Mitigating Deanonymization Attacks via Iterative Language Translation for Social Networks

Jasmine D. Bowers

North Carolina A&T State University

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department: Computer Science

Major: Computer Science

Major Professor: Dr. Kelvin Bryant

Greensboro, North Carolina

2015

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

The Graduate School
North Carolina Agricultural and Technical State University

This is to certify that the Master's Thesis of

Jasmine D. Bowers

has met the thesis requirements of
North Carolina Agricultural and Technical State University

Greensboro, North Carolina
2015

Approved by:

Dr. Kelvin Bryant Major Professor	Dr. Kenneth Williams Committee Member
Dr. Regina Williams Davis Committee Member	Dr. Gerry Dozier Department Chair

Dr. Sanjiv Sarin
Dean, The Graduate School

Biographical Sketch

Jasmine D. Bowers, born in Charlotte, NC August 4 1991, received a Bachelor of Science in Computer Science and Mathematics from Fort Valley State University in 2013. Jasmine is currently working towards obtaining her Master's degree from N.C. A&T State University and is expected to graduate in May 2015. She has worked as a co-op with the Department of Defense and as a summer intern at Lawrence Livermore National Laboratory as a Cyber Defender. Miss Bowers is a member of the Association for Computing Machinery (ACM), the Society of Women Engineers (SWE), the National Society of Black Engineers (NSBE), and the Upsilon Pi Epsilon International Honor Society (UPE) for the computing and information Disciplines.

Dedication

I dedicate this thesis to my mother who continuously sacrificed for me to reach my goals.

Acknowledgments

I would like to acknowledge Dr. Kelvin Bryant, Dr. Kenneth Williams, Dr. Regina Williams Davis, and Dr. Gerry Dozier for their commitment to my success. In addition, I would like to thank the entire Computer Science department, especially the faculty and staff who supported me and provided the love and motivation that I needed to finish.

Table of Contents

List of Figures

List of Tables

Abstract

An author's writing style is an identity attribute that may not be considered when taking steps to remain anonymous. In addition, anonymous social network (ASN) providers continue to encourage users to be assured that their privacy and anonymity is protected. However, ASNs can still collect other personal data (e.g. location, device ID, call information, etc.) about the user. By combining the different types of personal data, an application can generate a profile and then associate it with the author's text. In addition, the writing style of the user remains intact and unaltered, which can result in a deanonymization attack. A deanonymization occurs when the ASN compares the posted text to text online to identify the author. However, if a user were equipped with a method that could conceal his/her writing style when posting text, it would be harder for an ASN provider or a peer on the ASN to reverse engineer the concealment process to identify the author. The goal is to protect users not only from other users but also from the application ASN providers. This research presents an author concealment technique in the form of iterative language translation (ILT), that is effective in preserving author anonymity. Specifically, ILT will be utilized in order to mask an author's writing style and thus decrease his or her recognition rate. The results show that translation of text between languages with higher grammatical differences has a higher success rate for masking an author's identity, e.g. English and Arabic. In addition, if the ASN provider discovers that the text is concealed and attempts to reverse engineer the concealment, three important factors would be required: number of languages (n), number of translators (m), and number of combinations (nm!). This research shows that with three languages (n=3) and one translator (m=1) a tremendous decrease in recognition occurs.

CHAPTER 1

Introduction

1.1 Anonymous Social Networks

Since 2010, anonymous social network (ASN) providers have implemented methods designed to protect their users' identity [5][6][7][8][9][12][14][15]. Whether it is a burning question that one may want to ask (e.g. Ask.fm [6]) or a joke about a teacher (e.g. Fess [7]) parenthetical citation!!!!!! ASN providers continue to encourage users to be assured that their privacy and anonymity is protected. Table 1 provides results of a survey of seven social networking sites that advertise author anonymity to their respective users. The sites include Whisper [8], Secret [9], Fess [7], Omegle [12], Ask.fm [6], Spring.me [14], and YikYak [15]. Each site is categorized by the permissions requested by the ASN provider prior to installation. Many users neglect to thoroughly review the list before installing.

The ASNs shown in Table 1 operate as follows. Whisper [8] allows its users to share posts that consist of text superimposed on an image (of their choice). The average post is a sentence in length. Secret [9] also allows a user to share anonymous posts, but solely with specific Secret users found in his/her address book. The application's privacy implementations include setting a minimum number of Secret friends required for user access to (1) posts of their Secret friends and (2) the source's subgroup (friend or friend of friend). In addition, secrets are posted at random intervals to decrease the likelihood of a user being identified by another user who also posts a secret in close proximity.

Fess [10] is a secret posting site specifically designed for high school students to share secrets with their peers, as text superimposed on a solid colored square. The application verifies the user's age via Facebook [11]. Omegle [12] allows users to chat anonymously with random

people. Ask.fm [13] and Spring.me [14] allow users to post and receive anonymous questions and answers. YikYak [15][16] allows users to post and read anonymous posts of users within the same area based on their device's geo-location. The survey results alone show that although an identifier such as name or email address is not shared when a message is sent, the application will collect other personal data (e.g. location, device ID, call information, etc.) about the user. By combining the different types of personal data, an application can generate a profile and then associate it with the author's text. However, if users were equipped with a method of concealment for his/her writing style when posting text online, it would be harder for an ASN provider or a peer on the ASN to correlate between their text and their personal data that has been collected. The goal is to protect users not only from other users but also from the ASN providers.

