact of stylometry on pseudonymity and determining that 6500 words of writing is enough to leak the identity of an author [Rao and Rohatgi 2000]. Others have looked at the potential of automatically obfuscating a document to preserve anonymity and determined that in the case of the Federalist Papers it took just 14 changes per 1000 words to shift authorship from Madison to Hamilton [Kacmarcik and Gamon 2006]. This work, however, does not modify the actual text in the documents. They instead modified the numerical feature vectors after they have been extracted from the original text.

Previous research in the field has also looked at authorship recognition and pastiche by comparing the work of Gilbert Adair to that of Lewis Caroll, whom he was trying to imitate by writing follow-up stories to Alice in Wonderland [Somers and Tweedie 2003]. However, our work is the first to apply adversarial stylometry to actual documents written by humans to defeat stylometry and test the results against multiple methods and features sets. Patrick Juola validated the effectiveness of adversarial writing on stylometry by evaluating the Brennan-Greenstadt corpus with JGAAP[1] and also demonstrated some methods' resistance to such writing samples [Juola and Vescovi 2010].

2.1. The Role of Stylometry in Privacy and Anonymity

A multidisciplinary approach to privacy has long demanded greater attention that includes much more than traditional location-based circumvention and anonymity tools [Adams 2006]. Users of the Internet with the desire to publish anonymously may also desire to hide their writing style to circumvent stylometry techniques. The goal of these authors is to keep their identity private. With the growth of stylometry as an accurate means of determining authorship, the question of how it affects privacy and anonymity in the information age is becoming increasingly important. Privacy and anonymity are held in high regard by many activists, journalists, and law enforcement officers. The introduction of adversarial stylometry and the use of circumvention passages on stylometry are possible means of ensuring the privacy and anonymity of an individual. The largest example of the value of anonymous speech to these groups is the Tor Project, an anonymous communication tool originally developed by the Naval Research Laboratory that is utilized by hundreds of thousands of individuals including law enforcement, journalists, activists, businesses, and more [The Tor Project 2012a, 2012b].

The privacy issues concerning stylometry can be summarized through an example scenario: Alice the anonymous blogger versus Bob the abusive employer: Alice is an employee at Bob's company, the Widget Design Corporation. Alice, a long-time employee, wishes to draw attention to the various systemic abuses that she has observed at the company under Bob's management such as harassment, unpaid overtime, and employees being encouraged to rip off competing widget designers. She decides to do this by publishing an open anonymous letter on the Web she has personally authored detailing these abuses. She takes great care to use privacy enhancing technologies as outlined by Dr. Rao and Dr. Rohatgi so that it is very difficult if not impossible to trace her post back to her identity or her IP address [Rao and Rohatgi 2000].

The letter draws criticism of the company from the press, and Bob decides to discover the author, believing it came from within his company due to the details it revealed. In this case, Bob's company has about 100 employees. If Bob has access to a stylometry system such as Writeprints that has very high accuracy for large numbers of potential authors [Abbasi and Chen 2008], he can collect 6500 words from each employee's writing (probably though various reports and emails that they have written in their time at the company) and come to believe that Alice is the writer of the document with over 90% probability. He can then take action against Alice that may compromise her job.

[1] JGAAP is available at http://jgaap.com.

This hypothetical scenario represents a reasonable threat that is within the range of ability for current methods of stylometry. However, the threat presented to anonymity is not purely hypothetical. In his 2011 book, *Inside Wikileaks*, former Wikileaks spokesman Daniel Domscheitt-Berg discussed the potential impact of stylometry on the organization after attending a presentation on adversarial stylometry at the 26th Chaos Communication Congress.[2]

> "If someone had run WikiLeaks documents through such a program, he would have discovered that the same two people were behind all the various press releases, document summaries, and correspondence issued by the project. The official number of volunteers we had was also, to put it mildly, grotesquely exaggerated." [Domscheit-Berg et al. 2011]

The reality of using stylometry to identify individuals who wish to remain anonymous can also be seen from the perspective of law enforcement. It is highlighted in the recently commissioned FBI report *State-of-the-Art Biometric Excellence Roadmap*(SABER).

> "As non-handwritten communications become more prevalent, such as blogging, text messaging and emails, there is a growing need to identify writers not by their written script, but by analysis of the typed content. Currently, there are some studies in the area of writerÔs colloquial analysis that may lead to the emerging technology of writer identification in the blogosphere." [Colosimo et al. 2009]

Recent work by the authors of this article examines the domain of deception and writing style in real-world scenarios like the "Gay Girl in Damascus" blog in which an American writer masqueraded as a Syrian woman. This fraudulent blog was uncovered through forensic means but this research shows that stylometry demonstrates a strong correlation between the pseudonymous blog and writing samples by the true identity of the author [Afroz et al. 2012].

The technical ability for a method of stylometry to present such a threat to anonymity is explained in the next section. The methodology used in Writeprints in particular has demonstrated the potential for identifying a single author among up to 100 unique identities. As a result, bloggers and others may have reason to circumvent stylometry to protect their privacy and anonymity. Adversarial stylometry can be viewed in this way as a means for maintaining privacy or anonymity.

3. METHODOLOGY

We collected writing samples from human subjects that represented both their unmodified writing style and attempts to create adversarial passages to circumvent stylometry methods. We identified three methods of creating documents to circumvent stylometry, two of which are written by human subjects and one of which is an automated method. We implemented three methods of stylometry to analyze the effect of these adversarial passages by attempting to recognize the original author.

The next section explains our method for creating a survey and collecting the data required for the unique adversarial corpus upon which this study was based. We then detail the three types of circumvention passages that have been examined in this study. Finally we summarize the implementation of each method used in the study and their corresponding feature sets.

[2] http://events.ccc.de/congress/2009/wiki/.

3.1. Study Setup and Format

While there are a virtually limitless number of corpora with known author information that can be used for stylometry research, we had to develop our own unique corpus for evaluating the effects of adversarial passages. This was necessary because in addition to a large number of texts from unique authors we also needed adversarial texts written by those same authors. The collection of this data was performed through the use of a survey by voluntary participants of various backgrounds, though all participants held at least some college education. The Brennan-Greenstadt adversarial stylometry corpus is freely available.[3]

There were three basic elements for participation in the survey. First, each author had to submit at least 6500 words of preexisting sample writing. Each writing sample had to be from some sort of formal source, such as essays for school, reports for work, and other professional and academic correspondence. This was intended to eliminate slang and abbreviations, instead concentrating on consistent, formal writing style of everyone involved. This also helped to limit possible errors that are not a result of the malicious attempts but nonetheless could have an effect on the accuracy of the authorship attribution. Participants submitted at least two sample documents, with a median submission of 5 documents per author, which were combined and split into approximately 500-word sample passages.