Table 1

Survey of 7 Anonymous Social Networking Sites

Anonymous Social Media Sites							
	Whisper	Secret	Fees	Omegle	Ask.fm	Spring.me	YikYak
Device App History	X						
Identity: Find Accounts on Device	X	X	X		X	X	
Contacts/ Calendar		X					
Location	X	X	X				X
Photos/Media/Files	X	X	X	X	X	X	X
Camera/ Microphone	X					X	
Wi-fi Connection Information	X				X		X
Device ID and Call Information	X	X	X		X		X
Receive Data from Internet	X	X	X		X	X	X
Control Vibration	X		X		X	X	
Full Network Access	X	X	X	X	X	X	X
View Network Connections	X	X	X	X	X	X	X
Close Other Apps	X						
Prevent Device from Sleeping	X	X	X		X	X	
Use accounts on the Device	X						
Read Google service Configuration							X
Run at Startup		X	X				
Send Sticky Broadcast					X		

1.2 Deanonymization Attack

A deanonymization attack occurs when text of unknown authorship is compared to a set of text of known authorship, whether online or in a database [2]. In this research, a deanonymization attack occurs when an ASN provider or ASN user attempts to compare the anonymous posts of a user to text on the web in an effort to recover the user's true identity. Although ASN users are able to provide pseudo names and passwords to the ASN provider, the writing style is still exploitable. In response, this research is designed to combat the vulnerabilities those attacks. An ASN provider attack can occur when a provider collects bodies of text of a particular user. Although the ASN provider may only know the user's pseudo name, the provider can compare the writing style of the text of unknown authorship to text of known authorship from sources such as blogs or websites. Since the ASN provider may have access to facts such as the user's geo-location, the search could be easily filtered. For example, if an ASN provider collected text of an author located in Texas, the ASN provider could narrow down their search by eliminating websites and blogs that do not pertain to Texas. A peer attack can occur when an ASN user collects bodies of text written by a peer and compares the writing style to text of known authorship.

The scope of this research solely encompasses the task of masking writing style. Since each respective language translator has its own respective set of rules that it abides by, translations may vary from translator to translator. For consistency purposes, this research contains translated text from only one respective translator. Additionally, certain idioms, culture-based jargon, or other language-dependent writing style components, including tone, may not translate in a uniform way across all translators. Therefore, some translations may not translate the text with the exact tone or mood. However, preservation of such factors could deter the

masking process, since idioms and jargon can serve as identifiers in writing style. Consequently, this research is solely focused on masking.

The remainder of this paper is as follows. Section II discusses author identification and Section III describes the feature extractors used in the experiments. Section IV provides details of the experiments and Section V contains the results. Section VI discusses our conclusions and future work.

CHAPTER 2

Literature Review

2.1 Author Identification

Measured textual features (i.e. number of function words, sentence length, etc.) are utilized to differentiate text written by different authors [2][17][18] in a process known as authorship attribution. The authorship attribution problem [18] is defined as follows: determine the author of a text where the author is unknown. According to Stamatatos [18], computer scientists generally refer to the authorship attribution problem as author identification [1][4][17]. The following text discusses some of the author identification research performed within the last three decades.

Narayanan [2] discusses his author identification techniques that he performed by comparing writing styles of authors. He notes that any manually generated text will ultimately possess specific traits of the author, which creates an advantage when comparing text of unknown authorship to a set of text of known authorship. He refers to the process as a "deanonymization attack." His process includes a nearest neighbor classifier, which successfully matched one blog against 100,000 blogs at a 12% success rate.

In [17], Green discusses his author identification technique performed on short text of 140 characters or less. He categorizes author identification as a subfield of natural language processing (NLP) that utilizes machine learning to identify an author based on features such as parts of speech, frequency, etc. In his research, he utilized a support vector machine (SVM) and extracted both bag-of-words (BOW) and style marker feature sets. A support vector machine is a classification technique that can handle high-dimensional data. The bag-of-words method counts the frequency of each word in a sentence and then represents each unique word count in a vector

[29]. The For example, the sentence “I love big dogs and small dogs.” would be represented as [1, 1, 1, 2, 1, 1], where “I”, “love”, “big”, “and”, and “small” are all found once, while “dogs” is written twice. His results show that the style marker approach outperformed the BOW approach, since the small amount of text limited the vocabulary of the author.

Madigan [19] notes that some researchers depend on topic specific text containing topic-dependent function words to identify an author. Conversely, law enforcement and cyber forensic analysts rely on topic independent comparisons. Coulthard [20] discusses the use of linguistic fingerprinting to perform linguistic investigation of authorship for forensic purposes – a discipline implemented in recent cases involving author identification of a text. A linguistic fingerprint is established when the word-choice tendencies of the speaker or author are apparent in their work (e.g. appalled vs. shocked). Although some techniques require a minimum amount of text to successfully identify an author, MacLeod [21] discusses the development of automating forensic linguistic techniques that have been used in court cases to successfully identify an author of a shorter body of text.

In [23], Stamatatos describes author identification as a prediction of the most likely author of a text when given a set of pre-defined possible authors and a certain number of text samples per author. He goes on to describe stylometry, a subset of author identification, which involves the extraction of writing style-based features. Examples of such feature extraction include quantifying vocabulary richness, as well as and counting the frequency of words and parts of speech. His contribution consisted of a feature selection method for variable-length n-grams, vs. the traditional fixed length n-grams (3-gram, 4-gram, and 5-gram).

In [4], de Vel discusses author identification in terms of forensic analysis. Specifically, he focuses on how email content, as well as headers, attachments, etc., will remain consistent with

an author's writing style, as would the traditional writing style characteristics. His definition of email author identification is similar to that of Stamatatos' - identifying an author from a collection in which the author is included versus identifying an author from a larger collection of largely unknown authors.

Narayanan [2] mentions that manually generated text often reveals additional attributes of an author's writing style (e.g. spelling errors or idiosyncrasies) and therefore links the author to separate pieces of text. For anonymous authors, such information can prevent the writer from truly being anonymous. Such deanonymization attacks can be further mitigated by the approach presented in our research. Li [24] complements this explanation by officially identifying these attributes as an author's writeprint (in comparison to the traditional fingerprint). He includes vocabulary richness, sentence length, keywords, and paragraph layout as contributors to the body of each respective writeprint. Their earlier work in [25] focused on building an authorship identification framework, now writeprint, by extracting lexical, syntactic, structural, and content-specific features. The goal of Li's research was to answer the following questions:

1. Can the author identification technique be applied to online content?
2. Which writing style features are most effective in author identification of online content?
3. Which classification techniques are most effective in author identification of online content?
4. How well can author identification perform in multi-language online content?
5. How effective is author identification when applied to online content with various numbers of authors and messages?

The research of Narayanan and Li shows structural and content-specific features were the most effective features when identifying an author via writing style. In [25], Zheng expresses the

importance of including multiple languages in author identification research due to the fact that writing style features are language dependent.

CHAPTER 3

Methodology

3.1 Feature Extraction

This section discusses the two respective feature extraction methods implemented in this research: uni-gram and stylometric.

3.1.1 Uni-Gram. Uni-Gram feature extraction counts the total number of characters in a sample of text. It also counts the frequency of each character. The feature extractor (FE) then divides each character frequency by the total number of characters and creates a Feature Vector (FV) for each sample of data within the set. One such FE was used by Forsyth [3] in which 95 of the total characters in the ascii set are used. Table 2 shows a table of the 95 characters used.

Table 2

The subset of Unicode characters used in the Uni-Gram style FE by R. S. Forsyth

(space)	!	“	#	$	%	&	‘	(	)
*	+	,	-	.	/	0	1	2	3
4	5	6	7	8	9	:	;	<	=
>	?	@	A	B	C	D	E	F	G
H	I	J	K	L	M	N	O	P	Q
R	S	T	U	V	W	X	Y	Z	[
\	]	^	_	`	a	b	c	d	e
f	g	h	i	j	k	l	m	n	o
p	q	r	S	t	u	v	w	x	y
z	{	|	}	~					

3.1.2 Stylometric-based. Stylometric features are attributes of an author's writing style, such as average word length and average sentence length. Structural features are specific to the document type structure, such as the signature in an email. The FE proposed by O. de. Vel et al. [4] is an example of a stylometric and structural based FE. This FE uses 170 stylometric features and 21 structural features. The 21 structural features are specific to the structure of an email. These features include the number of attachments, greeting/salutation, etc. The style-based features include sentence length, function words, vocabulary richness, etc. Table 3 displays the 170 style marker attributes, where M = total number of tokens (i.e. words), V = total number of types (i.e. distinct words), and C = total number of characters in an e-mail body.

Table 3

The stylometric features proposed by O. de. Vel

Stylometric Features
Number of blank lines/total number of lines
Average sentence length
Average word length (number of characters)
Vocabulary richness i.e., V/M
Total number of function words/M
Function words (122)
Total number of short words/M
Count of hapax legomena/M
Count of hapax legomena/V
Number of characters in words/C
Number of alphabetic characters in words/C
Number of upper-case chars/C
Number of digit characters in words/C
Number of white space characters/C
Number of space characters/C
Number of space characters/white space characters
Number of tab spaces/C
Number of tabs spaces/number of white spaces
Number of punctuations/C
Word length frequency distribution/M (30)

3.2 Iterative Language Translation

According to [26], the transactional model of communication is the process of communication between a sender and a receiver through a medium, such as a cellular device. The receiver is then tasked with doing their best to decode and construct the message. Figure 1 provides a detailed visual explanation. Although the message may be decoded successfully, the interpretation of the message may have been misconstrued. For example, sender A sends the following message to receiver B in a jovial tone: “We need to talk”. However, neither tone nor mood is explicitly stated. Although, receiver B could mistakenly assume that the tone is angry. Therefore, the message was successfully transmitted via text messaging, but the intended meaning was misunderstood. Such misunderstanding can occur when translating text between languages.

Although machine translation technology has advanced, tone and mood are not always accurately interpreted. In addition, languages possess varying grammatical rules and sometimes-complex morphologies [28]. In the Arabic language, for example, words are inflected based on gender, number, and grammatical case. Since the morphology of the Arabic language is more complex than that of the English language, translation from Arabic to English would be more difficult and possibly less accurate than from English to Arabic. Therefore, the meaning, mood, and tense can be misinterpreted, and can ultimately alter the style of writing. See Appendix A for an example of English-Spanish-English translation.

The Systran Business Translator [30] software was used for all translations in this research. Specifically, iterative language translation was performed in order to mask an author’s writing style and thus decrease writing style recognition rate. The author’s original English text

will be compared to text translated into Arabic, Chinese, and Spanish, respectively, and back to English. This cycle is then performed two additional times, resulting in cycles I, II, and III. The results show that as the text is translated into another language and back to English, the meaning and the writing style are altered.