Second, each author had to write an obfuscation passage in which he tried to hide his identity through his writing style. This was accomplished by writing a new 500-word passage on a specific topic. The task given to them was to write a description of their neighborhood to someone who has never been there as part of a college application. This task was designed to encourage thought about the writing style and not weigh down the author with having to think about a complex topic or be too creative with her content.

Third, each author had to write an imitation passage in which she tries to imitate another author’s style. For this task the participants were given a 2500-word sample from *The Road* by Cormac McCarthy to model their passage after. This selection was made for a variety of reasons. Imitation and pastiche is a common writing task, as can be seen through contest such as the International Hemingway Imitation Competition and the Faux Faulkner contest [Afroz et al. 2012]. In addition, Cormac McCarthy has a somewhat distinct writing style.

> “On the far side of the river valley the road passed through a stark black burn. Charred and limbless trunks of trees stretching away on every side. Ash moving over the road and the sagging hands of blind wire strung from the blackened lightpoles whining thinly in the wind.” [McCarthy 2008]

Since the participants are not professional writers this would allow them to make a fair attempt at emulating another style. The popularity of the book was also a factor in its selection for this study as we wanted to engage participants with reading samples that had broad appeal to a variety of readers. In engaging participants we sought more effort in writing passages and thus more robust results. The writing task given to participants was to narrate their day from the point at which they get out of bed, and to do so using a third-person perspective. This is also similar to the events in the sample text. For testing purposes an additional 2500 words were taken from *The Road* and used as training text for Cormac McCarthy along with the original sample. It should also be noted that the excerpt distributed to readers is freely available as a promotional passage from the book.

[3] https://psal.cs.drexel.edu/tissec.

Asking the participants to write the obfuscation passages before the imitation passage was intentional. We were concerned that if participants chose to write the imitation passage first then all of the obfuscation passages would simply read as a second Cormac McCarthy imitation attempt.

3.2. The Brennan-Greenstadt and Extended-Brennan-Greenstadt Corpora

We have published two freely available research corpora. The first is the Brennan-Greenstadt corpus, which is based on a survey conducted through Drexel University and contains 12 authors who volunteered their time and were not compensated for their efforts. This corpus was the basis for our original work on adversarial stylometry [Brennan and Greenstadt 2009]. The second is the Extended-Brennan-Greenstadt corpus containing 45 authors solicited through the Amazon Mechanical Turk platform. Submissions were vetted against a series of guidelines to ensure the quality of the content, as described next.

3.2.1. Brennan-Greenstadt Corpus. Participants for the Brennan-Greenstadt corpus were solicited through classes at Drexel University, colleagues, and other personal relationships. This provided us with submissions from 12 authors. The Brennan-Greenstadt corpus used an earlier version of the survey which had two relaxed requirements. Authors were only required to submit 5000 words of preexisting writing and they were not required to fill out a demographic survey.

While this corpus was sufficient for preliminary results presented in earlier work [Brennan and Greenstadt 2009], we desired a more robust corpus in order to confirm our original findings in a larger author space with a greater diversity of writers and tweaked survey requirements.

3.2.2. Extended-Brennan-Greenstadt Corpus. We utilized the Amazon Mechanical Turk (AMT) platform to create a large and diverse corpus that could be used for more robust analysis.

Amazon Mechanical Turk[4] is a platform that provides access to a large and diverse population that is willing to perform human intelligence tasks. Participants choose tasks that they would like to complete in exchange for a sum of money decided by a task creator.

Submission quality is a serious consideration when using the AMT platform as the completion of a task does not necessarily indicate that the worker has followed the directions and completed it correctly. In order to ensure that the submissions were acceptable we reviewed every submission and judged their acceptability by scrutinizing them according to the guidelines and requirements listed on the submission form. We only removed authors from the dataset who did not adhere to the directions of the survey. We did not remove authors because of the quality of their writing, demographic information, or anything other than their ability to follow directions.

In addition to the existing requirements we published four guidelines that submissions should adhere to.

(1) The submitted preexisting writing was to be “scholarly” in nature (i.e., a persuasive piece, opinion paper, research paper, journals, etc.).
(2) Anything that is not the writing content of the work should be removed (i.e., citations, urls, section headings, editing notes, etc.).
(3) The papers/samples should have a minimal amount of dialog/quotations.

[4] https://www.mturk.com.

(4) Please refrain from submitting samples of less than 500 words, laboratory and other overly scientific reports, Q&A-style samples such as exams, and anything written in another person's style.

As an added incentive for authors to take care with their submissions we offered a bonus payment of two dollars on top of an original payment of three dollars if their submission adhered to the quality guidelines. Of the 100 submissions we received, 45 satisfied the requirements of the survey. These 45 submissions make up the Extended-Brennan-Greenstadt adversarial stylometry corpus and are the basis of the evaluation for this research. It took about one hour on average for a participant to finish the complete task. The entire instruction set for participation is available online.[5]

3.3. Circumventing Stylometry

We developed three methods of circumvention against stylometry techniques in the form of obfuscation, imitation, and machine translation passages. Two of these, obfuscation and imitation, were manually written by human subjects. These passages were very effective at circumventing attempts at authorship recognition. Machine translation passages are automated attempts at obfuscation utilizing machine translation services. These passages were not sufficient in obfuscating the identity of an author. The full results and effectiveness of these circumvention methods are detailed in the evaluation section.

3.3.1. Obfuscation. In the obfuscation approach the author attempts to write a document in such a way that her personal writing style will not be recognized. There is no guidance for how to do this and there is no specific target for the writing sample. An ideal obfuscated document would be difficult to attribute to any author. For our study, however, we only look at whether or not it successfully deters recognition of the true author.

3.3.2. Imitation. The imitation approach is when an author attempts to write a document such that her writing style will be recognized as that of another specific author. The target is decided upon before a document is written and success is measured both by how successful the document is in deterring authorship recognition systems and how successful it is in imitating the target author. This could also be thought of as a "framing" attack.

3.3.3. Machine Translation. The machine translation approach translates an unmodified passage written in english to another language, or to two other languages, and then back to english. Our hypothesis was that this would sufficiently alter the writing style and obfuscate the identity of the original author. We did not find this to be the case.

We studied this problem through a variety of translation services and languages. We measured the effectiveness of the translation as an automated method as well as the accuracy of the translation in producing comprehensible, coherent obfuscation passages.