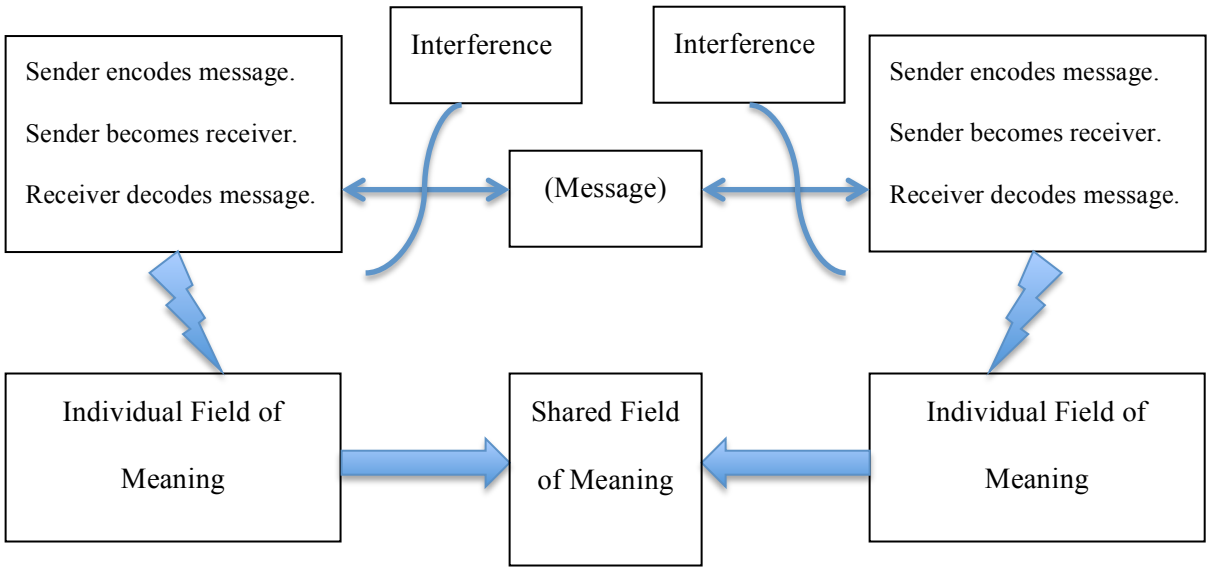

Figure 1: Transactional Model of Communication

If the ASN provider attempts to reverse engineer the Iterative Language Translation (ILT), three factors must be known: number of languages (n), number of translators (m), and number of combinations (nm!). This research shows that with three languages (n=3) and one translator (m=1) a tremendous decrease in recognition occurs. Hence, as n and m increase, the attempt to recover the original writing style becomes more complex (n*m*nm!). In continuation of this research, the tool Author CAAT [31] has been developed which allows n of three languages, m = 1 translators, and results in 3*3! difficulty. In addition, the tool allows the user to translate the text provided at any point. For example, a user wants to translate two sentences. The user can input a sentence, translate the sentence, and then continue typing the second sentence. Once the user types the second sentence, the entire body of text is translated again. Therefore, the first

sentence was translated twice and the second sentence was translated once. This process will further obfuscate the original text, resulting in the inability of the ASN provider to pin point the number of translation iterations that each individual word or sentence encountered.

3.3 Experiments

Experiments were performed on a dataset of 1000 blogs associated with 1000 authors. ILT was implemented by translating each blog into one of our choice languages (i.e. Spanish, Chinese, and Arabic) and back to English. This process was executed three times, resulting in three cycles or iterations. The blogs in the dataset, which consisted of an average of two paragraphs (eight to ten sentences), were then broken into percentages (10%, 30%, 50%, 70%, and 90%) of text, respectively. The amount of text in the original text files was counted and then split by percentage of text. Each percentage was taken starting from the first word in the original file, as seen in Figure 2. The datasets containing 10% of the original text (one to two sentences) can be compared to a body of text of an ASN user. The purpose of the percentage breakdown is to represent various bodies of text, from a two sentence ASN post to paragraphs in a blog, and to show the significance that the amount of text has in terms on identifying its author.