We performed three language experiments in addition to the English baseline. In all cases the original and final language were English. We performed single-step translations from English to German and back to English as well as English to Japanese and back to English. We then performed two-step translations from English to German to Japanese and then back to English. German was chosen for its linguistic similarities to English and Japanese for its differences.

[5] Online appendix, including participation instructions: https://psal.cs.drexe.edu/tissec.

Table I. Methods and Feature Sets

Method	Feature Set
Neural Network	Basic-9
SVM	Writeprints-Static
Synonym-Based	Vocabulary Frequency

The methods and feature sets examined in this study.

The two machine translation services we compared were Google Translate[6] and Bing Translator.[7] Both services are free and based on statistical machine translation.

3.4. Methods and Feature Sets

We selected a series of stylometry techniques that represent a variety of potential approaches both in machine learning methodology and feature selection. The feature selections range from basic to comprehensive and the methods from simple and novel to robust and unique (see Table I).

3.4.1. Neural Network and the Basic-9 Feature Set. The most straightforward stylometry techniques are those that use traditional machine learning methods with some set of linguistic features [Kacmarcik and Gamon 2006; Rao and Rohatgi 2000]. The effectiveness of neural networks [Tweedie et al. 1996] in stylometry established machine learning as an integral part of modern authorship analysis.We implemented a neural network with a simple, straightforward feature set. The purpose of the simple feature set and basic machine learning approach is to demonstrate the effectiveness of stylometry even with a limited representation of something as complex as writing style.

The features used for the neural network and SVM classifiers, which we will call the "Basic-9" feature set, include nine linguistic measurements: number of unique words, lexical density, Gunning-Fog readability index,[8] character count without whitespace, average syllables per word, sentence count, average sentence length, and the Flesch-Kincaid Readability Test.[9] The number of hidden layers in the neural network classifier was based on the number of features and the number of classes: $(number\_of\_features + number\_of\_classes)/2$. The feature extraction for this set was done with the JStylo tool.[10]

3.4.2. Synonym-Based Approach. Developed by Clark and Hannon, the synonym-based approach demonstrates the continuing value of novel techniques in stylometry. This method exploits the choice of a specific word given all the possible alternatives that exist. The theory behind this method is that when a word has a large number of synonyms, the choice the author makes is significant in understanding his or her writing style [Clark and Hannon 2007]. An example analysis of three sentences can be seen in Figure 1. The synonym-based approach represents the potential effectiveness of using a single type of feature vector for stylometric analysis.

The method called for a vocabulary-based feature set. A feature vector is created for each word w in a text, having two elements: the number of synonyms s that the word has according to Princeton's WordNet lexical database [Miller 1995], and the shared frequency n of the word w between the sample text and the training text of a known author. The match value for a sample text u from an unknown author and a reference

[6] http://translate.google.com.

[7] http://www.microsofttranslator.com.

[8] http://en.wikipedia.org/wiki/Gunning_fog_index.

[9] http://en.wikipedia.org/wiki/Flesch-Kincaid_readability_test.

[10] JStylo is available at https://psal.cs.drexel.edu.

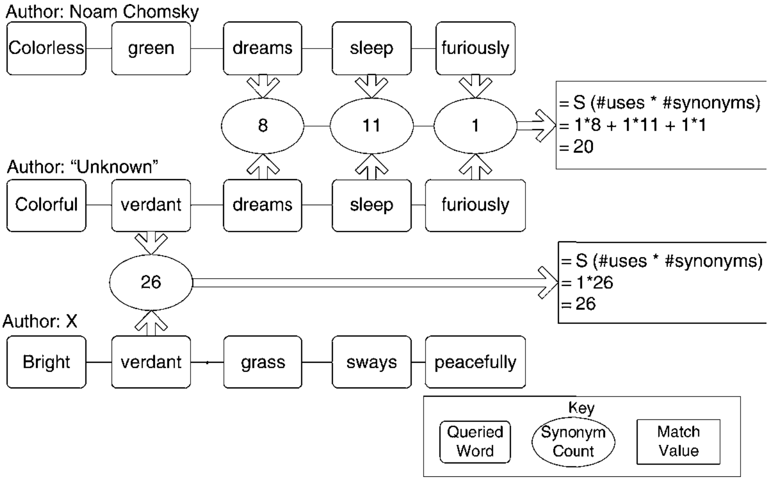

Fig. 1. An example from Clark and Hannon demonstrating how values are calculated to compare two known sentences to an unknown sentence. In this case the unknown sentence is correctly attributed to author X because of the highly salient feature of choosing the word "verdant" [Clark and Hannon 2007].

text k from a known author is then the sum of $n \times s$ for all shared words between the two texts. Authorship is attributed to a text based on the known author with the highest match value to the sample text. The method also takes into account the overall frequency of a word in all of the available text as well as removing words that appear on a stop-list of the 319 most common words in the English language. A sample text is attributed to the author with the highest match value over all samples from that author.

3.4.3. Support Vector Machine and the Writeprints-Static Approach. Writeprints is one of the most successful methods of stylometry that has been published to date because of its high levels of accuracy on a range of datasets with large numbers of unique authors. Unfortunately this accuracy comes at a high computation cost. In order to perform the robust experiments we designed for this study we have created our own approximation of the Writeprints algorithm that performs comparably but has much lower computation cost that allows us to run large numbers of experiments in a reasonable timeframe. We will summarize the Writeprints method and highlight the feature sets created for this approach and how we merged those feature sets into our approach.

The Writeprints technique constructs a single classifier using feature sets that are specific to each individual author rather than being generalized across the set of potential authors. The method has two major parts: writeprint creation and pattern disruption. The writeprint creation step constructs n-dimensional hyperplanes that represent an individual author's writing style, where n is the number of features in the feature set. The pattern disruption step identifies zero usage features and shifts the writeprint representation further away from writeprints that have nonzero values for the same features which in turn decreases the level of stylistic similarity between two separate authors. There are many nuances to this approach that we will not discuss here but are described in detail in the original research paper on Writeprints.

One of the most valuable pieces of research that has come from the creation of Writeprints are the baseline and extended feature sets [Abbasi and Chen 2008]. The baseline dataset has 327 features whereas the extended set contains tens of thousands. The primary difference between these two datasets, however, is that the baseline set

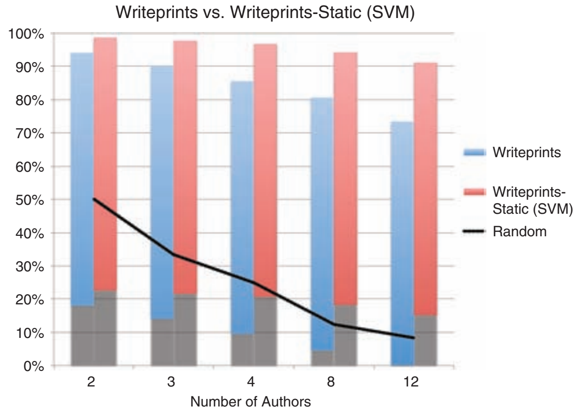

Fig. 2. The Writeprints approach versus Writeprints Static on the Brennan-Greenstadt corpus.

contains only static features in that the contents do not change with the addition and removal of documents. The extended feature set contains many elements that are based on the documents being classified, and is much larger as a consequence. Examples of these dynamic features are the most common misspellings and character bigrams in the corpus. We used many pieces of the feature set created by Writeprints.

To mitigate the issue of Writeprints' high computation cost we have combined a hybrid version of the Writeprints feature sets with a support vector machine. This results in a faster method that has a higher, but comparable, precision on our corpus. We validated the effectiveness of this approach by comparing it to the original Writeprints approach for select datasets from the original Brennan-Greenstadt corpus and found that the precision of our approach is comparable to the precision of the complete Writeprints method, as can be seen in Figure 2. This is also in line with results from the original Writeprints research which compared the approach with a variety of others, including SVMs using the same feature set.

The feature set we use combines the brevity and static nature of the baseline set with some of the more complex features of the extended set. We call this the Writeprints Static feature set. It contains 557 static features, detailed in Table II. We applied this feature set to a Support Vector Machine (SVM) classifier in the form of a Sequential Minimal Optimization (SMO) with a polynomial kernel using Weka machine learning software.[11]

4. EVALUATION

There are two ways to think of “success” when evaluating how stylometric methods respond to adversarial writing samples. One way is to measure the success of the method in identifying the true author of a document intended to circumvent stylometry and the other is to measure the success of the circumvention passage in preserving the anonymity of the author. We will examine the results from both angles.

To test the success we look at the performance on different sets of unique authors. Our dataset consisted of a total of 45 unique authors. This is a larger number of unique authors than almost all of stylometry studies cited throughout this article and is in line with the current state-of-the-art, Writeprints, which looked at writing samples of 25, 50, and 100 unique authors.

[11] http://www.cs.waikato.ac.nz/ml/weka/.

Table II. Writeprints Static Feature Set

Group	Category	No. of Features	Description
Lexical	Word level	3	Total words, average word length, number of short words
	Character level	3	Total char, percentage of digits, percentage of uppercase letters
	Special Character	21	Occurrence of special characters
	Letters	26	Letter frequency
	Digit	10	Digit frequency (e.g., 1,2,3)
	Character bigram	39	Percentage of common bigrams
	Character trigram	20	Percentage of common trigrams
	Vocabulary Richness	2	Ratio of hapax legomena and dis legomena
Syntactic	Function Words	403	frequency of function words
	POS tags	22	Frequency of Parts of speech tag
	Punctuation	8	Frequency and percentage of colon, semicolon, qmark, period, exclamation, comma

The Writeprints-Static feature set, adopted from the Writeprints approach [Abbasi and Chen 2008].

In order to evaluate the corpus we set up test sets for 1000 unique sets of 5, 10, 15, 20, 25, 30, 35, and 40 authors out of a total sample pool of 45 authors from the Extended-Brennan-Greenstadt corpus. The precision measurement discussed for any author count throughout this section refers to the average across all 1000 sets. All of the baseline results are based on tenfold cross-validation. The precision for classifying obfuscation, imitation, and translation passages is measured by training each classifier on the entire unmodified corpus for the authors in a test set and testing that classifier on the corresponding circumvention passages. The graphs in this article refer to the precision because we believe that is the most important and intuitive measurement when determining the authorship of an individual, unknown document. We have made additional graphs available reflecting recall and f-measure on our Web site[12].

The high number of combinations is uncommon in stylometry research but we believe it is important. Accuracy between different sets of authors can vary significantly depending on the specific authors chosen. By viewing the potential combinations of authors as the sample space and a specific combination of authors as a sample selection we are able to make robust accuracy claims with minimal standard error. Standard error in this case is 3.1%.

4.1. Evaluating the Extended-Brennan-Greenstadt Corpus

In order to substantiate the results we present in this article as being in line with our previous work we evaluated all of the methods presented in this work on our original dataset, the Brennan-Greenstadt corpus. We utilized the author counts available in our original paper given the smaller dataset. We found that the precision for each approach is comparable on all datasets. The Basic-9 neural network approach saw a slight drop as can be seen in Figure 3. The others were nearly identical as seen with the Writeprints Static approach in Figure 5 and the synonym based approach in Figure 4.

4.2. Baseline

Figure 6 demonstrates the effectiveness of the four methods we tested and the accuracy of random chance classification. The random chance line in all figures represents what

[12] Precision, recall and F-measure graphs are available at https://psal.cs.drexel.edu/tissec.

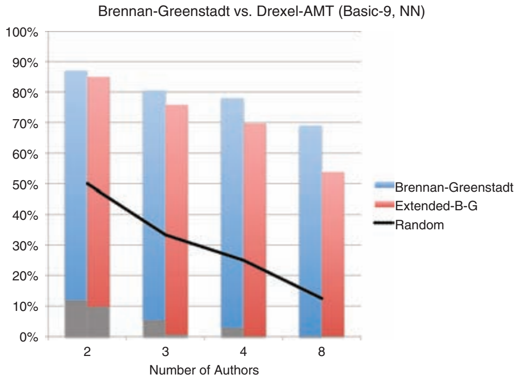

Fig. 3. Brennan-Greenstadt and Extended-Brennan-Greenstadt with the Basic-9 neural network approach.

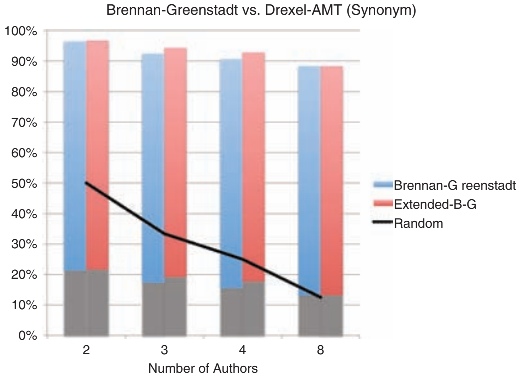

Fig. 4. Brennan-Greenstadt and Extended-Brennan-Greenstadt with the synonym-based approach.