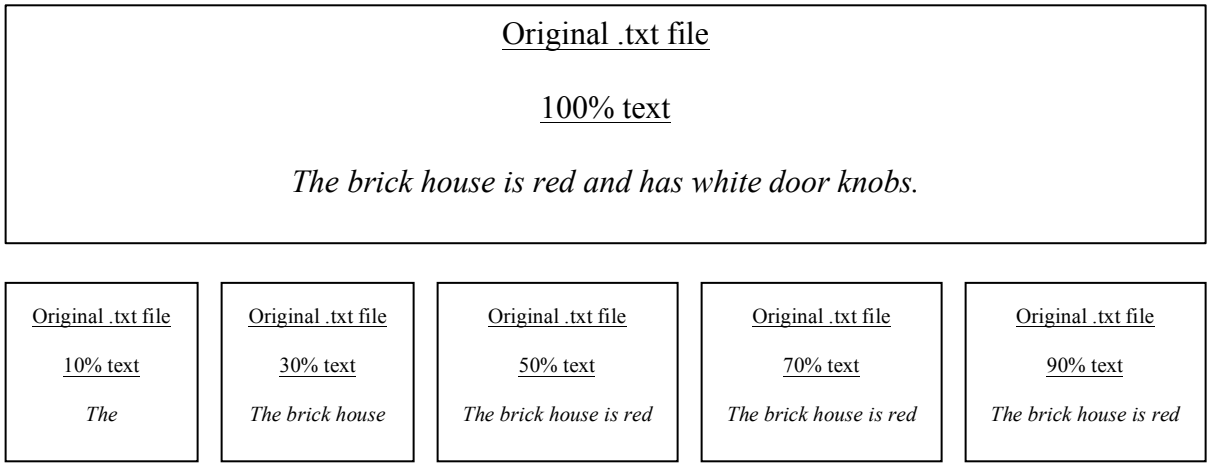

Figure 2. Text Percentage Diagram

As seen in Figure 3, each blog (10-100%) was evenly divided into four separate sub-samples (instances). The first instance was designated as the probe and the remaining three instances were designated as the gallery. Each dataset was labeled in the format of probe-gallery. For example, the original English text was labeled as E-E and the three cycles of Spanish were labeled as E-ESE, E-ESESE, and E-ESESESE. The same process was performed with the Chinese and Arabic languages, resulting in the following datasets: E-EAE, E-EAEAE, E-EAEAEAE, E-EC, E-ECECE, and E-ECECECE.

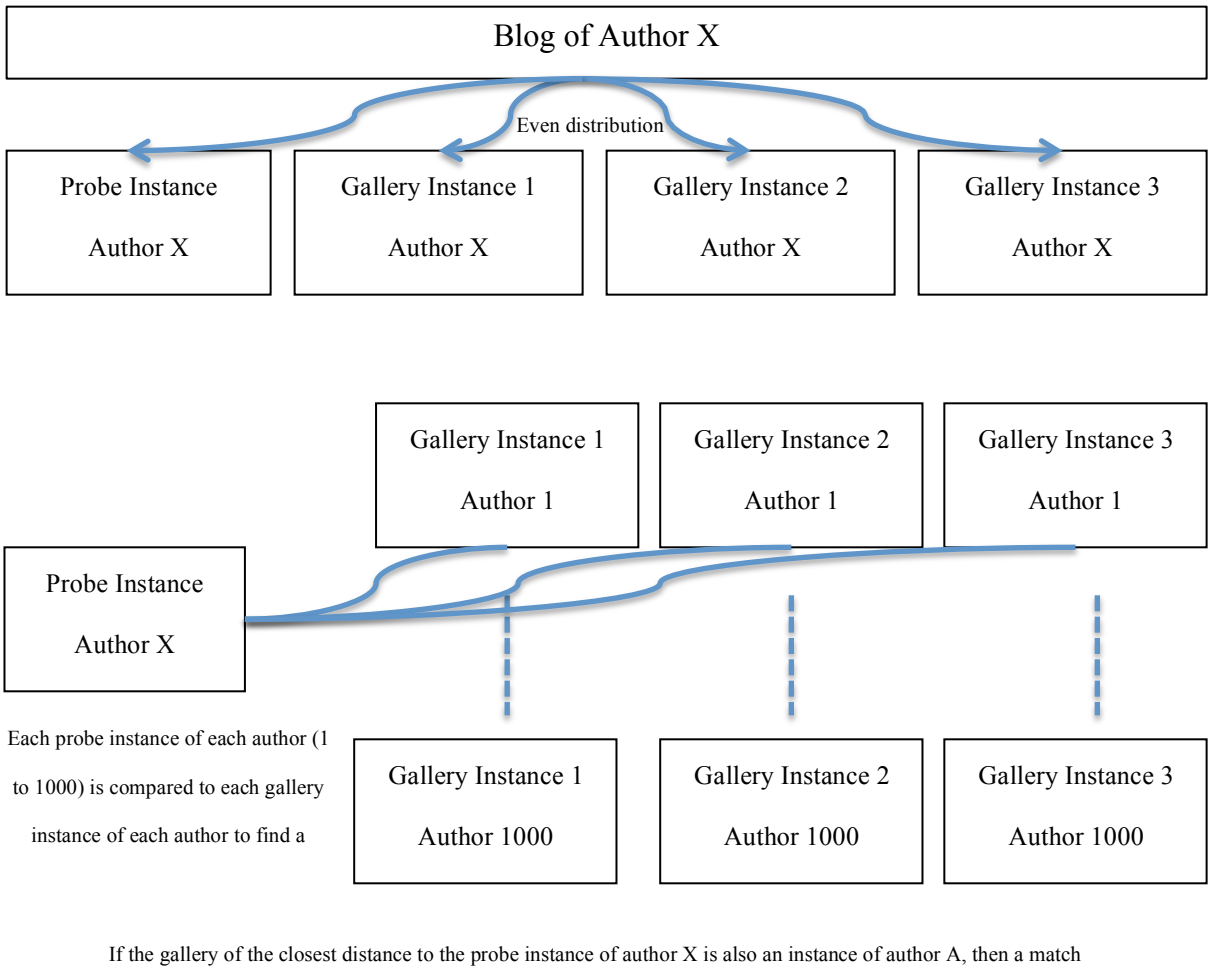

Figure 3. Probe and Gallery Diagram

The Manhattan distance metric was implemented to compare an instance in the probe set to all instances in the gallery. The formula is as follows:

$$dm\ (v1, v2)\ =\ \Sigma|\ v1\ -\ v2\ |$$

where v_1 and v_2 are two distinct feature vectors. The gallery instance with the shortest distance to the respective probe instance was declared the match. If the matching gallery instance is of the same authorship as the probe instance, the match was successful.