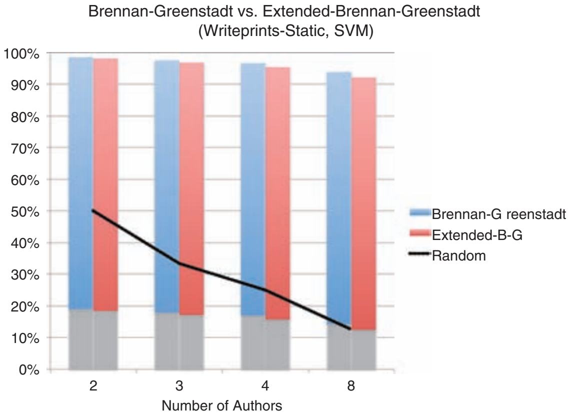

Fig. 5. Brennan-Greenstadt and Extended-Brennan-Greenstadt with the Writeprints Static SVM approach.

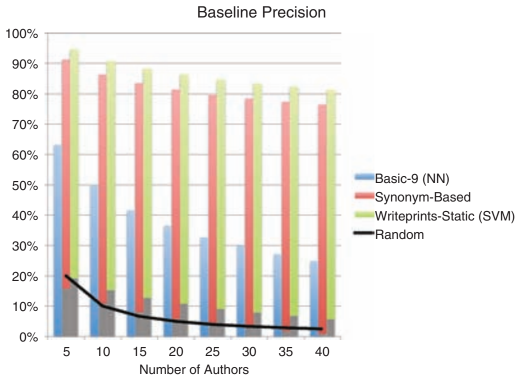

Fig. 6. Baseline accuracy.

the precision would be if the authorship of a document was determined by randomly selecting one of the potential authors. All of the results for the baseline precision measurements are statistically significant over random chance. All methods show a degradation of precision as the number of unique authors increases but the effectiveness is still quite substantial at even the largest author set. The Writeprints Static feature set utilizing an SVM demonstrates the highest precision overall. The synonym-based approach is also very effective. The Basic-9 feature set does poorly compared to the other two methods, but is still far above that of random chance. This is important and confirms our hypothesis that even a very simple measurement of writing style is effective for small numbers of authors and still demonstrates significant ability for deanonymization with larger numbers of authors.

Our SVM approach was evaluated using tenfold cross-validation in the same manner as the rest of our experiments. This is more robust than the approach utilized by the Writeprints authors. In addition, the amount of training data available per author is much lower in our dataset than most of the sets used in the Writeprints research. Our classification is based on approximately 6500 words per author (we allow for error within 250 words of 6500 in order to not break-up sentences) as opposed to anywhere from 23,000 to 43,000 in most datasets used in the Writeprints study. These points are important because while our method does not achieve the same levels of precision at the Writeprints approach, it still approaches those numbers despite the relaxed constraints of our evaluation.

The neural network approach displays accuracies varying between 63% and 25%. This is not as effective as the other approaches but is high enough to be an effective form of authorship recognition and a legitimate threat to privacy and anonymity. The synonym-based approach, however, performs exceedingly well even at higher numbers of unique authors. The degradation in precision follows the same decline as the Writeprints Static feature set approach.

The synonym and Writeprints Static approaches follow a smooth degradation curve as the number of authors in each set increases. This confirms hypotheses of our original research that the degradation curve we observed using the original Brennan-Greenstadt corpus would extend to larger numbers of authors.

4.3. Obfuscation and Imitation Circumvention Approaches

Attempting to recognize the authors of the obfuscation passages results in a drop of accuracy to around that of chance classification as can be seen in Figure 7. Only the SVM (Writeprints Static) approach displayed an effectiveness above that of random chance.

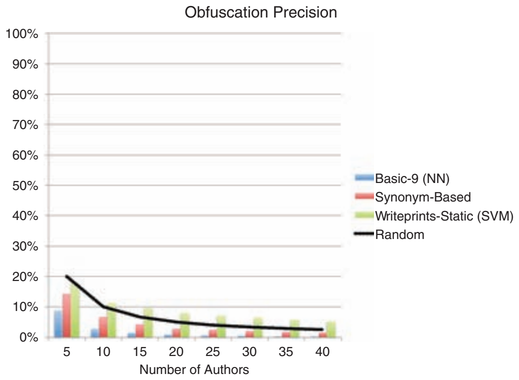

Fig. 7. Detection of obfuscation attacks.

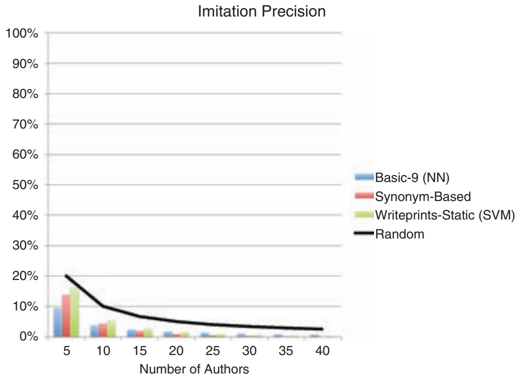

Fig. 8. Accuracy in imitation detection.

This demonstrates the weakness of these methods in detecting writing style designed to be adversarial. The fact that the obfuscated passages track closely to random chances is exactly what we would expect for well-constructed obfuscation passages.

Imitation passages demonstrated a significant drop in effectiveness beyond the obfuscation approach for the neural-network-and SVM-based approaches, and maintained chance-level accuracy for Writeprints and the synonym-based methods as is seen in Figure 8. The lower-than-chance results for SVM and neural network approaches demonstrates a difference in how resistant different methods are to adversarial passages. This might indicate that the features modified in attempting to imitate Cormac McCarthy are well represented in the Basic-9 feature set but not so much by the extensive Writeprints Static feature set or the specific but unique synonym analysis.

Furthermore the ability of the imitation passages to successfully imitate Cormac McCarthy was high. This is demonstrated in Figure 9. The synonym-based approach, despite very high precision in the baseline tests, was just as susceptible to misclassification of imitation passages towards Cormac McCarthy as the neural network approach that it greatly outperformed in the baseline evaluation.