CHAPTER 4

Results

Figure 4 provides a detailed comparison of the Uni-Gram and stylometric FE performance on the original E-E datasets. The horizontal axis contains the respective percentage breakdown of the dataset (10% to 100%). The vertical axis measures the identification accuracy. The figure presents a correlation between the amount of text and identification accuracy.
With 100% of the text, the original E-E dataset's accuracy was 12.70% via the Uni-Gram FE and 3.5% via the stylometric FE. As shown, the Uni-Gram FE out-performed the stylometric FE by 9.2%. Therefore, the remaining results consist solely of the performance of the Uni-Gram FE.

Similar to Figure 4, Figures 5-7 show the results of the respective Spanish, Chinese, and Arabic cycles. Figure 5 displays the results of the Spanish iterations. As shown in Figure 4, the original E-E uni-gram results ranged from 3.00% to 12.70%. However, once the text was translated, the results decreased significantly. With 100% of the text (8-10 sentences), the results decreased from 12.70% to 7.60% during the first Spanish iteration. The 5.1% decline represents a 40.16% decrease in recognition rate, which demonstrates the significance of the language translation and its impact on the writing style. As the percentage of text decreases, the recognition rate decreases. Therefore, the less text provided, the more difficult it is to match the writing style to text of known authorship. The results also show that after the first translation iteration occurs, the recognition rate decrease was not as significant, meaning that the first iteration has the most impact.

Figures 6 and 7 provide similar results. With 100% of the text (eight to ten sentences), the results decreased from 12.70% to 6.40% during the first Chinese iteration and 12.70% to 2.50% in the first Arabic iteration. Figure 7 shows that the Arabic results outperformed both the Chinese

and Spanish Results, with an 80.31% decrease in recognition rate. Hence, of the three languages, the Arabic language would be the most effective when attempting to conceal writing style via iterative language translation (see Figure 8).

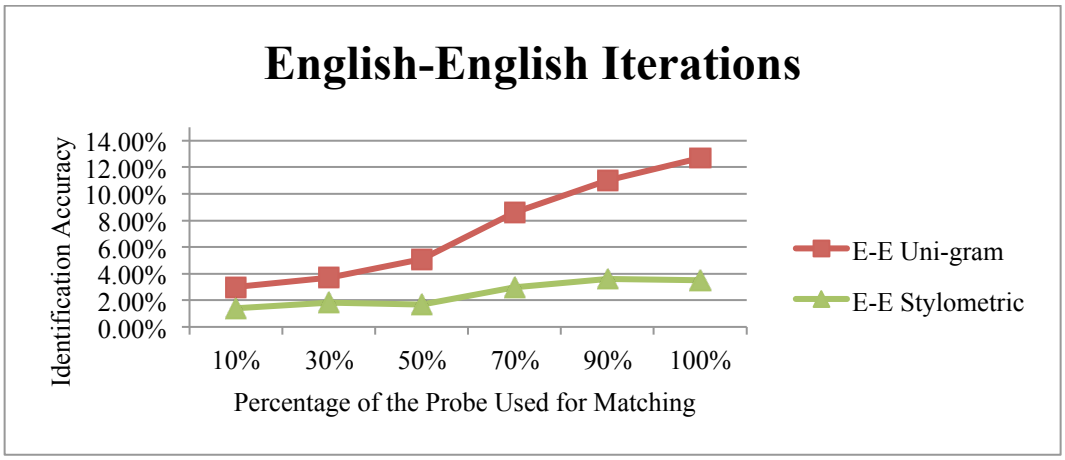

Figure 4. English Results: E-E

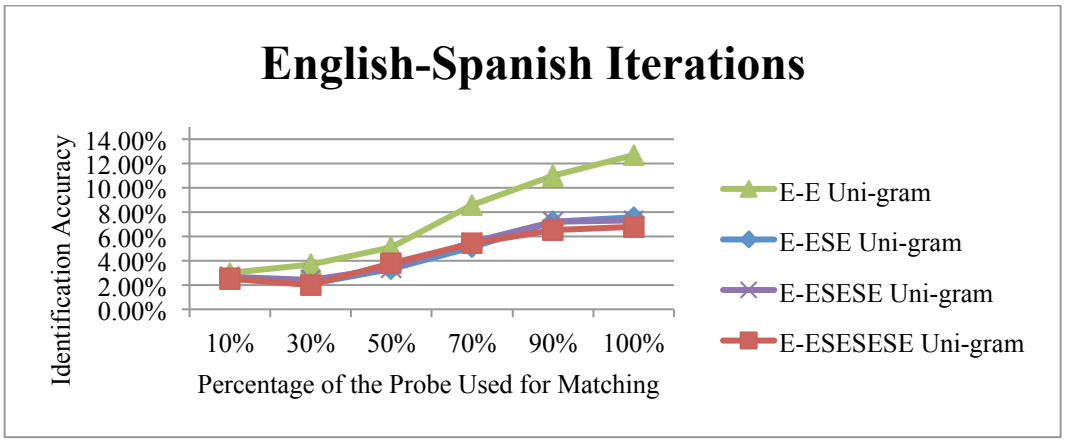

Figure 5. Spanish Results: E-ESE, E-ESESE, E-ESESESE

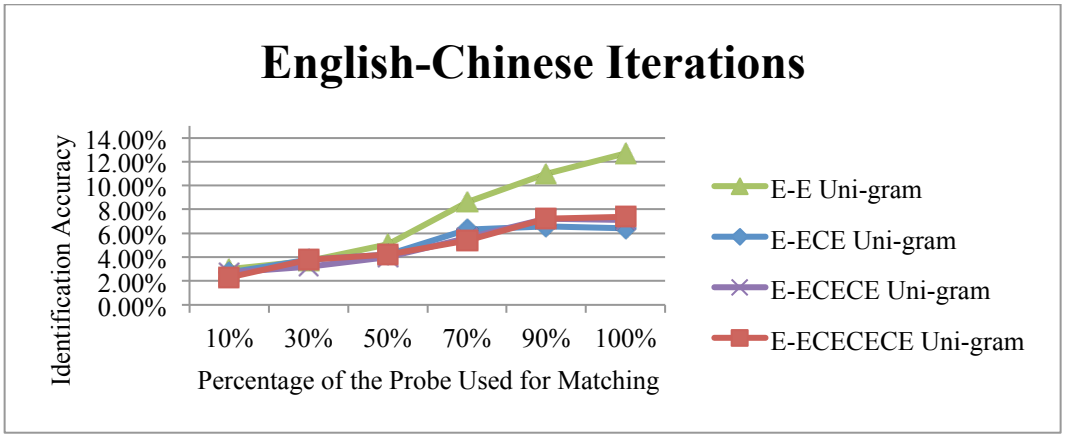

Figure 6. Chinese Results: E-ECE, E-ECECE, E-ECECECE

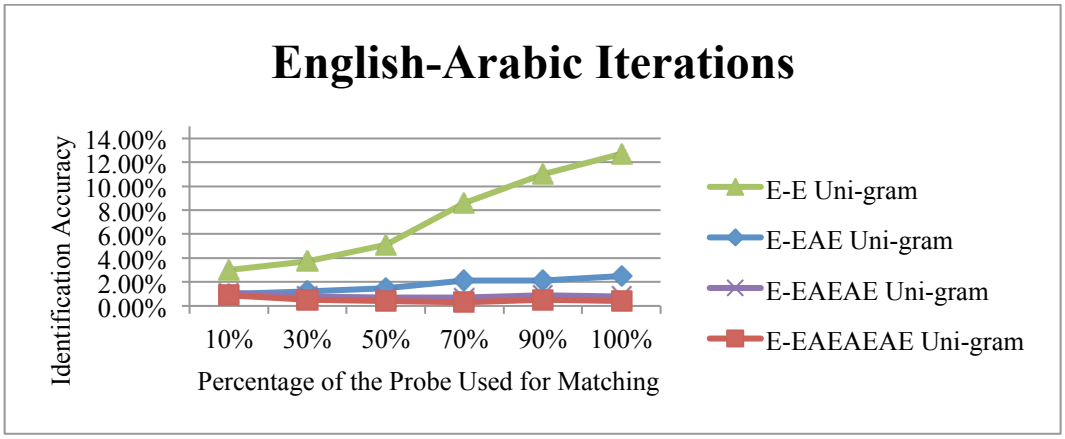

Figure 7: Arabic Results: E-EAE, E-EAEAE, E-EAEAEAE

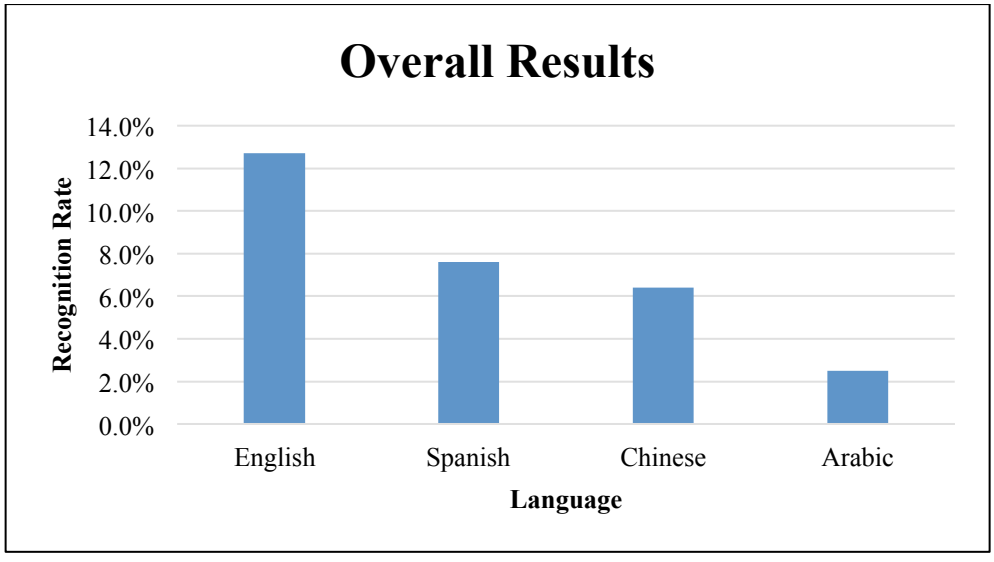

Figure 8: Overall Results: 100% Text

CHAPTER 5

Discussion and Future Research

In conclusion, this research strongly suggests that an author's writing style is altered once the text is translated into another language and back to the original language. This technique can be potentially implemented as a mobile application component that allows users to manipulate their text in order to increase the probability of remaining anonymous. Such anonymity would be an asset to entities, such as employees providing detailed feedback to their employer or students providing feedback to their professors. Future research includes the incorporation of other languages in order to determine the most effective translation language in terms of masking an author's identity.