It is important to note that the testing of imitation success required a modification to the evaluation process. In order to test for the success of an imitation passage in representing Cormac McCarthy's writing styles we needed to include training texts for Cormac McCarthy when building the classifier. This is only done for the measurement

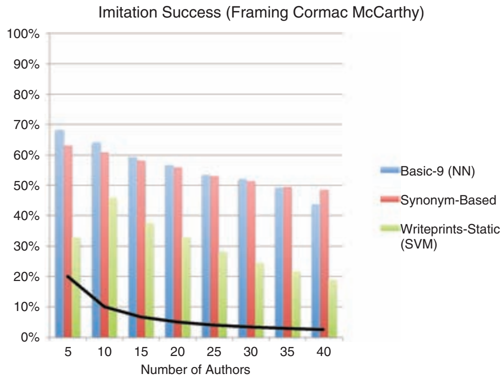

Fig. 9. Success of imitation attacks.

of imitation success; no Cormac McCarthy text is used in the training process for any other tests.

While stylometry methods that are resistant to adversarial passages have been hypothesized [Brennan and Greenstadt 2009; Juola and Vescovi 2010], we can see the first true instance of it here with the results from the SVM approach. The precision when classifying obfuscation passages never drops below that of random chance. Furthermore it can be considered resistant to wrongly classifying imitation passages as their intended target as it generally is only half as likely as the other methods to classify an imitation passage as having been written by Cormac McCarthy. It is likely, however, that more sophisticated circumvention methods could pose a more significant threat to this approach.

4.4. Circumvention by Machine Translation

Since publication of our initial 2009 research on this topic, circumvention via machine translation has been consistently questioned as a possible approach to maintaining anonymity. Our evaluation suggests that machine translation in its current form does not provide a reliable means of circumventing stylometry.

The translation evaluation was performed using passages from the Brennan-Greenstadt corpus. The single-step translations provided a drop in precision of anywhere from 10% to 30%. The two-step translation caused drops from 30% to 60%. Generally, Bing was better at discerning authorship after Japanese translations and Google was better with German translations. Bing, however, performed very poorly with the two-step translations described shortly. We found that the synonym method was especially resistant to these passages and the neural network and SVM methods performed similarly throughout.

4.4.1. One-Step Translations. One-step translations to and from German and Japanese were generally not effective at hiding the writing style of an author. While there was a reduction and it was sometimes statistically significant, such as Bing translations from English to German to Japanese to English on neural network and SVM classifiers, the drop in accuracy is not large enough to grant much comfort to those looking to maintain their anonymity as it generally only decreases the chance of identification from about 65% to about 50% in SVM and neural network approaches and only from 85% to 75% in the case of the synonym-based method. These numbers are not strong enough to warrant a claim that they are effective in providing an anonymizing effect on a document.

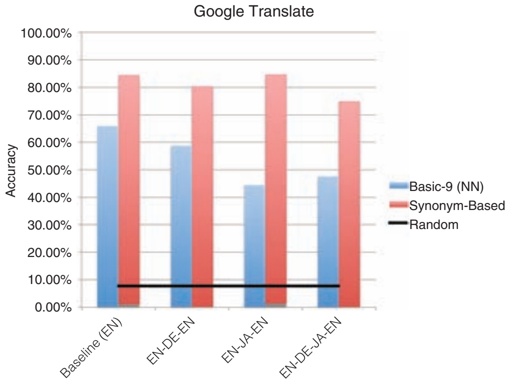

Fig. 10. Google translation results.

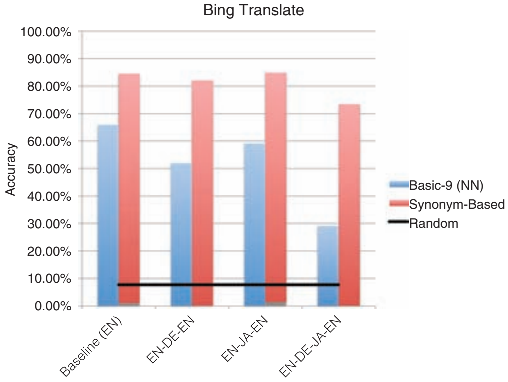

Fig. 11. Bing translation results.

4.4.2. Two-Step Translations. Two-step translations from English to German to Japanese and then back to English were generally no more effective at hiding the writing style of an author than a one-step translation with Japanese except for the case of Bing Translator using SVMs and neural networks as is explained next.

4.4.3. Translation Service Comparison. Single-step translation approaches to and from German were less effective using Google Translate and Japanese translations were less effective on Bing Translator. Bing seemed to produce very effective two-step translation passages for the NN and SVM methods. Overall it appears as though Bing's ability to construct adversarial passages well from an accuracy standpoint is greater than Google Translate. Bing's average accuracy across correctly identifying adversarial translation passages is 12 points lower than Google and when the especially effective synonym-based method is removed the difference increases to 17 points. This would indicate Bing is better for attempting a machine translation-based circumvention approach but overall the accuracies are still not low enough to suggest it would be sufficient to protect privacy and anonymity.

4.4.4. Classifier Method Comparison. The most effective of the three for all experiments as well as the baseline was the synonym-based method. This method was demonstrated to have high accuracy in past work but there is no previous work indicating that accuracy persists when looking at larger numbers of unique authors until now. The baseline of 84.5% for the synonym-based method was the highest of the three classifiers. The SVM and neural network using the Basic-9 feature set had baseline accuracies of 65.1% and 65.9%, respectively.

Using Google Translate the synonym-based method maintained an accuracy of 84.8% for Japanese and 80.5% for German one-step translations. The accuracy dropped only to 75% for the two-step translation. Similar results were found with the Bing Translator. The SVM and neural network methods dropped to 46% and 44.4% for Japanese, respectively, and 50.1% and 58.7% for German. They also saw accuracies similar to Japanese one-step translations for the two-step translation experiment.

Overall these results are, for the most part, not statistically significant in favor of the translation having an anonymizing effect on the writing style of an author and we believe the reduction in accuracy is not enough to warrant calling this an effective approach to circumventing stylometry.

4.4.5. Effectiveness of Translated Documents. Even if we were to accept a drop in accuracy by 15 to 35 points as sufficient for aiding the anonymization of a document, would the resulting translated passage be acceptable for communication purposes or publication? We observed that the answer to this question depends heavily on the complexity of the language being translated. Here is an example sentence from Cormac McCarthy that appeared in his novel *The Road*, along with each translation.

> (Original)
> Just remember that the things you put into your head are there forever, he said.
> (English ⇒ German ⇒ English)
> Remember that the things that you are dead set on always there, he said.
> (English ⇒ Japanese ⇒ English)
> But things are there forever remember what you put in your head, he said.
> (English ⇒ German ⇒ Japanese ⇒ English)
> You are dead, that there always is set, please do not forget what he said.

The original sentence was reasonably complex and did not fare well through the translation process. While the translated sentences were coherent, the meaning was fundamentally changed in each one. But when we look at a simpler sentence from that same passage we find more consistent results.

> (Original)
> They passed through the city at noon of the day following.
> (English ⇒ German ⇒ English)
> They crossed the city at noon the following day.
> (English ⇒ Japanese ⇒ English)
> They passed the city at noon the following day.
> (English ⇒ German ⇒ Japanese ⇒ English)
> They crossed the city at noon the next day.

The translations of the simpler sentence are more effective but lack obfuscation. The goal of the translation approach is to alter the writing style while retaining the meaning. There are many examples of this that can be found in the translated passages, such as “Fighting was tough, with each house and factory fiercely contested.” being translated to “The fight was hard, fought hard with every home and factory.” But these are outweighed by the number of significantly altered meanings, incoherent translations, and very good but nonobfuscated translations.

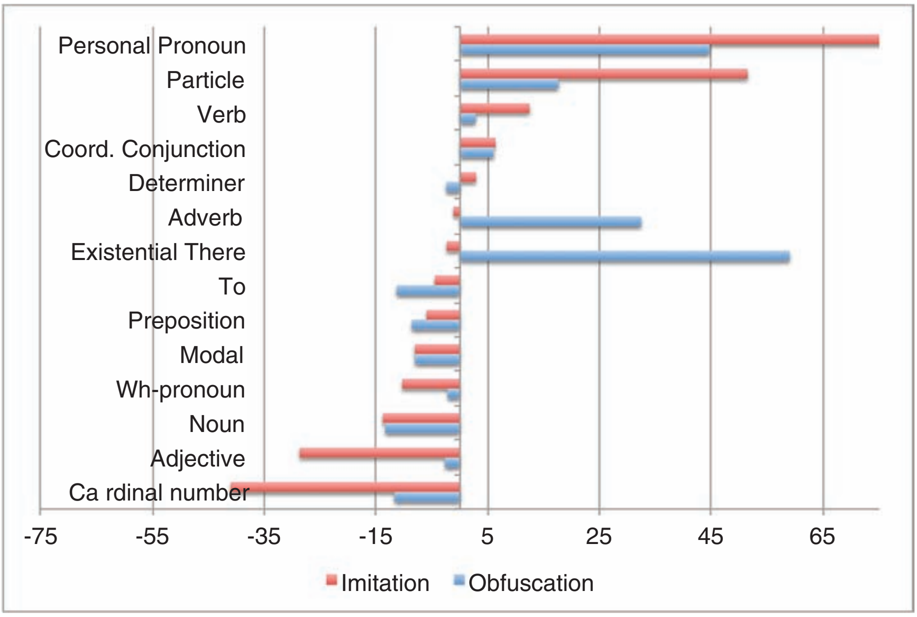

Fig. 12. Changes made in obfuscation and imitation passages in the Extended-Brennan-Greenstadt corpus. The amount to the right of the y-axis represents increases of a feature, and to the left are decreases.

4.5. Adjusting Writing Style

In informal discussions with participants after completing the study, we found that many of them tried to obfuscate their style by "dumbing down" their writing by using shorter sentences and less descriptive words. When imitating the writing style of Cormac McCarthy the participants described attempting to use descriptive and grim language. We can verify at least some of these claims through analysis of the feature frequency changes in the obfuscation and imitation passages.

We compare the normalized frequencies of the features to understand which ones people change to hide their writing style. The changes made to a selection of features for both obfuscation and imitation passages can be seen in Figure 12. This graph illustrates the changes in frequencies for each feature. The y-axis contains a list of features that have been adjusted in the passages and x-axis of the graph denotes the change in each feature. We compute the change in feature using the following formula.

Change in Feature f, $C_f = 100 * (f_{adv} - f_{reg})/(f_{reg} + 1)$

where
f_{adv} = Average values of feature F in adversarial documents.
f_{reg} = Average values of feature F in regular documents.

We add one with f_{reg} in the denominator to avoid divide-by-zero error, as f_{reg} can be zero for some features. The amount to the right of the y-axis represents the increases in a feature and the amount to the left represents the decreases.

In our experiments, the most changed features are average syllables, average word length, sentence count, average sentence length, usage of personal pronouns, adjectives, and verbs, and readability index. We do see hints that authors are "dumbing down" their writing style with shorter sentences, lower readability scores, and less complex words. Most of these are the features in our Basic-9 feature set which may explain why that feature set can be effective despite its relatively small size.

In imitation passages, all the participants use more personal pronouns and verbs and shorter sentences than the regular cases. The personal pronouns can likely be attributed to the direction of the writing survey which asked participants to describe their day in the third person. The shorter sentences, however, are likely a result of imitating Cormac McCarthy who often uses short sentences in his prose. In both imitation and obfuscation passages, participants use shorter and simpler words (those with only one or two syllables) and shorter sentences. As a result, adversarial writings are easier to read than regular writings.

5. DISCUSSION AND FUTURE WORK

5.1. Authors, Topics and Skill

The amount of training text for each author is not exceptionally large, but the total number of authors in the study is significantly larger than most studies of stylometry methods and on par with new methods such as Writeprints. Having such a large number of authors to work with allowed us to do more extensive testing by choosing random groups of authors and averaging the accuracies across them. This also allowed us to see interesting patterns in the study such as which authors did a better job than others in creating successful adversarial passages. Through our anecdotal observations it was clear that certain authors did a poor job in writing their adversarial passages.

Additionally, certain authors also had a style that seemed to be particularly susceptible to obfuscation attempts. Authorship of obfuscation passages were often attributed to these authors when they were a member of a test set. An interesting avenue of research would be to determine if it is possible to create a generic writing style by automated means.

The domain of possible content in this study was fairly open. Participants were allowed to present samples from a variety of subjects so long as they were scholarly in nature. It could be beneficial to study the effects of adversarial attacks in stricter domains where there is less room for maneuvering and thus less options for how an author could hide his or her identity. Stylometry used in restricted domains may prove less susceptible to our attacks.

In general the participants in this study were unskilled in the field of stylometry and were not professional writers. Despite this lack of expertise they were able to consistently circumvent the authorship attribution methods that were put in place. This strengthens our findings as it would be reasonable to expect authors with expertise in such areas could do a better job at attacking the system.