References

[1] *(2014, Octoer 2014) Fact Sheet 35: Social Networking Privacy: How to be Safe, Secure and Social.* [Online]. Available: https://www.privacyrights.org/social-networking-privacy-how-be-safe-secure-and-social

[2] A. Narayanan et al, "On the Feasibility of Internet-Scale Author Identification. Security and Privacy (SP)," *in 2012 IEEE Symposium*, 2012, pp.300, 314.

[3] R. Forsyth, "Short substrings as document discriminators. *Joint International Conference of the Association for Computers* and the *Humanities and the Association for Literary & Linguistic Computing*," Queen's University, Kingston, Ontario, 1997.

[4] O. de Vel et al, "Mining e-mail content for author identification forensics," in *ACM SIGMOD* Record 30(4), 2001, pp. 55-64.

[5] *Ask.fm.* Available: http://www.linkedin.com/company/ask-fm

[6] *Ask.fm: A New View on Safety*. Available: http://ask.fm/about/safety/core-values

[7] *Fess: Anonymous High School Confessions. Available:* http://www.fessapp.com/

[8] *Whisper*. Available: https://play.google.com/store/apps/details?id=sh.whisper

[9] *Secret - FAQ*. Available: https://www.secret.ly/faq

[10] *Fess*. Available: https://play.google.com/store/apps/details?id=com.yml.fess

[11] *Fess Terms of Service*. Web. 25 Sept. 2014, from http://www.fessapp.com/fesstos

[12] *Omegle*. Available: http://www.omegle.com/

[13] *Ask.fm Privacy*. Available: http://ask.fm/about/policy/privacy-policy

[14] (2013, October 18) *Welcome to Spring.me!* Available: http://new.spring.me/#!/page/terms

[15] *Yik Yak Terms*. Available: http://www.yikyakapp.com/terms/

[16] *Yik Yak Features*. Available: http://www.yikyakapp.com/features/

[17] R. Green and J. Sheppard, "Comparing Frequency- and Style-Based Features for Twitter Author Identification" in *International Florida Artificial Intelligence Research Society Conference*, 2013.

[18] E. Stamatatos, "A survey of modern authorship attribution methods," in *Journal of the American Society for Information Science and Technology, 60*(3), 2009, 538-556.

[19] D. Madigan et al, "Author identification on the large scale," In *CSNA-05*, 2005.

[20] M. Coulthard, "Author Identification, Idiolect and Linguistic Uniqueness." *Applied Linguistics,* 2004, 25, 4, p. 431-447.

[21] N. Macleod and T. Grant, "Whose Tweet? Authorship analysis of micro-blogs and other short-form messages," in *International Association of Forensic Linguists' tenth biennial conference*, Aston University, 2012, p. 210-224.

[22] The differences between English and Arabic. Available: http://esl.fis.edu/grammar/langdiff/arabic.htm

[23] J. Houvardas and E. Stamatatos, "N-Gram Feature Selection for Authorship Identification," in 12th *International Conference on Artificial Intelligence: Methodology, Systems, Applications*, 2006, 77-86.

[24] L. Jiexun et al, "From Fingerprint to Writeprint," in *Commun. ACM*49, 4, 2006, 76-82.

[25] R. Zheng et al, "A framework for authorship identification of online messages: Writing-style features and classification techniques," in The Journal of the American Society for Information Science and Technology (JASIST) 57, 3 (2006), 378–393

[26] R. Williams et al, *The Fundamentals of Speech Communication in the Digital Age*. Dubuque, IA: KendallHunt, 2012, 1, pp. 6.

[27] R. Wall, amd U. Niehaus, “Russian to English machine translation with simple logical processing,” in *American Institute of Electrical Engineers, Part I: Communication and Electronics, Transactions* of the , vol.76, no.6, 1958, pp.709,714.

[28] I. Badr et al, “Segmetaion for English-to Arabic Statistical Machine Translation” in *ACL’08*, 2008,153-156.

[29] *Bag-of-words model* [Online]. Available: http://en.wikipedia.org/wiki/Bag-of-words_model

[30] *SYSTRAN 7 Business Translator* [Online]. Available: http://www.systransoft.com/translation-products/desktop/systran-7-business-translator/

[31] H. Williams, Author CAAT, 2014.

Appendix

Example of original English text compared to text translated into to Spanish and back to English

English: Apple unveiled a lot of new stuff yesterday, but the notebook announcements caught my attention because I've been thinking about upgrading to a new 15-inch MacBook Pro. It's been a full year since Apple has updated the MacBook Pro range, if you don't count the launch of the expensive Retina display models with a sharper screen but no DVD drive.
Apple wants people to source all their software and content from the iTunes store.

English-Spanish-English: Apple revealed much new matter yesterday, but the notices of the notebook took my attention because there am been thinking about the new increase in 15 MacBook Pro inches. It has been a complete year since Apple has bought up to date range of MacBook Pro, if you do not tell to the launching of the expensive models of exhibition of the retina with a acute screen but any impulsion of the DVD. Apple wants to people to the source all their software and content of the store of iTunes.