5.2. Open Problems in Adversarial Stylometry

Given the evidence in this research that hiding one's writing style is an effective means to circumventing authorship recognition, one of the next logical steps is to develop end-user software that can assist users in modifying their writing. This is being addressed with the release of Anonymouth.[13] Anonymouth augments the writing-style modification process with intelligent suggestions driven by implementations and analysis of stylometry techniques outlined in this research and elsewhere. Open research avenues include identifying the most effective structured approach for writing-style modification, resolving the trade-off between comprehensive modification and overfitting the changes for specific recognition methods, and identifying which features may be heavily automated and which must rely greatly on manual input.

[13] https://psal.cs.drexel.edu/anonymouth.

Another important part of continued research in this area is larger and more defined corpora in different languages. Our general corpus satisfies a number of reasonable demands for consistency, length, and focus but there are many other more specific domains that could produce different results when examining adversarial passages. For example, will writing style modifications be more or less effective in a highly restricted domain such as complex scientific research papers or in a very broad domain such as fictional short stories? Are the most salient features for identification in other languages similar to those in English?

6. CONCLUSION

This study demonstrates the effectiveness of adversarial writing against modern methods of stylometry. The analysis of stylometry techniques and their weaknesses to adversarial writing demonstrates that we must test stylometry methods for their resistance to adversaries in situations where their presence is likely. We advocate a stronger stance of not relying on stylometry in sensitive situations where the authorship of an unknown document must be known with a high degree of certainty unless the possibility of a modified writing style is negligible.

The obfuscation approach weakens all methods to the point that they are no better than randomly guessing the correct author of a document. The imitation approach was widely successful in causing authorship to be attributed to the intended imitation target. Additionally, these passages were generated by participants in very short periods of time by amateur writers who lacked expertise in stylometry. Translation with widely available machine translation services does not appear to be a viable mode of circumvention. Our evaluation did not demonstrate sufficient anonymization and the translated document has, at best, questionable grammar and quality.

There has long been a case to be made for a multidisciplinary approach to privacy and anonymity. This research shows both the necessity of considering writing-style analysis as a component of that approach and demonstrates the possibility for privacy-conscious individuals to take steps to maintain their anonymity in the face of advanced stylometric techniques.

This work provides further evidence that learning techniques used in adversarial settings need to be tested with adversarial test sets. This research also has implications for machine translation research through the use of stylometry as a method for testing the effectiveness of machine translation. If a machine translated dataset shows comparable accuracy in an adversarial stylometry setting then the results may be used to validate the translation method.

This study also strengthens the original claims of high accuracies by validating the methods on a large set of new data produced for a variety of purposes. When these methods are used in situations where adversaries are not considered to be a threat, they perform quite well.

ACKNOWLEDGMENTS

The authors wish to acknowledge the hard work of Privacy, Security and Automation Lab researchers Andrew McDonald, Aylin Caliskan, and Ariel Stolerman.

REFERENCES

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

Adams, C. 2006. With a little help from my friends (and colleagues): The multidisciplinary requirement for privacy. http://www.idtrail.org/content/view/402/42/.

Afroz, S., Brennan, M., and Greenstadt, R. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *Proceedings of the IEEE Symposium on Security and Privacy*.

BRENNAN, M. AND GREENSTADT, R. 2009. Practical attacks on authorship recognition techniques. *Innov. Appl. Artif. Intell.*

CHASKI, C. E. 2005. Who's at the keyboard: Authorship attribution in digital evidence investigations. In *Proceedings of the 8th Biennial Conference on Forensic Linguistics/Language and Law*.

CLARK, J. H. AND HANNON, C. J. 2007. A classifier system for author recognition using synonym-based features. In Lecture Notes in Computer Science, vol. 4827. Springer, 839–849.

COLOSIMO, M., GRAEF, R., LAMPERT, S., AND PETERSON, M. 2009. State of the art biometrics excellence roadmap. Tech. rep., U.S. Department of Justice Federal Bureau of Investigation.

DOMSCHEIT-BERG, D., KLOPP, T., AND CHASE, J. 2011. *Inside Wikileaks: My Time with Julian Assange at the World's Most Dangerous Website*. Crown Publishers.

HOLMES, D. AND FORSYTH, R. 1995. The federalist revisited: New directions in authorship attribution. *Liter. Linguist. Comput. 10*, 111–127.

JUOLA, P. 2006. Authorship attribution. *Found. Trends Inf. Retr. 1*, 233–334.

JUOLA, P. AND VESCOVI, D. 2010. Empirical evaluation of authorship obfuscation using jgaap. In *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security (AISec '10)*. ACM, New York, 14–18.

KACMARCIK, G. AND GAMON, M. 2006. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL* Main Conference Poster Sessions. Association for Computational Linguistics, Morristown, NJ, USA, 444–451.

KLARREICH, E. 2003. Bookish math. *Sci. News 164,* 25.

MALYUTOV, M. 2006. Information transfer and combinatories. In Lecture Notes in Computer Science, vol. 4123. Springer, 3.

MATTHEWS, R. 1993. Linguistics on trial: Forensic scientists have fiercely condemned a technique used in court to show that confessions have been tampered with. *New Sci. 1887*.

MCCARTHY, C. 2008. *The Road*. Vintage International. Knopf Doubleday Publishing Group.

MILLER, G. 1995. Wordnet: A lexical database for english. *Comm. ACM 38*, 39–41.

MORTON, A. AND MICHAELSON, S. 1996. The qsum plot. Internal rep. CSR-3-90.

OAKES, M. P. 2004. Ant colony optimisation for stylometry: The federalist papers. In *Proceedings of the 5th International Conference on Recent Advances in Soft Computing.* 86–91.

RAO, J. R. AND ROHATGI, P. 2000. Can pseudonymity really guarantee privacy? In *Proceedings of the 9th Conference on USENIX Security Symposium*.

SOMERS, H. AND TWEEDIE, F. 2003. Authorship attribution and pastiche. *Comput. Humanit. 37*, 407–429.

THE INSTITUTE FOR LINGUISTIC EVIDENCE. 2008. Mission & philosophy. www.linguisticevidence.org.

THE TOR PROJECT. 2012. Tor metrics portal: Users. https://metrics.torproject.org/users.html.

THE TOR PROJECT. 2012. Who uses tor? https://www.torproject.org/about/torusers.html.en.

TWEEDIE, F. J., SINGH, S., AND HOLMES, D. 1996. Neural network applications in stylometry: The federalist papers. *Comput. Humanit. 30,* 1, 1–10.

UZUNER, U. AND KATZ, B. 2005. A comparative study of language models for book and author recognition. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*. 969.

Received February 2012; revised May 2012; accepted July 2012